

Lecture Notes in Statistics 215
Proceedings

Mei-Ling Ting Lee · Mitchell Gail
Ruth Pfeiffer · Glen Satten
Tianxi Cai · Axel Gandy *Editors*

Risk Assessment and Evaluation of Predictions

 Springer

Edited by

P. Bickel, P. Diggle, S. E. Fienberg, U. Gather, I. Olkin, S. Zeger

For further volumes:

<http://www.springer.com/series/694>

Mei-Ling Ting Lee • Mitchell Gail
Ruth Pfeiffer • Glen Satten • Tianxi Cai
Axel Gandy
Editors

Risk Assessment and Evaluation of Predictions

 Springer

Editors

Mei-Ling Ting Lee
Department of Epidemiology
and Biostatistics
University of Maryland
College Park, MD, USA

Mitchell Gail
Division of Cancer Epidemiology
and Genetics
National Cancer Institute
Bethesda, MD, USA

Ruth Pfeiffer
National Cancer Institute
NCI Shady Grove, Bethesda, MD, USA

Glen Satten
Centers for Disease Control and Prevention
Atlanta, GA, USA

Tianxi Cai
Department of Biostatistics
Harvard School of Public Health
Boston, MA, USA

Axel Gandy
Department of Mathematics
Imperial College London
London, UK

ISSN 0930-0325

ISSN 2197-7186 (electronic)

ISBN 978-1-4614-8980-1

ISBN 978-1-4614-8981-8 (eBook)

DOI 10.1007/978-1-4614-8981-8

Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013953296

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

On October 12–14, 2011, the Biostatistics and Risk Assessment Center (BRAC) and the Department of Epidemiology and Biostatistics of the University of Maryland hosted an international conference entitled “Risk Assessment and Evaluation of Predictions.” The conference was held in Silver Spring, Maryland.

In assembling this volume, we invited conference participants to contribute their articles. All papers were peer-reviewed, by anonymous reviewers, and revised before final editing and acceptance. Although this process was quite time-consuming, we believe that it greatly improved the volume as a whole, making this book a valuable contribution to the field of research in risk assessment and evaluation of predictions.

This volume presents a broad spectrum of articles presented at the conference. It includes 21 chapters organized into three parts:

- Part I: Risk Assessment in Lifetime Data Analysis,
- Part II: Evaluation of Predictions, and
- Part III: Applications.

Part I includes different methods for risk assessment in survival analysis such as accelerated failure time models; threshold regression models; residual survival models; competing risks; Neyman, Markov processes and survival analysis; and nonparametric inference. Part II covers many important issues related to evaluation of risk predictions as well as recent advances in the development of receiver-operating characteristic (ROC) curves. Part III presents a variety of applications from genetics, competing risk models and breast cancer, product life cycle evaluation, environmental exposure biomarkers, extreme wind speed, and consumer services.

We hope that this volume will serve as a valuable reference for researchers in these important areas.

Maryland, USA
Bethesda, MD, USA
Bethesda, MD, USA
Atlanta, GA, USA
Boston, MA, USA
London, UK

Mei-Ling Ting Lee
Mitchell Gail
Ruth Pfeiffer
Glen Satten
Tianxi Cai
Axel Gandy

Acknowledgments

The conference on Risk Assessment and Evaluation of Predictions and this book volume would not have been possible without the help, support, and hard work of many people. I would like to thank (in alphabetical order) Chao Chen (EPA), Andrew N. Freedman (NCI), Mitchell H. Gail (NCI), Robert T. O’Neill (FDA), Ruth Pfeiffer (NCI), Antonio Possolo (NIST), and Sue-Jane Wang (FDA) for their help in organizing an interesting program. The conference was cosponsored by the Division of Cancer Control and Population Sciences at NCI, as well as the following divisions, schools and colleges, and centers at the University of Maryland at College Park, including Division of Research; School of Public Health, College of Agriculture and Natural Sciences; College of Computer, Mathematical, and Natural Sciences; Department of Mathematics; Department of Epidemiology and Biostatistics; Biostatistics and Risk Assessment Center; and the Maryland Institute of Applied Environmental Health. I also thank Mitchell Gail and Ruth Pfeiffer for giving a tutorial on “Absolute Risk Prediction”, and Margret Pepe for giving a tutorial on “Current Methods for Evaluating Prediction Performance of Biomarkers and Tests.” Both tutorials were well attended and we all learned a lot from them. The editors are grateful to the authors and the many anonymous reviewers for their efforts in preparing the manuscripts in this volume. Finally, I thank Xin He and Raul Cruz-Cano who helped to organize the conference logistics and ensure its success.

University of Maryland

Mei-Ling Ting Lee

Contents

Part I Risk Assessment in Lifetime Data Analysis

Non-proportionality of Hazards in the Competing Risks Framework	3
Alvaro Muñoz, Alison G. Abraham, Matthew Matheson, and Nikolas Wada	
Semiparametric Inference on the Absolute Risk Reduction and the Restricted Mean Survival Difference	23
Song Yang	
Connecting Threshold Regression and Accelerated Failure Time Models	47
Xin He and G.A. Whitmore	
Residuals and Functional Form in Accelerated Life Regression Models	61
Stein Aaserud, Jan Terje Kvaløy, and Bo Henry Lindqvist	
Neyman, Markov Processes and Survival Analysis	67
Grace Yang	
Quantiles of Residual Survival	87
Christopher Cox, Michael F. Schneider, and Alvaro Muñoz	

Part II Evaluation of Predictions

Methods for Evaluating Prediction Performance of Biomarkers and Tests	107
Margaret Pepe and Holly Janes	
Estimating Improvement in Prediction with Matched Case-Control Designs	143
Aasthaa Bansal and Margaret Sullivan Pepe	

ROC Analysis for Multiple Markers with Tree-Based Classification	179
Mei-Cheng Wang and Shanshan Li	
Assessing Discrimination of Risk Prediction Rules in a Clustered Data Setting	199
Bernard Rosner, Weiliang Qiu, and Mei-Ling Ting Lee	
Time-Dependent AUC with Right-Censored Data: A Survey	239
Paul Blanche, Aurélien Latouche, and Vivian Viallon	
Subgroup Specific Incremental Value of New Markers for Risk Prediction	253
Q. Zhou, Y. Zheng, and T. Cai	
Part III Applications	
Assessing the Effects of Imprinting and Maternal Genotypes on Complex Genetic Traits	285
Shili Lin	
Competing Risks Models and Breast Cancer: A Brief Review	301
Sharareh Taghipour, Dragan Banjevic, Anthony Miller, and Bart Harvey	
Quantifying Relative Potency in Dose-Response Studies	315
Gregg E. Dinse and David M. Umbach	
Development and Validation of Exposure Biomarkers to Dietary Contaminants Mycotoxins: A Case for Aflatoxin and Impaired Child Growth	333
Paul Craig Turner and Barbara Zappe Pasturel	
Pharmaceutical Risk Assessment and Predictive Enrichment to Maximize Benefit and Minimize Risk: Issues in Product Life Cycle Evaluation	349
Robert T. O'Neill	
A Multiple Imputation Approach for the Evaluation of Surrogate Markers in the Principal Stratification Causal Inference Framework	361
Xiaopeng Miao, Xiaoming Li, Peter B. Gilbert, and Ivan S.F. Chan	
Mapping Return Values of Extreme Wind Speeds	383
Adam L. Pintar and Franklin T. Lombardo	

Decision Analysis Methods for Selecting Consumer Services with Attribute Value Uncertainty	405
Dennis D. Leber and Jeffrey W. Herrmann	
Evaluating Incremental Values from New Predictors with Net Reclassification Improvement in Survival Analysis	425
Yingye Zheng, Layla Parast, Tianxi Cai, and Marshall Brown	
Erratum	E1

Part I
Risk Assessment in Lifetime Data Analysis

Non-proportionality of Hazards in the Competing Risks Framework

Alvaro Muñoz, Alison G. Abraham, Matthew Matheson, and Nikolas Wada

Abstract The simplest means of determining the effect of an exposure on the frequency and timing of two competing events is to contrast the cumulative incidences between the exposed and unexposed groups for each event type. Methods and software are widely available to semi-parametrically model the sub-hazards of the cumulative incidences as proportional and to test whether the constant relative sub-hazards (a_1 and a_2) are different from 1. In this chapter, we show that a_1 and a_2 are tethered by a strong relationship which is independent of the timing of the competing events; the relationship is fully determined by the overall frequencies of events, and a_1 and a_2 must be on opposite sides of 1. When violations of proportionality occur, separate analyses for the two competing events often yield an inadmissible result in which the estimates of a_1 and a_2 are on the same side of 1, and may even exhibit statistical significance. We further characterize the compatibility of concurrent proportionality of cause-specific hazards and sub-hazards, and show that strong tethering also occurs among these quantities; and that, of the sub-hazards

A. Muñoz (✉)

Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street – E7648,
Baltimore, MD 21205, USA
e-mail: amunoz@jhsph.edu

A.G. Abraham

Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street – E7640,
Baltimore, MD 21205, USA
e-mail: aabraham@jhsph.edu

M. Matheson

Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street – E7009,
Baltimore, MD 21205, USA
e-mail: mmatheso@jhsph.edu

N. Wada

Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street – E7650,
Baltimore, MD 21205, USA
e-mail: nwada@jhsph.edu

and cause-specific hazards, at most two of the four can be proportional, but without restriction on which two. Because proportionality rarely holds in practice, the default analytical approach should allow the relative hazards to depend on time, which can be easily carried out with widely available software. However, the statistical power of this approach is limited in the case of large numbers of event-free observations. An application using data from a North American cohort study of children with kidney disease is presented.

Introduction

The problem of competing risks has been addressed in the literature in several ways [1]. The most common approach, that of cause-specific hazards, partitions the hazard of the composite event (instantaneous probability of failing from any event among survivors of all events) as the sum of mutually exclusive cause-specific hazards (instantaneous probability of failing from a specific cause among survivors of all events). This approach allows for estimation and testing of relative cause-specific hazards by treating the times to one event as censored observations for the other. The cause-specific hazards approach has been well developed in the literature [1–7] and, under the assumption of proportionality of the cause-specific hazards, can be carried out using standard software for proportional hazards. However, interpretative challenges arise because effects of exposures on cause-specific hazards may not mirror effects on cumulative incidences [1].

A second approach, centered on the sub-hazards of the cause-specific cumulative incidences (often referred to as subdistribution hazards), is to consider those who experience the competing event as immune to the event of interest. For example, those who died of a cardiovascular cause persist in the analysis of renal failure death as part of the risk set, but of course cannot experience renal death for the remainder of the study. The appeal of such a strategy is that it reflects the reality of events in the study population, in which some will have the event of interest and others never will as a result of competing events. Within this framework, cumulative incidence functions in the case of two event types, $I_1(t)$ and $I_2(t)$, are directly estimated, and comparisons between groups are made via the associated sub-hazards $\lambda_1(t)$ and $\lambda_2(t)$ [8–10].

The most widely-used model for the effect of an exposure on the sub-hazards assumes proportional sub-hazards, as described by Fine and Gray [8]. This method distributes the information from censored times to competing events using a weighting procedure. Software to carry out this method is widely available and simple to use (e.g., `stcrreg` of Stata). However, proportionality of sub-hazards rarely holds. Indeed, in their seminal paper, Fine and Gray had the foresight to warn that “In applications, we anticipate time \times covariate interactions” [8]. Furthermore, since the total cumulative incidence must sum to 1 at $t = \infty$, any exposure-induced increase in the cumulative incidence of one event type must be offset by a decrease in the cumulative incidence of the other event type. This induces interdependence, or

tethering, of the relative sub-hazards with important consequences if proportionality is assumed for both events. To circumvent this tethering, Beyersmann et al. [11] recommend assuming proportionality for only one of the events. However, in many epidemiological studies [12–16], determination of the effect of exposures on all event types is of primary interest.

In the “[Methods and Models](#)” section of this chapter, we characterize the tethering of relative sub-hazards under the assumption of simultaneous proportionality for both events, and we extend results about the compatibility of concurrent proportionality of cause-specific and sub-hazards for the two event types by proving that at most two of the four measures can be proportional, but without restriction on which two. In the “[Simulation](#)” section we illustrate the consequences of tethering of relative sub-hazards using data simulated from a mixture of the conditional distributions of the times of the competing events [17–19]. In the “[Application](#)” section, data from a North American cohort study of children with chronic kidney disease are used to illustrate a case congruent with proportionality of sub-hazards and a case with strong time dependency of the relative sub-hazards. Limitations posed by data with heavy censoring are included in the discussion.

Methods and Models

Sub-hazards

The one-sample survival analysis problem for two competing events ($E=1$ and $E=2$) wherein only one event is observed per subject can be fully described by a mixture of two distributions determined by (1) the mixture parameter $\pi = P(E=1) = 1 - P(E=2)$, which describes the overall frequency of each event; and (2) the conditional distribution functions, $F_1(t) = P(T \leq t | E=1)$ and $F_2(t) = P(T \leq t | E=2)$, which govern the timing of the events, with T representing the time for the composite event. Hereafter, for simplicity, we drop the notation “(t)” from functions except in cases where necessary for clarity. The cumulative incidence functions (I_1 and I_2) for the two events follow from weighting the conditional distributions by the mixture parameter π as: $I_1 = \pi F_1$ and $I_2 = (1 - \pi) F_2$. Thus, as $t \rightarrow \infty$, I_1 approaches π and I_2 approaches $1 - \pi$. The sub-hazards λ_1 and λ_2 that correspond to the cumulative incidences are $\lambda_i = I_i' / (1 - I_i)$, where I_i' is the derivative of I_i for $i = 1, 2$. This representation of the sub-hazards can be re-expressed as

$$\lambda_i(t) = \frac{P(T \in dt, E = i)}{P(T > t) + P(T \leq t, E \neq i)} = \frac{P(T \in dt, E = i)}{P(T > t, E = i) + P(E \neq i)}.$$

Thus, the sub-hazard is estimated by the number of individuals who experienced the event of interest at time t , divided by all those who remained free of any event

at t plus those who had experienced the competing event prior to t ; or equivalently, divided by those who experienced the event of interest after t plus all those who ever experienced the competing event. From the second representation in the above equation, it is easily seen that the sub-hazards are smaller than the conditional hazards defined by $P(T \in dt, E = i)/P(T > t, E = i) = F_i'(t)/(1 - F_i(t))$.

If λ_1 and λ_2 are known (hereafter using $\int_0^t \lambda_i$ to represent $\int_0^t \lambda_i(x)dx$), the cumulative incidences and the mixture parameter can be recovered by $I_i(t) = 1 - e^{-\int_0^t \lambda_i}$ for $i = 1, 2$; and $\pi = 1 - e^{-\int_0^\infty \lambda_1} = e^{-\int_0^\infty \lambda_2}$. Therefore, $1 - \pi = 1 - e^{-\int_0^\infty \lambda_2} = e^{-\int_0^\infty \lambda_1}$. Hence, λ_1 and λ_2 must satisfy

$$\exp\left(-\int_0^\infty \lambda_1\right) + \exp\left(-\int_0^\infty \lambda_2\right) = 1. \quad (1)$$

Relative Sub-hazards in the Two-Sample Setting and the Case of Proportional Sub-hazards

The two-sample problem (unexposed vs. exposed) expands on the previously described relationships. In general, an exposure (functions and parameters of the exposed group are identified hereafter with $*$) may result in a change of the mixture parameter (from π to π^*) and/or the timing of one or both events (from F_i to F_i^*). Let $A_i(t) = \lambda_i^*(t)/\lambda_i(t)$ denote the relative sub-hazards for event i for $i = 1, 2$. Since the sub-hazards in the exposed group must also fulfill Eq. 1, it follows that A_1 and A_2 must satisfy $\exp\left(-\int_0^\infty A_1(t)\lambda_1(t)dt\right) + \exp\left(-\int_0^\infty A_2(t)\lambda_2(t)dt\right) = 1$.

This relationship is fairly flexible as long as $A_1(t)$ and $A_2(t)$ remain functions of time. However, if proportionality is assumed such that $A_1(t) \equiv a_1$ and $A_2(t) \equiv a_2$, then $1 - I_i^*(t) = \exp\left(-a_i \int_0^t \lambda_i\right) = (1 - I_i(t))^{a_i}$ for $i = 1, 2$. Evaluating these equations as $t \rightarrow \infty$ yields $1 - \pi^* = (1 - \pi)^{a_1}$ and $\pi^* = \pi^{a_2}$. Hence, it follows that the constant relative sub-hazards a_1 and a_2 are fully determined by the mixture parameters as $a_1 = \log(1 - \pi^*)/\log(1 - \pi)$ and $a_2 = \log(\pi^*)/\log(\pi)$. Therefore, if the sub-hazards are proportional, the relative sub-hazards do not depend on the timing of the events, but simply depend on the overall frequencies of each event type in the exposed and unexposed groups. Furthermore, except for the null setting, the constant relative sub-hazards must lie on opposite sides of 1 (i.e., if $\pi^* > \pi$ then $a_1 > 1$ and $a_2 < 1$; and if $\pi^* < \pi$ then $a_1 < 1$ and $a_2 > 1$). Intuitively, any exposure-induced increase in the cumulative incidence of one event type must be offset by a decrease in the cumulative incidence of the other event type because the total cumulative incidences for the unexposed and the exposed groups must sum to 1 at $t = \infty$. This tethering of the two relative sub-hazards means that there is effectively only one relative sub-hazard, since the other is then completely determined [11].

Compatibility of Proportionality of One Sub-hazard and One Cause-Specific Hazard

The tethered relationship between the relative sub-hazards highlighted in the previous section is one example of the strong interdependences that exist among the sub-hazards and cause-specific hazards when proportionality is assumed. These relationships have been explored to some extent in the literature. Beyersmann et al. [11] described methods to simulate data from cause-specific hazards models that are consistent with proportional sub-hazards for the event of interest. These methods provide a general approach centered on data generation for probing the bounds of consistency between proportional cause-specific hazards and sub-hazards. In the sections “[Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type](#)” and “[Compatibility of Proportionality of Sub-hazards for One Event Type and Cause-Specific Hazards for the Other Event Type](#)” we fully characterize the tethering relationships that arise from proportionality shared between the cause-specific hazards and sub-hazards, and we also provide explicit expressions for the cumulative incidences, which in turn simplify procedures for data simulation. Further, in the sections “[Incompatibility of Proportional Sub-hazards for One Event Type and Proportional Cause-Specific Hazards for Both Event Types](#)” and “[Incompatibility of Proportional Sub-hazards for Both Event Types and Proportional Cause-Specific Hazards for One Event Type](#)”, we show that no combination of three of the four hazards can be simultaneously proportional. An immediate consequence of this result is that proportionality of the two cause-specific hazards and of the two sub-hazards cannot simultaneously occur, confirming previous reports [20–22].

The cause-specific hazards μ_i of the unexposed group corresponding to the proportions of individuals experiencing event type i among those remaining free of any event are defined by $\mu_i = I_i' / (1 - I_1 - I_2)$ for $i = 1, 2$. A similar definition follows for the cause-specific hazards μ_i^* of the exposed group, and we denote the relative cause-specific hazards by $B_i(t) = \mu_i^*(t) / \mu_i(t)$. Throughout, we let π , F_1 , and F_2 define the mixture of the competing events for the reference group, and use lowercase letters a_i and b_i to denote constant relative sub-hazards and constant relative cause-specific hazards, respectively.

Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type

If we allow the sub-hazards of event type 1 (without loss of generality) to be proportional, then from the section “[Relative Sub-hazards in the Two-Sample Setting and the Case of Proportional Sub-hazards](#)” we have $I_1^* = 1 - (1 - I_1)^{a_1}$ and thus $I_1^{*'} = a_1(1 - I_1)^{a_1 - 1} I_1'$. If we further allow $B_1(t) \equiv b_1$, then $b_1 \equiv a_1(1 - I_1)^{a_1 - 1} (1 - I_1 - I_2) / (1 - I_1^* - I_2^*)$, which, letting $t \rightarrow 0$, yields $b_1 = a_1$ and thus $1 - I_1^* - I_2^* = (1 - I_1)^{a_1 - 1} (1 - I_1 - I_2)$. Therefore, $I_2^* = (1 - I_1)^{a_1 - 1} I_2$. The

right-hand side of this equation is 0 at $t=0$ and converges to $(1-\pi)^{a_1}$ as $t \rightarrow \infty$. However, for it to be increasing (i.e., its derivative to be positive), a_1 must be $\leq 1 + \text{minimum}[(1-I_1)I_2'/(I_1'I_2)]$. This upper bound is always a finite number ≥ 1 .

Hence, in this case of the relative sub-hazards and relative cause-specific hazards for the same event type being constant, the tethering between the two constant relative hazards is the strongest, as they must be equal and the common constant has an upper bound ≥ 1 which is a function of I_1 and I_2 . For a_1 in the allowable range, the cumulative incidences for the exposed group are explicitly determined by $I_1^* = 1 - (1-I_1)^{a_1}$ and $I_2^* = (1-I_1)^{a_1-1}I_2$.

Compatibility of Proportionality of Sub-hazards for One Event Type and Cause-Specific Hazards for the Other Event Type

In this case, $\lambda_1^*/\lambda_1 \equiv a_1 \neq 1$ and $\mu_2^*/\mu_2 \equiv b_2 \neq 1$. From the former it follows that $I_1^* = 1 - (1-I_1)^{a_1}$ and $I_1^{*'} = a_1(1-I_1)^{a_1-1}I_1'$, and we need to determine I_2^* to fulfill the latter. From the equations

$$\mu_1^*(1-I_1^*-I_2^*) = I_1^{*'} \quad (2)$$

and

$$1-I_1^*-I_2^* = e^{-\int \mu_1^* + \mu_2^*} = e^{-\int \mu_1^*} e^{-\int \mu_2^*} = e^{-\int \mu_1^*} e^{-b_2 \int \mu_2} \quad (3)$$

we can solve for μ_1^* and I_2^* . Namely, substituting Eq. 3 into Eq. 2 yields $\mu_1^* e^{-\int \mu_1^*} = e^{b_2 \int \mu_2} I_1^{*'}$, and the left-hand side of this equation is equal to the derivative of $-e^{-\int \mu_1^*}$. Thus, integrating both sides from 0 to t yields

$$\exp\left(-\int_0^t \mu_1^*\right) = 1 - \int_0^t \exp\left(b_2 \int_0^x \mu_2\right) I_1^{*'}(x) dx. \quad (4)$$

For the right-hand side of this equation to remain positive, b_2 must remain beneath an upper bound which depends on π, F_1, F_2 and a_1 ; and, specifically, is inversely related to a_1 . Thus, the two constants are tethered, though in a weaker fashion than in the case of the two sub-hazards being proportional (section “[Relative Sub-hazards in the Two-Sample Setting and the Case of Proportional Sub-hazards](#)”) or the case of the sub-hazards and cause-specific hazards of the same event type being proportional (section “[Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type](#)”).

Substituting from Eq. 4 into Eq. 3 allows us to express I_2^* as

$$I_2^*(t) = 1 - I_1^*(t) - \exp\left(-b_2 \int_0^t \mu_2\right) \left[1 - \int_0^t \exp\left(b_2 \int_0^x \mu_2\right) I_1^{*'}(x) dx\right]$$

which fulfills the properties of a cumulative incidence, with everything on the right-hand side being a known quantity determined by π, F_1, F_2, a_1 and b_2 .

Incompatibility of Proportional Sub-hazards for One Event Type and Proportional Cause-Specific Hazards for Both Event Types

As shown in the section “[Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type](#)”, if the cause-specific hazards and the sub-hazards are both proportional for event type 1, then the relative sub-hazard a_1 in the allowable range must be equal to the relative cause-specific hazard b_1 , and

$$I_2^* = (1 - I_1)^{a_1 - 1} I_2 \quad (5)$$

and

$$1 - I_1^* - I_2^* = (1 - I_1)^{a_1 - 1} (1 - I_1 - I_2) \quad (6)$$

From Eq. 5, it follows that

$$\lim_{t \rightarrow 0} I_2^*/I_2 = \lim_{t \rightarrow 0} (1 - I_1)^{a_1 - 1} = 1. \quad (7)$$

If we further assume that the cause-specific hazards for type 2 events are proportional, then $b_2 \equiv \frac{\mu_2^*}{\mu_2} = \frac{I_2^{*'}}{I_2'} \frac{1 - I_1 - I_2}{1 - I_1^* - I_2^*} = \frac{I_2^{*'}}{I_2'} \frac{1}{(1 - I_1)^{a_1 - 1}}$ so that $I_2^{*'} = b_2 (1 - I_1)^{a_1 - 1} I_2'$; and from this equation it follows that

$$\lim_{t \rightarrow 0} I_2^{*'} / I_2' = \lim_{t \rightarrow 0} b_2 (1 - I_1)^{a_1 - 1} = b_2. \quad (8)$$

But, by l'Hôpital's rule, $\lim_{t \rightarrow 0} I_2^*/I_2 = \lim_{t \rightarrow 0} I_2^{*'} / I_2'$; hence, from Eqs. 7 and 8 we would obtain $b_2 = 1$. Since $\mu_1^* = a_1 \mu_1$ and $\mu_2^* = \mu_2$, then

$$1 - I_1^* - I_2^* = \left(e^{-\int \mu_1}\right)^{a_1} e^{-\int \mu_2} = (1 - I_1 - I_2) \left(e^{-\int \mu_1}\right)^{a_1 - 1} \quad (9)$$

Hence, from Eqs. 6 and 9, it follows that $(e^{-\int \mu_1})^{a_1-1} = (1 - I_1)^{a_1-1} = (e^{-\int \lambda_1})^{a_1-1}$; therefore, $\mu_1 = \lambda_1$, which is impossible because for any event type the cause-specific hazard is by definition greater than the sub-hazard.

Incompatibility of Proportional Sub-hazards for Both Event Types and Proportional Cause-Specific Hazards for One Event Type

We assume the type 1 event is the one for which both the cause-specific hazards and the sub-hazards are proportional. Therefore, $a_1 = b_1$, and $I_2^* = (1 - I_1)^{a_1-1} I_2$ with a_1 in the allowable range. If we further assume proportional sub-hazards for event type 2 ($\lambda_2^*/\lambda_2 \equiv a_2 \neq 1$), then $I_2^* = 1 - (1 - I_2)^{a_2}$. Equating the two expressions for I_2^* gives $(1 - I_1)^{a_1-1} I_2 = 1 - (1 - I_2)^{a_2}$. Hence, $(1 - I_1)^{a_1-1} = \frac{1 - (1 - I_2)^{a_2}}{I_2}$ for all $t > 0$. Taking the limit as t approaches 0 and applying l'Hôpital's rule, we obtain the result $a_2 = 1$, which is inadmissible since a_1 and a_2 must be on the opposite sides of 1, and $a_1 \neq 1$.

Methods to Simulate Data Fulfilling Proportionality of Hazards

A byproduct of the characterizations presented in the sections “[Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type](#)” and “[Compatibility of Proportionality of Sub-hazards for One Event Type and Cause-Specific Hazards for the Other Event Type](#)” is that, for simulating data with proportional sub-hazards for event type 1 and proportional cause-specific hazards for either event type, they provide a simple alternative approach to the general methods presented by Beyersmann et al. [20].

If the two cumulative incidence functions for a group are known, the simulation of competing risks data is straightforward. Specifically, to generate n observations from the unexposed group with given cumulative incidences I_1 and I_2 , sampling from a binomial $(n, I_1/I_1(\infty))$ yielding k will result in the need to generate k times from $F_1 = I_1/I_1(\infty)$ and $(n - k)$ times from $F_2 = I_2/I_2(\infty)$. For the exposed group, the sections “[Relative Sub-hazards in the Two-Sample Setting and the Case of Proportional Sub-hazards](#)” and “[Compatibility of Proportionality of Sub-hazards and of Cause-Specific Hazards for the Same Event Type](#)” provide explicit expressions for the cumulative incidences based on fixing a_1 when the two sub-hazards are proportional and when the sub-hazards and cause-specific hazards of the same type are proportional. The section “[Compatibility of Proportionality of Sub-hazards for One Event Type and Cause-Specific Hazards for the Other Event Type](#)” provides explicit expressions for the cumulative incidences based

on fixing a_1 and b_2 when sub-hazards of type 1 and cause-specific hazards of type 2 are proportional. The simulation of data for two relative cause-specific hazards being constant (i.e., $\equiv b_1$ and $\equiv b_2$, respectively) follows from the well-known fact that the cumulative incidences for the exposed group are determined by $I_i^* = b_i \int \mu_i e^{-\int (b_1 \mu_1 + b_2 \mu_2)}$ for $i = 1, 2$. Censoring times can be generated by standard procedures.

In summary, in this section we have shown that at most two of the four hazards (i.e., sub-hazards and cause-specific hazards of event types 1 and 2) can be proportional, but without restriction on which two. Furthermore, except for the case of proportionality of the two cause-specific hazards, the constant relative hazards are tethered, which in turn provides explicit and simple approaches to simulate data subjected to various forms of allowable proportionalities.

Simulation

A drawback of classical competing risks analysis via the method of Fine and Gray is that estimating the relative sub-hazards independently (i.e., untethered) can lead to results where both estimates are on the same side of 1. It is important to note that these undesirable results are not simply due to the incomplete information provided by censored observations. Even in situations in which all event times are observed, traditional analysis that incorrectly assumes proportional sub-hazards may lead to results that are theoretically inconsistent.

Consequences of incorrectly assuming proportionality of sub-hazards can be illustrated using a simulated example in which non-proportionality of sub-hazards holds. We restricted our example to the case of complete observation of event times to illustrate the drawbacks of the proportionality assumption even in the absence of censoring. Table 1 describes the components of the model used to simulate the data. Specifically, the mixture parameter was set to $\pi = 0.55$ in the unexposed group and $\pi^* = 0.50$ in the exposed group. The times of the two events in the unexposed group were drawn from the same exponential distribution with median = 20 (i.e., $F_1(t) = F_2(t) = 1 - \exp(-0.035t)$), and the times of the two events in the exposed group were shorter by a factor of 4 (i.e., median = 5, $F_1^*(t) = F_2^*(t) = 1 - \exp(-0.140t)$). In this setting, the cause-specific hazards are constant and thus proportional, and their ratios for the exposed to unexposed groups are $3.64 = 0.50 \times 0.140 / (0.55 \times 0.035)$ and $4.44 = 0.50 \times 0.140 / (0.45 \times 0.035)$ for type 1 and 2 events, respectively. In order to contrast the inferences drawn from different approaches, we generated 1,000 observations for each group.

Panels a and b of Fig. 1 display the sub-hazard functions for each event type among the exposed (continuous lines) and unexposed (dashed lines) groups. The four sub-hazards (fully described in Table 1) are decreasing, with those of the exposed group being steeper and crossing those of the unexposed group at times 15.5 and 18.0 for type 1 and type 2 events, respectively.

Table 1 Specification of true model as mixture of exponential distributions with non-proportional sub-hazards

	Type 1 event	Type 2 event
Unexposed		
Frequency	$\pi = 55\%$	$1 - \pi = 45\%$
Timing, $F_i(t)$	$1 - \exp(-0.035t)$	$1 - \exp(-0.035t)$
Cause-specific hazard, $\mu_i(t)$	0.55×0.035	0.45×0.035
Sub-hazard ^a , $\lambda_i(t)$	$\log(0.55 \times 0.035) - 0.035t - \log(0.55e^{-0.035t} + 0.45)$	$\log(0.45 \times 0.035) - 0.035t - \log(0.55 + 0.45e^{-0.035t})$
Exposed		
Frequency	$\pi^* = 50\%$	$1 - \pi^* = 50\%$
Timing, $F_i^*(t)$	$1 - \exp(-0.140t)$	$1 - \exp(-0.140t)$
Cause-specific hazard, $\mu_i^*(t)$	0.50×0.140	0.50×0.140
Sub-hazard ^a , $\lambda_i^*(t)$	$\log(0.50 \times 0.140) - 0.140t - \log(0.50e^{-0.140t} + 0.50)$	$\log(0.50 \times 0.140) - 0.140t - \log(0.50 + 0.50e^{-0.140t})$
Relative cause-specific hazard, $B_i(t)$	3.64	4.44
Relative sub-hazard ^a , $A_i(t)$	$1.291 - 0.105t - \log(0.50e^{-0.140t} + 0.50) + \log(0.55e^{-0.035t} + 0.45)$	$1.492 - 0.105t - \log(0.50 + 0.50e^{-0.140t}) + \log(0.55 + 0.45e^{-0.035t})$

^aLogarithmic scale

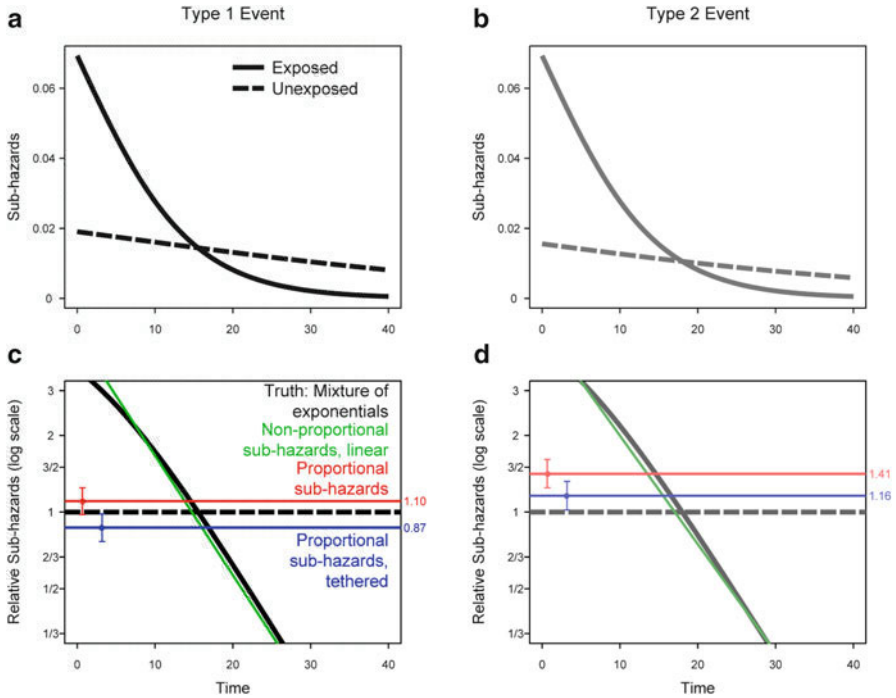


Fig. 1 Sub-hazards (panels **a** and **b**) and relative sub-hazards (panels **c** and **d**) for the competing risks setting defined by the true model in Table 1. Thick dashed and continuous lines correspond to the unexposed and exposed groups, respectively. Panels **c** and **d**, with a common legend shown in panel **c**, show results of three different approaches (non-proportional sub-hazards (*in green*), proportional sub-hazards (*in red*), and proportional sub-hazards with tethering (*in blue*)) for the analysis of 1000 simulated observations for each group

The logarithms of the relative sub-hazards for each event type are shown at the bottom of Table 1. Both are dominated by downward linear trends with equal slope (0.105) corresponding to the difference between the two hazards of the exponential distributions ($= 0.140 - 0.035$). The thick continuous lines in panels c and d of Fig. 1 depict the highly time-dependent true relative sub-hazards for each event type, with a dashed thick line at 1 for the unexposed (reference) group.

The first analytical approach for the simulated data was the traditional Fine and Gray method. In this case with no censored observations, the analysis reduces to a standard Cox regression where times for one event type are treated as censored observations past the largest observed time for the other event. The results of this analysis are presented in the first row of Table 2 and depicted as red horizontal lines in panels c and d of Fig. 1. Here, we found the theoretically inconsistent result of both constant relative sub-hazards estimates being above 1 ($1.10 = \exp(0.099)$)

Table 2 Results (logarithm of relative sub-hazards) of fitting models to 1000 unexposed and 1000 exposed observations from true models in Table 1

Model	Log relative sub-hazard \pm standard error	
	Type 1 event	Type 2 event
Proportional sub-hazards	0.099 \pm 0.062	0.346 \pm 0.065
Proportional sub-hazards, tethered	-0.142 \pm 0.063	0.148 \pm 0.066
Non-proportional sub-hazards ^a		
Intercept	1.591 \pm 0.116	1.671 \pm 0.117
Time	-0.108 \pm 0.008	-0.098 \pm 0.008

^aStata code for type 1 event: `stset time, failure(event==1)`

`sterreg exposure, compete(event==2) tvc(exposure) texp(_t) noshr`

for type 1 and $1.41 = \exp(0.346)$ for type 2), with p-values of 0.110 and <0.001 for event types 1 and 2, respectively. In this example, the analysis under the proportional sub-hazards assumption results in an estimate which provides a very poor summary of the impact of exposure on the risk of the two events.

The second analytical approach avoids both constant relative sub-hazards being on the same side of 1 by using $a_1 = \log(1 - \pi^*)/\log(1 - \pi)$ and $a_2 = \log(\pi^*)/\log(\pi)$. In the case of no censored data, estimates of π and π^* are simply the observed proportions of the event of interest in the unexposed ($\hat{\pi} = 0.55$) and exposed groups ($\hat{\pi}^* = 0.50$), respectively. The delta method was used to calculate the standard errors. The results of this analysis are presented in the second row of Table 2 and depicted as blue horizontal lines in panels c and d of Fig. 1. Although they indeed provide estimates of the relative sub-hazards on opposite sides of 1 ($0.87 = \exp(-0.142)$ for type 1 events and $1.16 = \exp(0.148)$ for type 2 events; $p < 0.05$ for both), the result is still a poor summary in comparison to the true time-varying relative sub-hazards (Fig. 1, Panels c and d).

An improvement to the summary of the relative sub-hazards can be achieved by including in the model an additional time-dependent term such that the total effect of exposure is a linear function of a fixed intercept and a time interaction, thus relaxing the rigid relationship enforced by the proportionality assumption. This single additional term, which introduces a linear time dependency, yields the flexibility to better summarize non-proportional sub-hazards, avoids theoretical inconsistency, and is easily implementable with standard software packages (e.g., by simply using the `tvc` and `texp` options of `sterreg` in Stata as shown in the footnote of Table 2). The results of this analysis are presented at the bottom of Table 2 and indicate a highly significant downward trend of the relative sub-hazards. The improved fit to the data using time-dependent relative sub-hazards as a linear trend with time is apparent in Fig. 1, panels c and d (green lines).

Application

Study Population, Outcomes, Exposures and Analytical Approaches

To explore the use of different approaches from the sub-hazards perspective, we analyzed data from the Chronic Kidney Disease in Children cohort study (CKiD), a North American study of chronic kidney disease in children [23]. Data at baseline were collected on 586 children between the ages of 1 and 16 with kidney function measured by glomerular filtration rate between 30 and 90 ml/min|1.73 m², and they were followed up at annual visits. For our analyses, the event of interest was end-stage renal disease (ESRD), defined as dialysis or a glomerular filtration rate less than 15 ml/min|1.73 m²; the competing event was kidney transplantation. There were 578 (99%) patients with adequate follow-up and event data. The time scale for our analysis was years since baseline visit, with a median of 3.1 years (upper quartile: 4.0 years). Patients with no event as of their most recent follow-up visit were censored at the last date seen, with a median follow-up of 3.3 years.

To illustrate the analytical approaches we selected two binary exposures. The first is a broad measure of socioeconomic status: household annual income greater than \$36,000. The second is a well-known biological predictor of progression of chronic kidney disease: nephrotic proteinuria, defined as a urine protein to creatinine ratio > 2. Information on proteinuria was available for the full cohort, while household income was available for 97.6% of the subjects. We used two semi-parametric approaches: the Fine and Gray model assuming proportional sub-hazards, and an extension of this model to allow linear time dependency in the logarithm of the relative sub-hazards.

Results

The top part of Table 3 provides the number of observed events among the 233 and the 331 children with annual household income less than or equal to \$36,000 and greater than \$36,000, respectively. The non-parametric cumulative incidences for ESRD and transplant are shown in Fig. 2, panels a and b, respectively. Traditional Fine and Gray regression yielded a constant relative sub-hazard estimate for ESRD, comparing those with household income greater than \$36,000 per year to those with less, of 0.5 ($= \exp(-0.656)$, 95% CI: 0.3, 0.8), as shown in Table 3. Allowing for a linear departure from proportionality, there was no indication of time interaction, as the coefficient for time shown at the bottom of Table 3 was only 0.006 and far from being statistically significant. In turn, the constant relative sub-hazard estimate for transplant was 1.7 ($= \exp(0.512)$ in Table 3) with 95% CI from 0.9 to 3.2. Similarly to ESRD, there was no departure from proportionality in the relative sub-hazard for

Table 3 Number of observed events and relative sub-hazards (logarithmic scale) of ESRD and transplant for household annual income in children with CKD

	Event type	
	ESRD	Transplant
Household annual income	Number of observed events	
≤\$36,000 (N = 233, reference)	36	13
>\$36,000 (N = 331)	29	32
Model	Log relative sub-hazard ± standard error	
Proportional sub-hazards	-0.656 ± 0.248	0.512 ± 0.329
Non-proportional sub-hazards		
Intercept	-0.665 ± 0.428	0.328 ± 0.876
Time		0.074 ± 0.331
	0.006 ± 0.210	

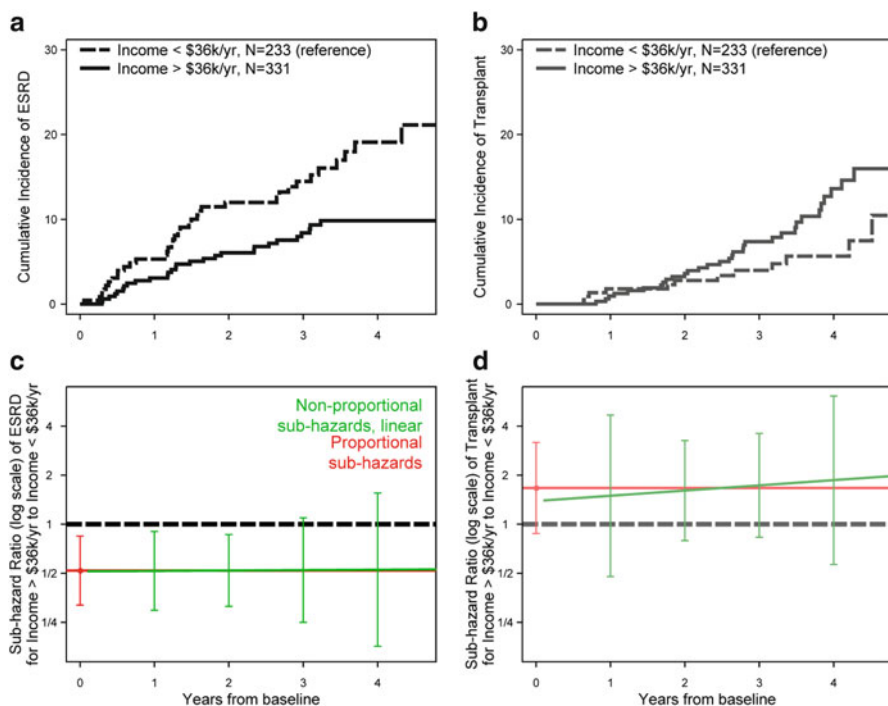


Fig. 2 Effect of annual income below/above \$36,000 on the competing events of end-stage renal disease (ESRD) and renal transplantation in the CKiD study. Panels **a** and **b** show estimates of cumulative incidences using non-parametric methods. Panels **c** and **d** depict relative sub-hazards under proportionality (*in red*) and linear departure from proportionality (*in green*)

Table 4 Number of observed events and relative sub-hazards (logarithmic scale) of ESRD and transplant for nephrotic proteinuria in children with CKD

	Event type	
	ESRD	Transplant
Nephrotic proteinuria	Number of observed events	
No (N = 501, reference)	30	31
Yes (N = 77)	38	16
Model	Log relative sub-hazard ± standard error	
Proportional sub-hazards	2.408 ± 0.246	1.206 ± 0.304
Non-proportional sub-hazards		
Intercept	3.609 ± 0.479	1.408 ± 0.707
Time	-0.721 ± 0.254	-0.079 ± 0.250

transplant (see lower right-hand entry in Table 3). In this example, the relative sub-hazards appear to fulfill the proportionality assumption and the estimates were on opposite sides of 1. The relative sub-hazard estimates with 95% confidence intervals for ESRD and transplant are shown in Fig. 2, panels c and d, respectively. All estimates were consistent with constant relative sub-hazards of 0.5 for ESRD and 1.7 for transplant when comparing households with an income above \$36,000 to those below.

The top part of Table 4 provides the number of observed events among the 501 and the 77 children without and with nephrotic proteinuria at baseline, respectively. The non-parametric cumulative incidences for ESRD and transplant are shown in Fig. 3, panels a and b, respectively. Traditional Fine and Gray regression yielded a constant relative sub-hazard estimate for ESRD, comparing those with nephrotic proteinuria to those without, of 11.1 (= $\exp(2.408)$, 95% CI: 6.9, 18.0), as shown in Table 4. Allowing for linear departure from proportionality (i.e., modeling the logarithm of the relative sub-hazards as a linear function of time), the relative sub-hazard for ESRD showed a steep and strongly significant downward trend (see last row of Table 4) moving from 37 (= $\exp(3.609)$) to nearly 1 in the four and a half years after baseline (see green line in Fig. 3, panel c). In turn, the constant relative sub-hazard estimate for transplant was 3.3 (= $\exp(1.206)$ in Table 4) with a 95% CI from 1.8 to 6.1, providing a case of the undesirable circumstance in which both estimates of constant relative sub-hazards are above 1 and statistically significant. Allowing for a linear departure from proportionality, the relative sub-hazards for transplant showed a mild and non-significant downward trend as shown in the last row of Table 4. The relative sub-hazard estimates with 95% confidence intervals for ESRD and transplant are shown in Fig. 3, panels c and d, respectively. Although the relative sub-hazards were both above 1 during the first four and a half years, there was a strong indication of a downward trend for ESRD.

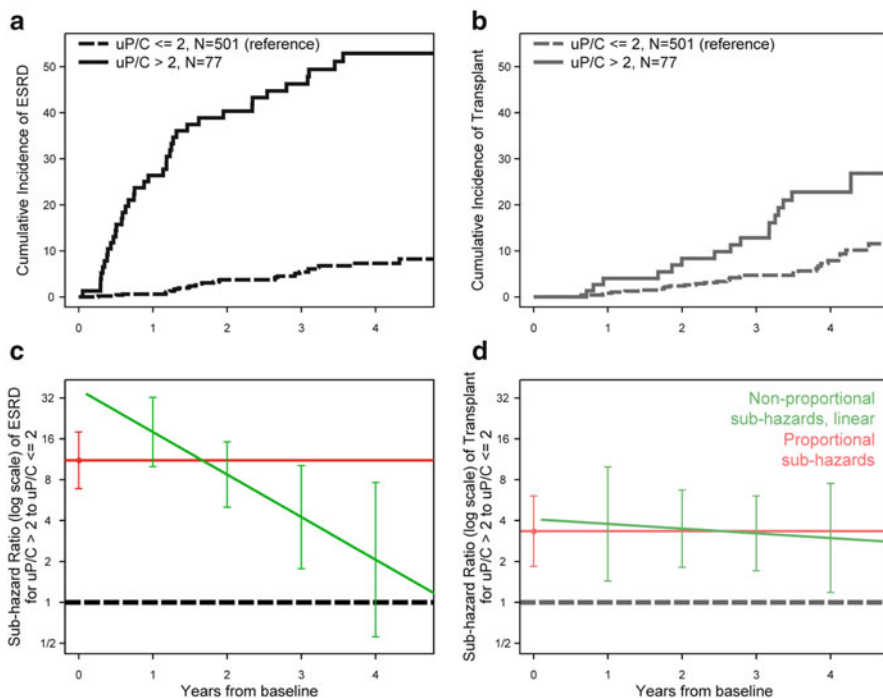


Fig. 3 Effect of nephrotic proteinuria (uP/C > 2) on the competing events of end-stage renal disease (ESRD) and renal transplantation in the CKiD study. Panels **a** and **b** show estimates of cumulative incidences using non-parametric methods. Panels **c** and **d** depict relative sub-hazards under proportionality (*in red*) and linear departure from proportionality (*in green*)

Discussion

In this chapter, we have shown that constant relative sub-hazards for two competing events are tethered by a strong relationship which is independent of the timing of the competing events, are fully determined by the overall frequencies of events, and must be on opposite sides of 1. When violations of proportionality occur, separate analyses under proportionality assumptions for the two competing events often yield results in which the estimates are on the same side of 1 [12–15, 21, 24, 25], and lead to misleading inferences unless explicit limitations about the time frame are included. In addition, we showed that of the sub-hazards and cause-specific hazards for two event types, at most two of the four can be proportional but without restriction on which two. Furthermore, we fully characterized the compatibility of concurrent proportionality of cause-specific hazards and sub-hazards and showed that strong tethering also occurs in those cases, except when the two cause-specific hazards are proportional.

Because proportionality rarely holds in practice, one may choose the inclusion of a time dependency of relative sub-hazards as the default analytical approach [8]; both tabular and graphical depictions of the time trends of relative sub-hazards are straightforward [26]. However, this does not assure correct interpretations in all cases since power may be limited to detect the true change in the relative sub-hazards over time, particularly when follow-up time is relatively short.

The relative sub-hazard has appealing properties as an estimator of the effect of an exposure on an event of interest in the presence of a competing event, and the Fine and Gray weighting procedure is easily implemented using standard software. However, care must be taken that ease of implementation does not lead to a cavalier assumption of proportionality in the sub-hazards. In this chapter, we have demonstrated the coarseness of the summary statistic as well as the inconsistency produced by assuming proportionality of sub-hazards when proportionality does not hold. Proportionality of the cause-specific hazards does not provide protection; rather, as we have illustrated here, it implies violation of the proportionality of the sub-hazards. Our simulated data set with non-proportional sub-hazards described in Table 1 highlights a case in which the relative cause-specific hazards are indeed constant (3.64 and 4.44 for type 1 and 2 events, respectively), but the sub-hazards are extremely non-proportional.

Given the tethered relationships caused by proportionality of sub-hazards, several authors have proposed approaches based on modeling the cumulative incidences directly. Klein [27] has argued that linear additive models are more natural because they intrinsically incorporate the fact that the sum of the cumulative incidences fulfills the requirement of being the cumulative incidence of the composite event. Others have suggested alternative summary measures including time-dependent ratios of the cumulative incidences themselves [17, 28]. We offer, for practical consideration, the incorporation of a simple time dependency in the model and also suggest limiting inferences to a finite interval [29], particularly when limited follow-up is available and nonlinear trends in relative sub-hazards may be hard to detect. Reporting relative sub-hazard results on the same side of 1 during a limited period is an acceptable summary; but, in point of fact, when estimates of relative sub-hazards are on the same side of 1, such a result should immediately alert the analyst to the presence of non-proportional sub-hazards over the full time span.

In cases where there is a substantial degree of right-censoring in the observed data, additional caution should be taken when analyzing competing risks data. First of all, mixtures of fully parametric distributions can yield very imprecise estimates of π and π^* [29, 30]. Second, power to detect departures from proportionality may be limited. A case in point is provided by the income data in our application because the apparent proportionality of the sub-hazards implies that the cause-specific hazards were time-dependent. However, the data limited to the first four and a half years of follow-up did not indicate departures from proportionality of the cause-specific hazards. This provides a case study in the need for restricting the analyses and inferences *only* up to a “time point located inside the support of the observed time variable” [29]. Hence, a more appropriate summary of the analysis would be that the data up to four and a half years from baseline are consistent

with the sub-hazards being proportional in that annual income greater than \$36,000 halves the risk of ESRD and increases the likelihood of transplantation by two thirds. In contrast, even in the case of heavy censoring, strong trends of relative sub-hazards can be detectable, as illustrated by the effect of proteinuria on ESRD in our application.

Although it is attractive to reduce inferences to one number corresponding to proportionality of measures of disease frequency, biological processes are often much more complex, and we have shown that in the setting of competing risks, the assumptions of proportionalities induce tethering of the relative hazards. If summaries based only on a single measure are desired, it is safer to rely on proportional cause-specific hazards as they are not subjected to tethering relationships, as argued and implemented by Wada et al. [16]. Another approach is to frame summaries as estimating least false parameters [31] or time-averaged effects [32].

It should be noted that when sub-hazards are truly proportional, simulation studies (data not shown) indicated that results from methods incorporating the tethering of the sub-hazards and those from the traditional Fine and Gray method yielded unbiased and equally efficient estimators. This is not a surprising result as the relative sub-hazards under proportionality are fully determined by the frequency of the two types of events and not by their timing.

In this chapter, we have restricted our discussion to the case of two competing events (i.e., $K = 2$). For the case of $K > 2$, the assumption of proportional sub-hazards will result in the relative sub-hazard of the type 1 event being unbounded, $K - 2$ of them having upper bounds determined by the overall frequencies, and the last one being tethered in a similar manner as the case of $K = 2$ (i.e., $a_K = \log(\sum_{i=1}^{K-1} \pi_i^*) / \log(\sum_{i=1}^{K-1} \pi_i)$).

In summary, although the sub-hazards approach has the appeal that covariate effects on the sub-hazard functions are consistent with the effects on the corresponding cumulative incidence functions, care should be taken to assure that violations of the proportionality assumption do not result in misleading or incorrect conclusions. Because proportionality rarely holds in practice, the default analytical approach should be to allow for the relative hazards to depend on time, though statistical power is limited in the case of large numbers of event-free observations. Restricting inferences to a finite period may also provide protection from reporting theoretically inconsistent results.

Acknowledgements This work and the Chronic Kidney Disease in Children Study are supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) at the National Institutes of Health (NIH), funded in collaboration the National Institute of Child Health and Human Development (NICHD) and the National Heart, Lung and Blood Institute (NHLBI) of NIH: Grant numbers U01-DK-66116, U01-DK-66143, U01-DK-66174, and U01-DK-82194.

References

1. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* **26**, 2389–2430 (2007). doi:[10.1002/sim.2712](https://doi.org/10.1002/sim.2712)
2. Cox, D.R., Oakes, D.: *The Analysis of Survival Data*, pp. 142–155. Chapman & Hall, New York (1984)
3. Gaynor, J.J., Feurer, E.J., Tan, C., Wu, D.H., Little, C.R., Straus, D.J., Clarkson, B.D., Brennan, M.F.: On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *J. Am. Stat. Assoc.* **88**, 400–409 (1993). doi:[10.1080/01621459.1993.10476289](https://doi.org/10.1080/01621459.1993.10476289)
4. Holt, J.D.: Competing risk analyses with special reference to matched pair experiments. *Biometrika* **65**, 159–165 (1978). doi:[10.1093/biomet/65.1.159](https://doi.org/10.1093/biomet/65.1.159)
5. Kalbfleisch, J.D., Prentice, R.L.: *The Statistical Analysis of Failure Time Data*. Wiley, New York (1980)
6. Larson, M.G.: Covariate analysis of competing-risks data with log-linear models. *Biometrics* **40**, 459–469 (1984)
7. Prentice, R.L., Kalbfleisch, J.D., Peterson Jr., A.V., Flournoy, N., Farewell, V.T., Breslow, N.E.: The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541–554 (1978)
8. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999). doi:[10.1080/01621459.1999.10474144](https://doi.org/10.1080/01621459.1999.10474144)
9. Gray, R.J.: A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **16**, 1141–1154 (1988). doi:[10.1214/aos/1176350951](https://doi.org/10.1214/aos/1176350951)
10. Pepe, M.S.: Inference for events with dependent risks in multiple endpoint studies. *J. Am. Stat. Assoc.* **86**, 770–778 (1991). doi:[10.1080/01621459.1991.10475108](https://doi.org/10.1080/01621459.1991.10475108)
11. Beyersmann, J., Allignol, A., Schumacher, M.: *Competing risks and multistate models with R*, pp. 89–153. Springer, New York (2012)
12. Babiker, A., Darbyshire, J., Pezzotti, P., Porter, K., Rezza, G., Walker, S.A., et al.: Changes over calendar time in the risk of specific first AIDS-defining events following HIV seroconversion, adjusting for competing risks. *Int. J. Epidemiol.* **31**, 951–958 (2002). doi:[10.1093/ije/31.5.951](https://doi.org/10.1093/ije/31.5.951)
13. del Amo, J., Perez-Hoyos, S., Moreno, A., Quintana, M., Ruiz, I., Cisneros, J.M., Ferreros, I., Gonzalez, C., de Olalla, P.G., Perez, R., Hernandez, I.: Trends in AIDS and mortality in HIV-infected subjects with hemophilia from 1985 to 2003: the competing risks for death between AIDS and liver disease. *J. Acquir. Immune Defic. Syndr.* **41**, 624–631 (2006). doi:[10.1097/01.qai.0000194232.85336.dc](https://doi.org/10.1097/01.qai.0000194232.85336.dc)
14. Lim, H.J., Zhang, X., Dyck, R., Osgood, N.: Methods of competing risks analysis of end-stage renal disease and mortality among people with diabetes. *BMC Med. Res. Methodol.* **10**, 97 (2010). doi:[10.1186/1471-2288-10-97](https://doi.org/10.1186/1471-2288-10-97)
15. Pacheco, A.G., Tuboi, S.H., May, S.B., Moreira, L.F.S., Ramadas, L., Nunes, E.P., Merçon, M., Faulhaber, J.C., Harrison, L.H., Schechter, M.: Temporal changes in causes of death among HIV-infected patients in the HAART era in Rio de Janeiro, Brazil. *J. Acquir. Immune Defic. Syndr.* **51**, 624–630 (2009). doi:[10.1097/QAI.0b013e3181a4ecf5](https://doi.org/10.1097/QAI.0b013e3181a4ecf5)
16. Wada, N., Jacobson, L.P., Cohen, M., French, A.L., Phair, J., Muñoz, A.: Cause-specific life expectancies after 35 years of age for human immunodeficiency syndrome-infected and human immunodeficiency syndrome-negative individuals followed simultaneously in long-term cohort studies: 1984–2008. *Am. J. Epidemiol.* **15**, 116–125 (2013). doi:[10.1093/aje/kws321](https://doi.org/10.1093/aje/kws321)
17. Checkley, W., Brower, R.G., Muñoz, A.: Inference for mutually exclusive competing events through a mixture of generalized gamma distributions. *Epidemiology* **21**, 557–565 (2010). doi:[10.1097/EDE.0b013e3181e090ed](https://doi.org/10.1097/EDE.0b013e3181e090ed)
18. Cole, S.R., Li, R., Anastos, K., Detels, R., Young, M., Chmiel, J.S., Muñoz, A.: Accounting for leadtime in cohort studies: evaluating when to initiate HIV therapies. *Stat. Med.* **23**, 3351–3363 (2004). doi:[10.1002/sim.1579](https://doi.org/10.1002/sim.1579)
19. Larson, M.G., Dinse, G.E.: A mixture model for the regression analysis of competing risks data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **34**, 201–211 (1985). doi:[10.2307/2347464](https://doi.org/10.2307/2347464)

20. Beyersmann, J., Latouche, A., Buchholz, A., Schumacher, M.: Simulating competing risks data in survival analysis. *Stat. Med.* **28**, 956–971 (2009). doi:[10.1002/sim.3516](https://doi.org/10.1002/sim.3516)
21. Latouche, A., Boisson, V., Chevret, S., Porcher, R.: Misspecified regression model for the sub-distribution hazard of a competing risk. *Stat. Med.* **26**, 965–974 (2007). doi:[10.1002/sim.2600](https://doi.org/10.1002/sim.2600)
22. Zhang, X., Zhang, M.J., Fine, J.: A proportional hazard regression model for the subdistribution with right-censored and left-truncated competing risks data. *Stat. Med.* **30**, 1933–1951 (2011). doi:[10.1002/sim.4264](https://doi.org/10.1002/sim.4264)
23. Furth, S.L., Cole, S.R., Moxey-Mims, M., Kaskel, F., Mak, R., Schwartz, G., Wong, C., Muñoz, A., Warady, B.A.: Design and methods of the Chronic Kidney Disease in Children (CKiD) prospective cohort study clinical. *J. Am. Soc. Nephrol.* **1**, 1006–1015 (2006). doi:[10.2215/CJN.01941205](https://doi.org/10.2215/CJN.01941205)
24. Geskus, R.B.: Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics* **67**, 39–49 (2011). doi:[10.1111/j.1541-0420.2010.01420.x](https://doi.org/10.1111/j.1541-0420.2010.01420.x)
25. Jeong, J.-H., Fine, J.P.: Parametric regression on cumulative incidence function. *Biostatistics* **8**, 184–196 (2007). doi:[10.1093/biostatistics/kxj040](https://doi.org/10.1093/biostatistics/kxj040)
26. del Amo, J., Jarring, I., May, M., Dabis, F., Crane, H., Podzamczar, D., et al.: Influence of geographical origin and ethnicity on mortality in patients on antiretroviral therapy in Canada, Europe and the United States. *Clin. Infect. Dis.* **54**, 1800–1809 (2013). doi:[10.1093/cid/cit111](https://doi.org/10.1093/cid/cit111)
27. Klein, J.P.: Modelling competing risks in cancer studies. *Stat. Med.* **25**, 1015–1034 (2006). doi:[10.1002/sim.2246](https://doi.org/10.1002/sim.2246)
28. Zhang, M.J., Fine, J.: Summarizing differences in cumulative incidence functions. *Stat. Med.* **27**, 4939–4949 (2008). doi:[10.1002/sim.3339](https://doi.org/10.1002/sim.3339)
29. Chang, W.-H., Wang, W.: Regression analysis for cumulative incidence probability under competing risks. *Statistica Sinica* **19**, 391–408 (2009)
30. Fine, J.P.: Analysing competing risks data with transformation models. *J. R. Stat. Soc. Ser. B* **61**, 817–830 (1999). doi:[10.1111/1467-9868.00204](https://doi.org/10.1111/1467-9868.00204)
31. Grambauer, N., Schumacher, M., Beyersmann, J.: Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Stat. Med.* **29**, 875–884 (2010). doi:[10.1002/sim.3786](https://doi.org/10.1002/sim.3786)
32. Lau, B., Cole, S.R., Gange, S.J.: Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry. *Stat. Med.* **30**, 654–665 (2011). doi:[10.1002/sim.4123](https://doi.org/10.1002/sim.4123) [correction in *Statistics in Medicine* 2012;31:1777–8. doi: [10.1002/sim.4468](https://doi.org/10.1002/sim.4468)]

Semiparametric Inference on the Absolute Risk Reduction and the Restricted Mean Survival Difference

Song Yang

Abstract For time-to-event data, when the hazards may be non-proportional, in addition to the hazard ratio, the absolute risk reduction and the restricted mean survival difference can be used to describe the time-dependent treatment effect. The absolute risk reduction measures the direct impact of the treatment on event rate or survival, and the restricted mean survival difference provides a way to evaluate the cumulative treatment effect. However, in the literature, available methods are limited for flexibly estimating these measures and making inference on them. In this article, point estimates, pointwise confidence intervals and simultaneous confidence bands of the absolute risk reduction and the restricted mean survival difference are established under a semiparametric model that can be used in a sufficiently wide range of applications. These methods are motivated by and illustrated for data from the Women’s Health Initiative estrogen plus progestin clinical trial.

Introduction

Comparison of two groups of survival data has wide applications in life testing, reliability studies, and clinical trials. Often the two sample proportional hazards model of Cox [4] is assumed and a single value of the hazard ratio is used to describe the group difference. When the hazard ratio is possibly time-dependent, a conventional approach is to give a hazard ratio estimate over each of a few time periods, by fitting a piece-wise proportional hazards model. Alternatively, a “defined” time-varying covariate can be used in a Cox regression model, resulting

S. Yang (✉)

Office of Biostatistics Research, National Heart, Lung, and Blood Institute,
6701 Rockledge Dr. MSC 7913, Bethesda, MD 20892, USA
e-mail: yangso@nhlbi.nih.gov

in a parametric form for the hazard ratio function (e.g. [6], Chap. 6). With these approaches, it may not be easy to pre-specify the partition of the time axis or the parametric form of the hazard ratio function.

In Yang and Prentice [22], a short-term and long-term hazards model was proposed. Assume absolutely continuous failure times and label the two groups control and treatment, with hazard functions $\lambda_C(t)$ and $\lambda_T(t)$, respectively. Then the short-term and long-term hazards model postulates that

$$\lambda_T(t) = \frac{1}{e^{-\beta_2} + (e^{-\beta_1} - e^{-\beta_2})S_C(t)} \lambda_C(t), \quad t < \tau_0, \quad (1)$$

where β_1 , β_2 are scalar parameters, S_C is the survivor function of the control group, and

$$\tau_0 = \sup\{x : \int_0^x \lambda_C(t) dt < \infty\}. \quad (2)$$

Under this model, $\lim_{t \downarrow 0} \lambda_T(t)/\lambda_C(t) = e^{\beta_1}$, $\lim_{t \uparrow \tau_0} \lambda_T(t)/\lambda_C(t) = e^{\beta_2}$. Thus various patterns of the hazard ratio can be realized, including proportional hazards, no initial effect, disappearing effect, and crossing hazards. In particular, model (1) includes the proportional hazards model and the proportional odds model as special cases. There is no need to specify a partition of the time axis or a parametric form of the hazard ratio function. For this model, Yang and Prentice [22] proposed a pseudo-likelihood method for estimating the parameters, and Yang and Prentice [23] studied inference procedures on the hazard ratio function. Extension of model (1) to the regression setting was also studied for current status data in Tong et al. [20].

In situations with non-proportional hazards, the hazard ratio is useful for assessing temporal trend of the treatment effect, but it may not directly translate to the survival experience. For example, the hazard ratio may be less than 1 in a region where there is no improvement in the survival probability. Also, there is no simple nonparametric estimator as a reference when comparing different estimators of the hazard ratio function. In the Women's Health Initiative estrogen plus progestin clinical trial [10, 21], the hazard ratio function was decidedly non-proportional for the outcomes of coronary heart disease, venous thromboembolism, and stroke. While the estimated hazard ratios from Prentice et al. [16] and Yang and Prentice [23] are in good agreement with each other for the outcomes of coronary heart disease and venous thromboembolism, they indicate somewhat different hazard ratio shapes for stroke. Under the piece-wise Cox model with the partition of 0–2, 2–5, and 5+ years (the partition used in [16]), the hazard ratio has an upside down U-shape. On the other hand, under the piece-wise Cox model using the partition of 0–3, 3–6, and 6+ years (a plausible partition since the maximum follow-up time was almost 9 years), the hazard ratio has a U-shape. The result from Yang and Prentice [23] shows a hazard ratio that is slightly decreasing over time. Thus for stroke, the temporal trend of the hazard ratio is portrayed somewhat differently under these models. These hazard ratio estimates are displayed in Fig. 1.

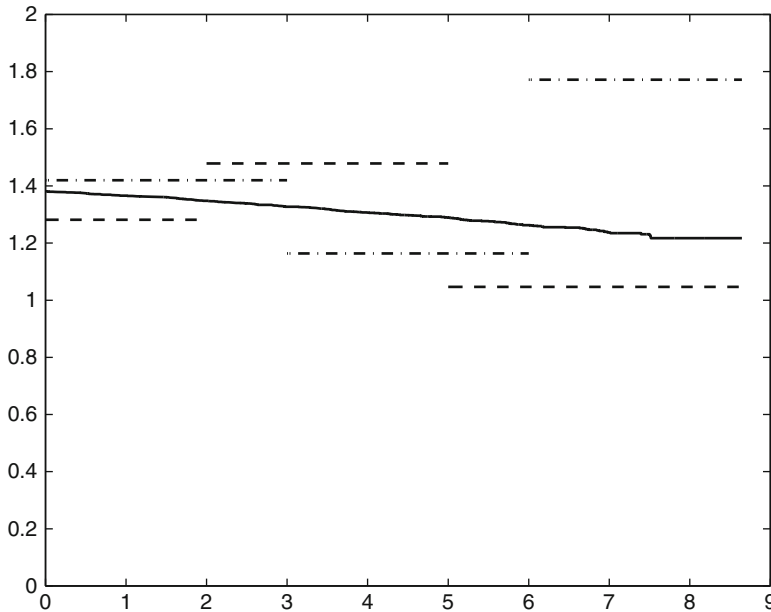


Fig. 1 Estimated hazard ratio for the WHI clinical trial stroke data: *Solid line*—Model (1); *Dashed line*—Piece-wise Cox model with cut points at 2 and 5 years; *Dash-dotted lines*—Piece-wise Cox model with cut points at 3 and 6 years

To help compare these different results, one can consider the absolute risk reduction by the treatment. Figure 2 displays various estimators of the absolute risk reduction. From Fig. 2, several observations can be made. Between the two piece-wise Cox models with different partitions, the partition with cut points 2 and 5 years results in a better agreement with the Kaplan-Meier [7] based estimator for the early to middle portion of the data range. The other partition results in a better agreement with the Kaplan-Meier based estimator for the range beyond 6 years. The estimator based on model (1) is a good compromise between the results from the two partitions. One more comparison of these models can be made through the restricted mean survival difference, displayed in Fig. 3. The different estimators are closer to each other and are also smoother. For the piece-wise Cox models, the partition with cut points 2 and 5 years results in an estimator that is closer to the Kaplan-Meier estimator for early part of the data range, but has a more noticeable deviation near the end. Again the estimator based on model (1) results in a good compromise between the two partitions.

In this article, we consider making semiparametric inference on the absolute risk reduction and the restricted mean survival difference for two sample time-to-event data, under model (1). The absolute risk reduction is directly related to the survival experience, and is a commonly used measure in epidemiological studies. The restricted mean survival time has been used as a summary measure in various works when the hazards are non-proportional. The restricted mean survival time

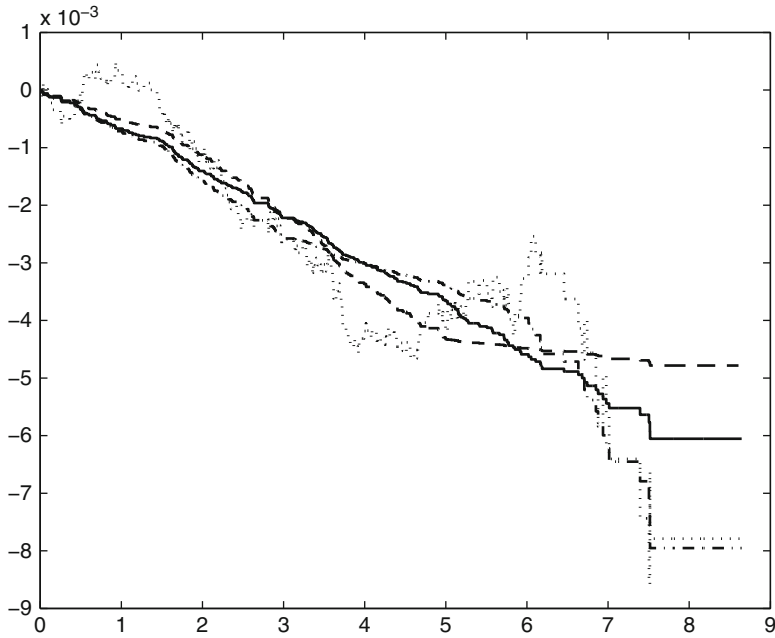


Fig. 2 Estimated absolute risk reduction for the WHI clinical trial stroke data: *Solid line*—Model (1); *Dotted line*: Kaplan-Meier; *Dashed line*—Piece-wise Cox model with cut points at 2 and 5 years; *Dash-dotted lines*—Piece-wise Cox model with cut points at 3 and 6 years

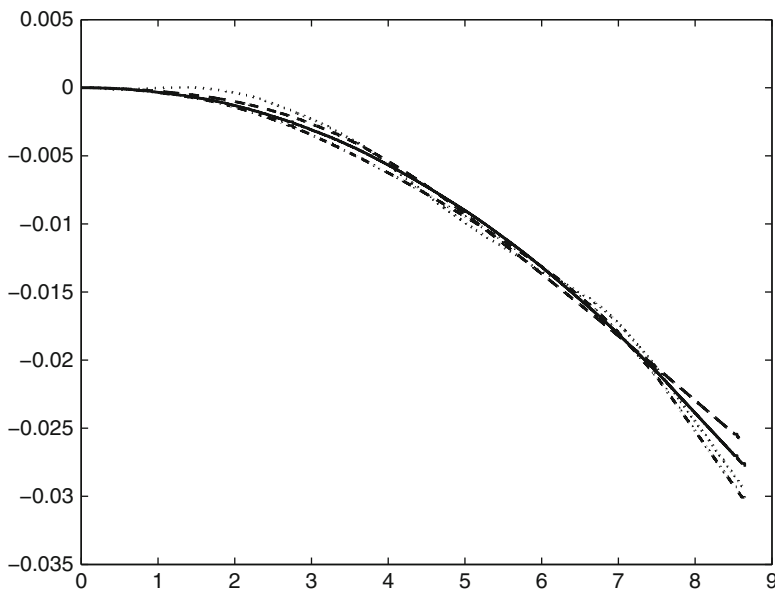


Fig. 3 Estimated mean restricted survival difference for the WHI clinical trial stroke data: *Solid line*—Model (1); *Dotted line*: Kaplan-Meier; *Dashed line*—Piece-wise Cox model with cut points at 2 and 5 years; *Dash-dotted lines*—Piece-wise Cox model with cut points at 3 and 6 years

up to t can be thought of as the ‘ t -year life expectancy’, and it approaches the unrestricted mean survival time as t approaches infinity. In clinical trials where the trial often ends after a pre-specified follow-up period, the restricted mean survival time is a more appropriate measure than the unrestricted mean survival time. In the subsequent development, the estimates, point-wise confidence intervals and simultaneous confidence bands of the absolute risk reduction and the restricted mean survival difference will be established under model (1). Such semiparametric inference procedures are sufficiently flexible for many applications, due to the various properties of model (1) mentioned before. These confidence intervals and confidence bands can be used to capture and graphically present the treatment effect. We illustrate these visual tools through applications to the clinical trial data from the Women’s Health Initiative.

There have been various works in the literature that are related to the problems considered here. Recently Schaubel and Wei [18] considered the restricted mean survival difference and other measures under dependent censoring. Royston and Parmar [17] considered inference on the restricted mean survival time by extending standard survival models to accommodate a wide range of baseline distributions. In both works, point-wise confidence intervals are constructed. In earlier works, Dabrowska et al. [5] introduced a relative change function defined in terms of cumulative hazards and found simultaneous bands for this function under the assumption of proportional hazards. Parzen et al. [13] constructed nonparametric simultaneous confidence bands for the survival probability difference. Cheng et al. [3] proposed pointwise and simultaneous confidence interval procedures for the survival probability under semiparametric transformation models. Zucker [24] and Chen and Tsiatis [2] compared the restricted mean survival time between two groups using Cox proportional hazards models. McKeague and Zhao [11] proposed simultaneous confidence bands for ratios of survival functions via the empirical likelihood method.

The article is organized as follows. In section “The Estimators and Their Asymptotic Properties” the short-term and long-term hazard ratio model and the parameter estimator are described. Pointwise confidence intervals are established for the absolute risk reduction and the restricted mean survival difference under the model. In section “Simultaneous Confidence Bands”, simultaneous confidence bands are developed for the absolute risk reduction and the restricted mean survival difference. Simulation results are presented in section “Simulation Studies”. Application to the stroke data from the Women’s Health Initiative trial is given in section “Application”. Some discussions are given in section “Discussion”.

The Estimators and Their Asymptotic Properties

Let T_1, \dots, T_n be the pooled lifetimes of the two groups, with T_1, \dots, T_{n_1} , $n_1 < n$, constituting the control group having the survivor function S_C . Let C_1, \dots, C_n be the censoring variables, and $Z_i = I(i > n_1)$, $i = 1, \dots, n$, where $I(\cdot)$ is the

indicator function. The available data consist of the independent triplets (X_i, δ_i, Z_i) , $i = 1, \dots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. We assume that T_i , C_i are independent given Z_i . The censoring variables (C_i 's) need not be identically distributed, and in particular the two groups may have different censoring patterns. For $t < \tau_0$ with τ_0 defined in (2), let $R(t)$ be the odds function $1/S_C(t) - 1$ of the control group. The model of Yang and Prentice [22] can be expressed as

$$\lambda_i(t) = \frac{1}{e^{-\beta_1 Z_i} + e^{-\beta_2 Z_i} R(t)} \frac{dR(t)}{dt}, \quad i = 1, \dots, n, t < \tau_0,$$

where $\lambda_i(t)$ is the hazard function for T_i given Z_i .

Let S_T be the survivor function of the treatment group. Then the absolute risk reduction is

$$\Phi(t) = S_T(t) - S_C(t).$$

This function is positive if the treatment reduces the event rate and negative if the treatment increases the event rate. Under model (1), $\Phi(t)$ depends on the parameter $\beta = (\beta_1, \beta_2)^T$ and the baseline function $R(t)$, where “ T ” denotes transpose. Yang and Prentice [22] studied a pseudo likelihood estimator $\hat{\beta}$ of β which we describe below.

Let $\tau < \tau_0$ be such that

$$\lim_n \sum_{i=1}^n I(X_i \geq \tau) > 0, \quad (3)$$

with probability 1. For $t \leq \tau$, define

$$\hat{P}(t; \mathbf{b}) = \prod_{s \leq t} \left(1 - \frac{\sum_{i=1}^n \delta_i e^{-b_2 Z_i} I(X_i = s)}{\sum_{i=1}^n I(X_i \geq s)} \right),$$

$$\hat{R}(t; \mathbf{b}) = \frac{1}{\hat{P}(t; \mathbf{b})} \int_0^t \frac{\hat{P}_-(s; \mathbf{b})}{\sum_{i=1}^n I(X_i \geq s)} d\left(\sum_{i=1}^n \delta_i e^{-b_1 Z_i} I(X_i \leq s) \right),$$

where $\hat{P}_-(s; \mathbf{b})$ denotes the left continuous (in s) version of $\hat{P}(s; \mathbf{b})$.

Let $L(\beta, R)$ be the likelihood function of β under model (1) when the function $R(t)$ is known, with the corresponding score vector $S(\beta, R) = \partial \ln L(\beta, R) / \partial \beta$. Define $Q(\mathbf{b}) = S(\mathbf{b}, R)|_{R(t) = \hat{R}_n(t; \mathbf{b})}$. Then the pseudo maximum likelihood estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$ of β is the zero of $Q(\mathbf{b})$.

Once $\hat{\beta}$ is obtained, $R(t)$ can be estimated by $\hat{R}(t; \hat{\beta})$. Thus under model (1), the absolute risk reduction $\Phi(t)$ can be estimated by

$$\hat{\Phi}(t) = \{1 + e^{-\hat{\beta}_2 + \hat{\beta}_1 \hat{R}(t; \hat{\beta})}\}^{-e^{\hat{\beta}_2}} - \frac{1}{1 + \hat{R}(t; \hat{\beta})}. \quad (4)$$

In Appendix 1, we show that $\hat{\Phi}(t)$ is strongly consistent for $\Phi(t)$ under model (1).

To study the distributional properties of $\hat{\Phi}(t)$, let

$$U_n(t) = \sqrt{n}(\hat{\Phi}(t) - \Phi(t)), \quad t \leq \tau.$$

Let $\xi_0(t) = 1 + R(t)$, $\xi(t) = e^{-\beta_1} + e^{-\beta_2}R(t)$, $\hat{\xi}_0(t) = 1 + \hat{R}(t; \beta)$, $\hat{\xi}(t) = e^{-\beta_1} + e^{-\beta_2}\hat{R}(t; \beta)$, and define

$$K_1(t) = \sum_{i \leq n_1} I(X_i \geq t), \quad K_2(t) = \sum_{i > n_1} I(X_i \geq t),$$

$$A(t) = \frac{1}{\hat{\xi}(t)}(e^{-\beta_1}, e^{-\beta_2}\hat{R}(t; \beta))^T,$$

$$B(t) = \int_t^\tau \frac{A(s)K_1(s)K_2(s)}{\hat{\xi}(s)\hat{P}(s; \beta)} \left(\frac{e^{-\beta_2}}{\xi(s)} - \frac{1}{\xi_0(s)} \right) dR(s).$$

In Appendix 1, it will be shown that, with probability 1,

$$Q(\beta) = \sum_{i \leq n_1} \int_0^\tau \{\mu_1(t) + o(1)\} dM_i(t) + \sum_{i > n_1} \int_0^\tau \{\mu_2(t) + o(1)\} dM_i(t), \quad (5)$$

uniformly in $t \leq \tau$ and $i \leq n$, where

$$\mu_1(t) = -\frac{\hat{\xi}_0(t)A(t)K_2(t)}{\hat{\xi}(t)K(t)} + \frac{\hat{\xi}_0(t)\hat{P}_-(t; \beta)}{K} B(t),$$

$$\mu_2(t) = A(t) \frac{K_1(t)}{K(t)} + \frac{\hat{\xi}(t)\hat{P}_-(t; \beta)}{K(t)} B(t), \quad (6)$$

$$M_i(t) = \delta_i I(X_i \leq t) - \int_0^t I(X_i \geq s) \frac{dR(s)}{e^{-\beta_1}Z_i + e^{-\beta_2}Z_i R(s)}, \quad i = 1, \dots, n.$$

By Lemma A3 of Yang and Prentice [22],

$$\sqrt{n}(\hat{R}(t; \beta) - R(t)) = \frac{1}{\sqrt{n}\hat{P}(t; \beta)} \left(\sum_{i \leq n_1} \int_0^t v_1 dM_i + \sum_{i > n_1} \int_0^t v_2 dM_i \right) \quad (7)$$

where

$$v_1(t) = \frac{n\hat{\xi}_0(t)\hat{P}_-(t; \beta)}{K(t)}, \quad v_2(t) = \frac{n\hat{\xi}(t)\hat{P}_-(t; \beta)}{K(t)}.$$

Let Λ_T be the cumulative hazard function of the treatment group and define

$$C(t) = \frac{1}{\hat{P}(t; \beta)} \left(\frac{1}{\xi_0^2(t)} - \frac{S_T(t)}{\xi(t)} \right), \quad \Omega = \left\{ -\frac{1}{n} \frac{\partial Q(\beta)}{\partial \beta} \right\}^{-1},$$

$$D(t) = C(t) \hat{P}(t; \beta) \frac{\partial \hat{R}(t; \beta)}{\partial \beta} - S_T(t) \left(\frac{R(t)}{\xi(t)}, \Lambda_T(t) - \frac{R(t)}{\xi(t)} \right)^T.$$

For $t \leq \tau$, define the process

$$\begin{aligned} \tilde{U}_n(t) &= \frac{D^T(t) \Omega}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^t \mu_1 dM_i + \sum_{i > n_1} \int_0^t \mu_2 dM_i \right) \\ &\quad + \frac{C(t)}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^t v_1 dM_i + \sum_{i > n_1} \int_0^t v_2 dM_i \right). \end{aligned} \quad (8)$$

With the representations for $Q(\beta)$ and $\sqrt{n}(\hat{R}(t; \beta) - R(t))$, in Appendix 2 it will be shown that U_n is asymptotically equivalent to \tilde{U}_n which converges weakly to a zero-mean Gaussian process U^* . The weak convergence of U_n thus follows. The limiting covariance function $\sigma_\Phi(s, t)$ of U^* involves the derivative vector $\partial \hat{R}(t; \beta) / \partial \beta$ and the derivative matrix in Ω . Although analytic forms of these derivatives are available, they are quite complicated and cumbersome. Here we approximate them by numerical derivatives. For the functions $C(t)$, $D(t)$, $\mu_1(t)$, $\mu_2(t)$, $v_1(t)$, $v_2(t)$, define corresponding $\hat{C}(t)$, $\hat{D}(t)$, ... by replacing β with $\hat{\beta}$, $R(t)$ with $\hat{R}(t; \hat{\beta})$, $S_T(t)$ and $\Lambda(t)$ with model based estimators, and $\partial \hat{R}(t; \beta) / \partial \beta$ with the numerical derivatives. Similarly, let $\hat{\Omega}$ be the numerical approximation of Ω . Simulation studies show that the results are fairly stable with respect to the choice of the jump size in the numerical derivatives, and that the choice of $n^{-1/2}$ works well. With these approximations, the covariation process $\sigma_\Phi(s, t)$, $s \leq t \leq \tau$, can be estimated by

$$\begin{aligned} \hat{\sigma}_\Phi(s, t) &= \hat{D}^T(s) \hat{\Omega} \left(\int_0^\tau \frac{\hat{\mu}_1(w) \hat{\mu}_1^T(w) K_1(w) d\hat{R}(w; \hat{\beta})}{n(1 + \hat{R}(w; \hat{\beta}))} \right. \\ &\quad + \int_0^\tau \frac{\hat{\mu}_2(w) \hat{\mu}_2^T(w) K_2(w) d\hat{R}(w; \hat{\beta})}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \Big) \hat{\Omega}^T \hat{D}(t) \\ &\quad + \hat{C}(s) \hat{C}(t) \left(\int_0^s \frac{\hat{v}_1^2(w) K_1(w) d\hat{R}(w; \hat{\beta})}{n(1 + \hat{R}(w; \hat{\beta}))} \right. \\ &\quad + \int_0^s \frac{\hat{v}_2^2(w) K_2(w) d\hat{R}(w; \hat{\beta})}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \Big) \\ &\quad + \hat{C}(t) \hat{D}^T(s) \hat{\Omega} \left(\int_0^t \frac{\hat{\mu}_1(w) \hat{v}_1(w) K_1(w) d\hat{R}(w; \hat{\beta})}{n(1 + \hat{R}(w; \hat{\beta}))} \right. \end{aligned}$$

$$\begin{aligned}
& + \int_0^t \frac{\hat{\mu}_2(w) \hat{v}_2(w) K_2(w) d\hat{R}(w; \hat{\beta})}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \\
& + \hat{C}(s) \hat{D}^T(t) \hat{\Omega} \left(\int_0^s \frac{\hat{\mu}_1(w) \hat{v}_1(w) K_1(w) d\hat{R}(w; \hat{\beta})}{n(1 + \hat{R}(w; \hat{\beta}))} \right. \\
& \left. + \int_0^s \frac{\hat{\mu}_2(w) \hat{v}_2(w) K_2(w) d\hat{R}(w; \hat{\beta})}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \right). \tag{9}
\end{aligned}$$

For a fixed $t_0 \leq \tau$, from the above results, an asymptotic $100(1 - \alpha)\%$ confidence interval for $\hat{\Phi}(t_0)$ is $\hat{\Phi}(t_0) \mp z_{\alpha/2} \sqrt{\hat{\sigma}_{\Phi}(t_0, t_0)/n}$, where $z_{\alpha/2}$ is the $100(1 - \alpha/2)\%$ percentile of the standard normal distribution.

Now let us look at the restricted mean survival difference

$$\Psi(t) = \int_0^t S_T(s) ds - \int_0^t S_C(s) ds.$$

Under model (1), $\Psi(t)$ is estimated by

$$\hat{\Psi}(t) = \int_0^t \hat{\Phi}(s) ds,$$

for $\hat{\Phi}(t)$ in (4). In Appendix 1, it will be shown that $\hat{\Psi}(t)$ is a strongly consistent estimator for $\Psi(t)$.

For $t \leq \tau$ define

$$V_n(t) = \sqrt{n}(\hat{\Psi}(t) - \Psi(t)),$$

and

$$\tilde{V}_n(t) = \int_0^t \tilde{U}_n(s) ds,$$

for \tilde{U}_n in (8). Exchanging the order of integration yields

$$\begin{aligned}
\tilde{V}_n(t) &= \frac{\int_0^t D^T(x) dx \Omega}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^\tau \mu_1(w) dM_i(w) + \sum_{i > n_1} \int_0^\tau \mu_2(w) dM_i(w) \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{i \leq n_1} \int_0^t v_1(w) \int_w^t C(x) dx dM_i(w) \\
&+ \frac{1}{\sqrt{n}} \sum_{i > n_1} \int_0^t v_2(w) \int_w^t C(x) dx dM_i(w). \tag{10}
\end{aligned}$$

In Appendix 2, it will be shown that the process $V_n(t)$ is asymptotically equivalent to the process $\tilde{V}_n(t)$ which converges weakly to the zero-mean Gaussian process

$V^*(t) = \int_0^t U^*(s)ds$. Thus $V_n(t)$ also converges weakly to $V^*(t)$. The covariation process $\sigma_\Psi(s, t)$ of $V^*(t)$ can be consistently estimated by

$$\begin{aligned}
\hat{\sigma}_\Psi(s, t) = & \int_0^s \hat{D}^T(x) dx \hat{\Omega} \left(\int_0^\tau \frac{\hat{\mu}_1(w) \hat{\mu}_1^T(w) K_1(w) d\hat{R}(w; \hat{\beta})}{n(1 + \hat{R}(w; \hat{\beta}))} \right. \\
& + \int_0^\tau \frac{\hat{\mu}_2(w) \hat{\mu}_2^T(w) K_2(w) d\hat{R}(w; \hat{\beta})}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \left. \right) \hat{\Omega}^T \int_0^t \hat{D}^T(x) dx \\
& + \int_0^s \frac{\hat{v}_1^2(w) K_1(w)}{n(1 + \hat{R}(w; \hat{\beta}))} \left(\int_w^s C(x) dx \right)^2 d\hat{R}(w; \hat{\beta}) \\
& + \int_0^s \frac{\hat{v}_2(w)^2 K_2(w)}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \left(\int_w^s C(x) dx \right)^2 d\hat{R}(w; \hat{\beta}) \\
& + \int_0^s \hat{D}^T(x) dx \hat{\Omega} \int_0^t \frac{\hat{\mu}_1(w) \hat{v}_1(w) K_1(w)}{n(1 + \hat{R}(w; \hat{\beta}))} \int_w^t C(x) dx d\hat{R}(w, \hat{\beta}) \\
& + \int_0^s \hat{D}^T(x) dx \hat{\Omega} \int_0^t \frac{\hat{\mu}_2(w) \hat{v}_2(w) K_2(w)}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \int_w^t C(x) dx d\hat{R}(w, \hat{\beta}) \\
& + \int_0^t \hat{D}^T(x) dx \hat{\Omega} \int_0^s \frac{\hat{\mu}_1(w) \hat{v}_1(w) K_1(w)}{n(1 + \hat{R}(w; \hat{\beta}))} \int_w^s C(x) dx d\hat{R}(w, \hat{\beta}) \\
& + \int_0^t \hat{D}^T(x) dx \hat{\Omega} \int_0^s \frac{\hat{\mu}_2(w) \hat{v}_2(w) K_2(w)}{n(e^{-\hat{\beta}_1} + e^{-\hat{\beta}_2} \hat{R}(w; \hat{\beta}))} \int_w^s C(x) dx d\hat{R}(w, \hat{\beta}) \quad (11)
\end{aligned}$$

From these results, an asymptotic $100(1 - \alpha)\%$ confidence interval for $\Psi(t_0)$ can be obtained as $\hat{\Psi}(t_0) \mp z_{\alpha/2} \sqrt{\hat{\sigma}_\Psi(t_0, t_0)/n}$.

Simultaneous Confidence Bands

To make simultaneous inference on $\Phi(t)$ over a time interval $I = [a, b] \subset [0, \tau]$, let $w_n(t)$ be a data-dependent function that converges in probability to a bounded function $w^*(t) > 0$, uniformly in t over I . Then U_n/w_n converges weakly U^*/w^* . If c_α is the upper α th percentile of $\sup_{t \in I} |U^*/w^*|$, an asymptotic $100(1 - \alpha)\%$ simultaneous confidence band for $\Phi(t)$, $t \in I$, can be obtained as

$$\left(\hat{\Phi}(t) - \frac{c_\alpha w_n(t)}{\sqrt{n}}, \hat{\Phi}(t) + \frac{c_\alpha w_n(t)}{\sqrt{n}} \right).$$

It is difficult to obtain c_α analytically. One obvious alternative would be the bootstrapping method. However, it is very time-consuming. More discussion on this will be given later on the application to data from the Women's Health Initiative estrogen plus progestin clinical trial. Here a normal resampling approximation will be used. Lin et al. [8] used the normal resampling approximation to simulate

the asymptotic distribution of sums of martingale residuals for checking the Cox regression model. This approach reduces computing time significantly, and has been used in many works, including Lin et al. [9], Cheng et al. [3], Tian et al. [19], and Peng and Huang [14].

For $t \leq \tau$, let $N_i(t) = \delta_i I(X_i \leq t)$, $i = 1, \dots, n$, and define the process

$$\begin{aligned} \hat{U}_n(t) &= \frac{\hat{D}^T(t)\hat{\Omega}}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^\tau \hat{\mu}_1 d(\epsilon_i N_i) + \sum_{i > n_1} \int_0^\tau \hat{\mu}_2 d(\epsilon_i N_i) \right) \\ &\quad + \frac{\hat{C}(t)}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^t \hat{\nu}_1 d(\epsilon_i N_i) + \sum_{i > n_1} \int_0^t \hat{\nu}_2 d(\epsilon_i N_i) \right) \\ &= \frac{\hat{D}^T(t)\hat{\Omega}}{\sqrt{n}} \left(\sum_{i \leq n_1} \epsilon_i \delta_i \hat{\mu}_1(X_i) I(X_i \leq \tau) + \sum_{i > n_1} \epsilon_i \delta_i \hat{\mu}_2(X_i) I(X_i \leq \tau) \right) \\ &\quad + \frac{\hat{C}(t)}{\sqrt{n}} \left(\sum_{i \leq n_1} \epsilon_i \delta_i \hat{\nu}_1(X_i) I(X_i \leq t) + \sum_{i > n_1} \epsilon_i \delta_i \hat{\nu}_2(X_i) I(X_i \leq t) \right), \quad (12) \end{aligned}$$

where ϵ_i , $i = 1, \dots, n$, are independent standard normal variables that are also independent of the data.

Conditional on (X_i, δ_i, Z_i) , $i = 1, \dots, n$, \hat{U}_n is a sum of n independent variables at each time point. In Appendix 2, it will be shown that \hat{U}_n given the data converges weakly to U^* . It follows that $\sup_{t \in I} |\hat{U}_n(t)/w_n(t)|$ given the data converges in distribution to $\sup_{t \in I} |U^*(t)/w^*(t)|$. Therefore, c_α can be estimated empirically from a large number of realizations of the conditional distribution of $\sup_{t \in I} |\hat{U}/w|$ given the data.

Motivated from recommendations in the literature for confidence bands of the survivor function and the cumulative hazard function in the one sample case, several choices of the weight w_n can be considered. The choice $w_n(t) = \sqrt{\hat{\sigma}_\Phi(t, t)}$ results in equal precision bands [12], which differ from pointwise confidence intervals in that c_α replaces $z_{\alpha/2}$. The choice $w_n(t) = 1 + \hat{\sigma}_\Phi(t, t)$ results in the Hall-Wellner type bands recommended by Bie et al. [1], which often have narrower widths in the middle of data range and wider widths near the extremes of data range [8]. One could also consider the unweighted version with $w_n(t) \equiv 1$. Compared with the previous two choices, this choice does not require $\hat{\sigma}_\Phi(t, t)$, and hence is easier to implement.

To obtain simultaneous confidence bands for $\Psi(t)$, again consider the weighted process $V_n(t)/w_n(t)$ which converges weakly to the limiting process V^*/w^* . If \tilde{c}_α is the upper α th percentile of $\sup_{t \in I} |V^*/w^*|$, an asymptotic $100(1 - \alpha)\%$ simultaneous confidence band for $\Psi(t)$, $t \in I$, can be obtained as

$$\left(\hat{\Psi}(t) - \frac{\tilde{c}_\alpha w_n(t)}{\sqrt{n}}, \hat{\Psi}(t) + \frac{\tilde{c}_\alpha w_n(t)}{\sqrt{n}} \right).$$

To approximate the critical value \tilde{c}_α , for $t \leq \tau$, define the process

$$\begin{aligned}
\hat{V}_n(t) &= \frac{\int_0^t \hat{D}^T(s) ds \hat{\Omega}}{\sqrt{n}} \left(\sum_{i \leq n_1} \int_0^\tau \hat{\mu}_1 d(\epsilon_i N_i) + \sum_{i > n_1} \int_0^\tau \hat{\mu}_2 d(\epsilon_i N_i) \right) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \leq n_1} \int_0^t \hat{v}_1(w) \int_w^t \hat{C}(x) dx d(\epsilon_i N_i(w)) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i > n_1} \int_0^t \hat{v}_2(w) \int_w^t \hat{C}(x) dx d(\epsilon_i N_i(w)) \\
&= \frac{\hat{D}^T(t) \hat{\Omega}}{\sqrt{n}} \left(\sum_{i \leq n_1} \epsilon_i \delta_i \hat{\mu}_1(X_i) I(X_i \leq \tau) + \sum_{i > n_1} \epsilon_i \delta_i \hat{\mu}_2(X_i) I(X_i \leq \tau) \right) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \leq n_1} \epsilon_i \delta_i \hat{v}_1(X_i) I(X_i \leq t) \int_{X_i}^t \hat{C}(x) dx \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i > n_1} \epsilon_i \delta_i \hat{v}_2(X_i) I(X_i \leq t) \int_{X_i}^t \hat{C}(x) dx, \tag{13}
\end{aligned}$$

where ϵ_i , $i = 1, \dots, n$, are independent standard normal variables that are also independent of the data. In Appendix 2, the process $\hat{V}_n(t)$ given the data is shown to converge weakly to $V^*(t)$. Thus \tilde{c}_α can be approximated empirically from a large number of realizations of the conditional distribution of $\sup_{t \in [a, b]} |\hat{V}(t)/w_n|$ given the data. Similarly to the case for \hat{U}_n , the weight function w_n can be chosen to yield equal precision, Hall-Wellner type, and unweighted confidence bands.

Simulation Studies

For stable moderate sample behavior, the range of the simultaneous confidence bands for both $\Phi(t)$ and $\Psi(t)$ needs to be restricted. Through a series of simulation studies, a data-dependent range was found to result in good performance for moderate samples. The range is obtained by truncating at the 25th percentile of the uncensored data at the lower end, and truncating at the 5th largest uncensored observation at the upper end. By this truncation, the confidence bands are given in a range where a reasonable amount of data are available. Also, in the estimating procedures, the function $\hat{P}(t; \mathbf{b})$ is replaced by an asymptotically equivalent form

$$\exp\left(-\int_0^t \frac{1}{\sum_{i=1}^n I(X_i \geq s)} d\left\{\sum_{i=1}^n \delta_i e^{-b_2 Z_i} I(X_i \leq s)\right\}\right).$$

For simulation studies reported here and for the real data application in section ‘‘Application’’, τ was set to include all data in calculating $\hat{\beta}$. All numerical

Table 1 Empirical coverage probabilities of the three types of simultaneous confidence bands HW, EP, and UW, for the absolute risk reduction Φ and the restricted mean survival difference Ψ , under model (1), based on 1,000 repetitions

Hazard ratio	Censoring (%)	n	Φ			Ψ		
			HW	EP	UW	HW	EP	UW
0.9 \uparrow 1.2	10	100	0.968	0.963	0.974	0.964	0.974	0.955
			0.968	0.963	0.974	0.956	0.972	0.946
			0.950	0.949	0.953	0.957	0.977	0.956
	30	200	0.961	0.964	0.958	0.951	0.965	0.944
			0.954	0.955	0.966	0.949	0.963	0.944
			0.940	0.942	0.945	0.940	0.962	0.937
	50	400	0.954	0.958	0.962	0.952	0.964	0.950
			0.961	0.961	0.967	0.951	0.969	0.946
			0.949	0.945	0.954	0.949	0.962	0.945
1.2 \downarrow 0.8	10	100	0.951	0.946	0.962	0.953	0.964	0.938
			0.951	0.947	0.969	0.959	0.979	0.956
			0.930	0.930	0.949	0.962	0.973	0.960
	30	200	0.956	0.956	0.955	0.952	0.969	0.947
			0.960	0.957	0.962	0.953	0.972	0.949
			0.942	0.933	0.947	0.943	0.960	0.940
	50	400	0.958	0.952	0.958	0.955	0.968	0.944
			0.954	0.955	0.954	0.949	0.966	0.951
			0.951	0.950	0.956	0.948	0.961	0.947

computations were done in *Matlab*. Some representative results are given in Table 1, where lifetime variables were generated with $R(t)$ chosen to yield the standard exponential distribution for the control group. The values of β were $(\log(0.9), \log(1.2))$ and $(\log(1.2), \log(0.8))$, representing 1/3 increase or decrease over time from the initial hazard ratio, respectively. The censoring variables were independent and identically distributed with the log-normal distribution, where the normal distribution had mean c and standard deviation 0.5, with c chosen to achieve various censoring rates. The data were split into the treatment and control groups by a 1:1 ratio. The empirical coverage probabilities were obtained from 1,000 repetitions, and for each repetition, the critical values c_α and \tilde{c}_α were calculated empirically from 1,000 realizations of relevant conditional distributions. For both $\Phi(t)$ and $\Psi(t)$, the equal precision bands, Hall-Wellner type bands and unweighted bands are denoted by EP, HW and UW respectively.

Note that with 1,000 repetitions and $1.96\sqrt{0.95 \cdot 0.05/1,000} = 0.0135$, we expect the empirical coverage probabilities to be mostly greater than 0.9365. In Table 1, the empirical coverage probabilities are greater than 0.9365 for all but three cases. Those three cases occurred for $\Phi(t)$, with 50% censoring and smaller sample sizes. The phenomenon disappeared when $n = 400$. Various additional simulation

Table 2 Empirical coverage probabilities of the three types of simultaneous confidence bands HW, EP, and UW, for the absolute risk reduction Φ and the restricted mean survival difference Ψ , under a monotone hazard ratio model, based on 1,000 repetitions

Hazard ratio	Censoring (%)	n	Φ			Ψ		
			HW	EP	UW	HW	EP	UW
0.9 \uparrow 1.2	10	100	0.973	0.977	0.975	0.964	0.975	0.955
			0.983	0.983	0.987	0.971	0.984	0.963
			0.969	0.973	0.971	0.967	0.986	0.965
	30	200	0.967	0.951	0.961	0.955	0.967	0.937
			0.966	0.965	0.975	0.956	0.971	0.950
			0.956	0.964	0.962	0.966	0.978	0.965
	50	400	0.956	0.916	0.978	0.963	0.965	0.948
			0.967	0.962	0.975	0.961	0.972	0.956
			0.984	0.982	0.979	0.970	0.984	0.969
1.2 \downarrow 0.8	10	100	0.974	0.970	0.979	0.964	0.975	0.953
			0.971	0.964	0.980	0.965	0.983	0.964
			0.966	0.971	0.978	0.976	0.989	0.974
	30	200	0.959	0.930	0.965	0.945	0.962	0.944
			0.971	0.972	0.971	0.947	0.967	0.937
			0.960	0.957	0.969	0.963	0.986	0.961
	50	400	0.935	0.872	0.975	0.960	0.968	0.953
			0.962	0.959	0.976	0.953	0.971	0.952
			0.966	0.958	0.982	0.961	0.974	0.958

studies indicated that the proposed procedures performed well for sample size close to 100 and up, with moderate censoring. Under heavy censoring, the results were still good with uncensored observations close to 50 and up in each treatment group.

To check how robust the procedures are against violation of model assumptions, various monotone hazard ratio models were also considered alternative to the model (1). The results indicated that the proposed procedures continued to perform well. For example, in Table 2, the control group lifetime variables were standard exponential. The hazard ratio was linear from 0 to the 90th percentile of the standard exponential, and continuous and constant afterwards. The initial and end hazard ratios again were (0.9, 1.2) and (1.2, 0.8) respectively, and the censoring variables were the same as before. It can be seen from Table 2 that the confidence bands performed satisfactorily.

To compare efficiency against the non-parametric alternatives based on the Kaplan-Meier estimators, for estimating $\Phi(t)$ and $\Psi(t)$ at various time points, the mean squared errors of the model based estimators and the Kaplan-Meier estimators were examined under model (1) in various simulation studies. Typically the model based estimators have smaller mean squared errors, more so for $\Phi(t)$ than for $\Psi(t)$. Also, the efficiency is higher under heavy censoring and for time points closer to the upper tail region. This is because the Kaplan-Meier becomes increasingly unstable

Table 3 Ratio of mean squared errors of the model based estimators over the Kaplan-meier estimators, for $\Phi(t)$ and $\Psi(t)$ under model (1), at $t = 0.5, 1, 1.5$ respectively, based on 1,000 repetitions

Hazard ratio	Censoring (%)	n	Φ			Ψ		
			0.5	1	1.5	0.5	1	1.5
0.9 \uparrow 1.2	10	100	0.6019	0.6731	0.6286	0.6341	0.7978	0.8470
	30		0.5527	0.6025	0.5248	0.5960	0.7303	0.7717
	50		0.4676	0.4008	0.2447	0.5232	0.6149	0.5872
	10	200	0.6438	0.7090	0.6865	0.6795	0.8436	0.9099
	30		0.5920	0.6436	0.5513	0.6368	0.7763	0.8407
	50		0.5150	0.4403	0.2672	0.5802	0.6744	0.6682
	10	400	0.6831	0.7191	0.6800	0.6975	0.9000	0.9530
	30		0.6321	0.6357	0.5425	0.6747	0.8271	0.8563
	50		0.5523	0.4222	0.2789	0.6255	0.7205	0.6777
1.2 \downarrow 0.8	10	100	0.6251	0.6509	0.6203	0.7195	0.8275	0.8349
	30		0.5897	0.6278	0.5613	0.6850	0.7825	0.8037
	50		0.4778	0.4180	0.2631	0.5662	0.6434	0.6182
	10	200	0.6650	0.7035	0.6902	0.7324	0.8535	0.8930
	30		0.6322	0.6648	0.5816	0.6982	0.8088	0.8467
	50		0.5434	0.4837	0.2893	0.6099	0.6992	0.7015
	10	400	0.7085	0.7081	0.6992	0.7432	0.8973	0.9289
	30		0.6753	0.6742	0.6079	0.7226	0.8560	0.8794
	50		0.6015	0.4661	0.3010	0.6458	0.7595	0.7256

near upper tail region and under heavy censoring. Some representative results are given in Table 3, in terms of the ratio of the mean squared errors of the model based estimators over the Kaplan-Meier estimators, under configurations the same as those for Table 1.

Application

For the Women’s Health Initiative (WHI) randomized controlled trial of combined (estrogen plus progestin) postmenopausal hormone therapy, an elevated coronary heart disease risk was reported, with overall unfavorable health benefits versus risks over an average of 5.6 year study period [10, 21]. Few research reports have stimulated as much public response, since preceding observational research literature suggested a 40–50% reduction in coronary heart disease incidence among women taking postmenopausal hormone therapy. Analysis of the WHI observational study shows a similar discrepancy with the WHI clinical trial for coronary heart disease, stroke, and venous thromboembolism, even after adjusting for confounding factors in the observational study. Following control for time from estrogen-plus-progestin

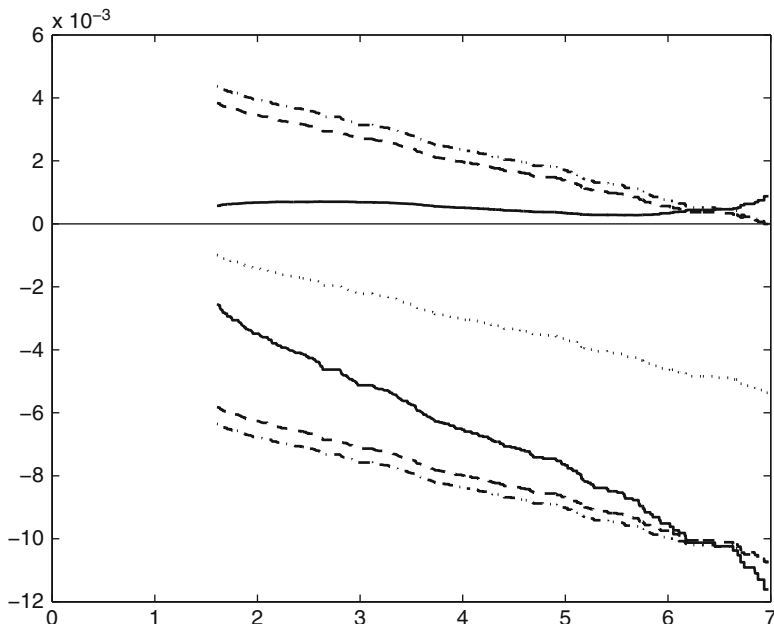


Fig. 4 Simultaneous 95% confidence bands of the absolute risk reduction for the WHI clinical trial stroke data: *Solid line*—equal precision confidence band; *Dashed line*—Hall-Wellner type confidence band; *Dash-dotted lines*—unweighted confidence band; *Dotted line*: Estimated absolute risk reduction

initiation and confounding, hazard ratio estimates were rather similar between the clinical trial and observational study components of WHI, although there was evidence of some remaining difference for stroke [16].

In the introduction, it was mentioned that for stroke, the estimated absolute risk reduction based on model (1) provides a good compromise between the results from the two partitioning approaches under the piece-wise Cox model. Let us illustrate the methods developed in the previous sections with the stroke data from the WHI clinical trial. Among the 16,608 postmenopausal women ($n_1 = 8,102$), there were 151 and 107 events observed in the treatment and control group respectively, implying about 98% censoring, primarily by the trial stopping time. Fitting model (1) to this data set, we get $\hat{\beta} = (0.32, -1.69)^T$. Plots of the model based survival curves and the Kaplan-Meier curves for the two groups show that the model is reasonable. The residual plot as mentioned in Yang and Prentice [23] also indicates a good model fit. These plots are not displayed here to save space. The three 95% simultaneous confidence bands for the absolute risk reduction are given in Fig. 4. From Fig. 4, it can be seen that both the Hall-Wellner type band and the unweighted band maintain a roughly constant width through the data range considered. In comparison, the equal precision band has width gradually increasing as the standard error of the estimated absolute risk reduction increases over time. Also, the width of

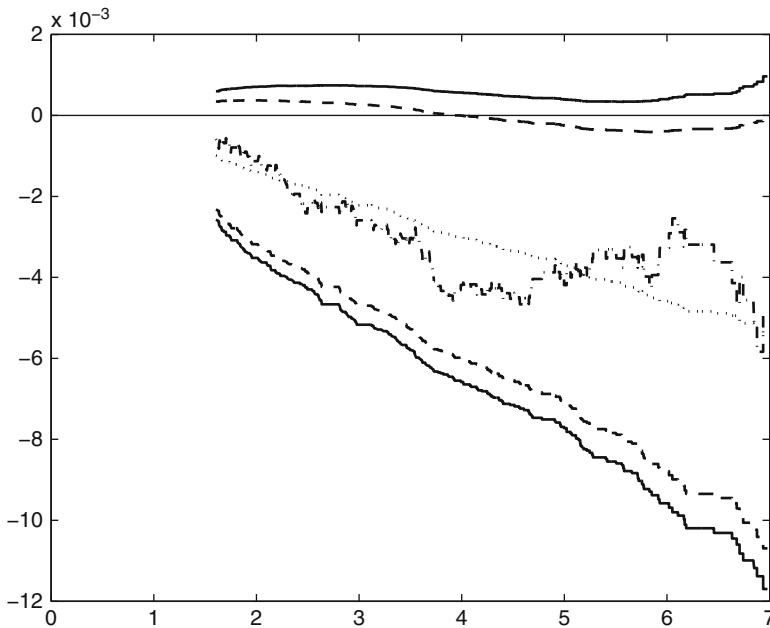


Fig. 5 95% equal precision confidence band and pointwise 95% confidence intervals of the absolute risk reduction for the WHI clinical trial stroke data: *Solid line*—equal precision confidence band; *Dashed line*—pointwise 95% confidence intervals; *Dotted line*: Model based estimator of the absolute risk reduction; *Dash-dotted lines*—Kaplan-Meier estimator of the absolute risk reduction

the equal precision band is narrower than those of the Hall-Wellner type band and the unweighted band through most of the range. Similar phenomena are often present in other applications not reported here. Thus it is recommended that the equal precision band be used in making inference on the absolute risk reduction under model (1). Note that the simple bootstrap method for approximating c_α when $w_n \equiv 1$ is already much more computationally intensive than the normal resampling approximation. With $w_n(t) = \sqrt{\hat{\sigma}_\Phi(t)}$, the bootstrap method would require one more level of bootstrapping samples, thus further increasing the computational burden. In comparison, once $\hat{\sigma}_\Phi(t)$ is obtained with the martingale structure, the normal resampling approximation only needs a small additional computation and programming cost. Similar remarks are also applicable to the case with the restricted mean survival difference.

To compare the point-wise confidence intervals and the simultaneous confidence band, Fig. 5 displays 95% point-wise confidence intervals and the simultaneous confidence band for the stroke data. The simultaneous confidence band is slightly wider than the point-wise confidence intervals and maintains the same rate of inflation in width throughout the range. The confidence intervals and confidence band indicate some evidence that the absolute risk reduction is negative in the

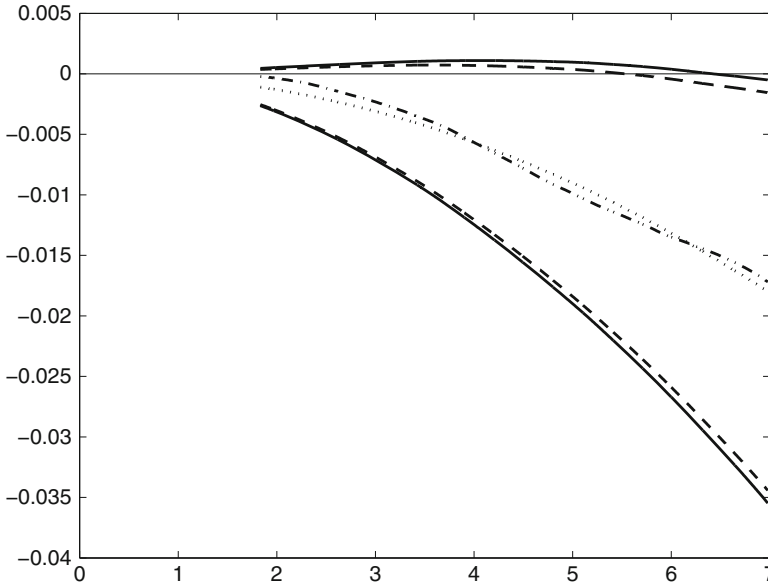


Fig. 6 95% equal precision confidence band and pointwise 95% confidence intervals of the mean restricted survival difference for the WHI clinical trial stroke data: *Solid line*—equal precision confidence band; *Dashed line*—pointwise 95% confidence intervals; *Dotted line*: Model based estimator of the mean restricted survival difference; *Dash-dotted lines*—Kaplan-Meier estimator of the mean restricted survival difference

range of 4–7 years, but the evidence is not very strong. Figure 5 also includes the Kaplan-Meier estimator. Between the semiparametric and nonparametric estimators, The model based estimator is smoother, the Kaplan-Meier estimator is more volatile and oscillates around the model based estimator. The model based estimator captures the general decreasing trend in the absolute risk reduction, and averages out the deviations from that trend, particularly in the range of 3.5 to 7 years.

For the restricted mean survival difference, Fig. 6 displays the estimator under model (1), the 95% point-wise confidence intervals and simultaneous equal precision confidence band for the stroke data. Since the restricted mean survival difference is a summary measure, the estimators are smoother compared with those for the absolute risk reduction. Also, the semiparametric and nonparametric estimators show a better agreement compared with the case for the absolute risk reduction. Furthermore, the inflation of width by the band over the point-wise confidence intervals is smaller compared with the situation in Fig. 5. This is possibly because the restricted mean survival difference, a summary measure, may have higher correlation at different time points compared with the absolute risk reduction at those same time points. From Fig. 6, there is some evidence that the restricted mean survival difference is negative towards the end of the data range.

Discussion

We have studied the asymptotic properties of the estimators for the absolute risk reduction and the restricted mean survival difference under the short-term and long-term hazards model. Point-wise confidence intervals and simultaneous confidence bands are developed for these measures. These procedures can have a sufficiently wide range of applications because of the flexibility of the model. In simulation studies, the confidence bands have good performance for moderate samples. Among the versions with different weights, the equal precision confidence band is recommended. It has width that is proportional to the standard error at each time point and often results in narrower width in most of the data range. It also demonstrates the inflation of the confidence interval width needed for simultaneous inference. For the restricted mean survival difference, often the measure at a fixed time, say t_0 years, with t_0 close to the maximum follow-up period of the clinical trial, is of interest. In those situations, the point-wise confidence intervals may suffice.

Compared with the nonparametric methods based on the Kaplan-Meier estimator, the semiparametric approach developed here produces more smooth estimators and more stable behaviors, especially near the end of the data range. Thus it provides a good alternative to the nonparametric approach should the model be appropriate. The model also permits inference on the hazard ratio function, as described in Yang and Prentice [23], where the nonparametric approach could result in wide confidence intervals at the tail regions. When the model provides good fit to the data, together the confidence intervals and bands on the hazard ratio, the absolute risk reduction and the restricted mean survival difference, present good visual tools for assessing the temporal pattern and cumulative effect of the treatment. It is also of interest to extend the results here to epidemiological studies by considering the regression setting and adjusting for covariate. These and other problems are worthy of further exploration.

Acknowledgements The original version of this article has previously been published in Lifetime Data Analysis in 2013.

Appendix 1: Consistency

Throughout the Appendices, we assume the following regularity conditions, which is a little weaker than the conditions used in Yang and Prentice [22].

Condition 1. $\lim_{n \rightarrow \infty} \frac{n_1}{n} = \rho \in (0, 1)$.

Condition 2. The survivor function G_i of C_i given Z_i is continuous and satisfies

$$\frac{1}{n} \sum_{i \leq n_1} G_i(t) \rightarrow \Gamma_1, \quad \frac{1}{n} \sum_{i > n_1} G_i(t) \rightarrow \Gamma_2,$$

uniformly for $t \leq \tau$, for some Γ_1, Γ_2 , and $\tau < \tau_0$ such that $\Gamma_j(\tau) > 0$, $j = 1, 2$.

Condition 3. The survivor functions S_C and S_T are absolutely continuous and $S_C(\tau) > 0$.

Under these conditions, the strong law of large numbers implies that (3) is satisfied.

For $t \leq \tau$, define

$$L(t) = \Gamma_1 S_C + \Gamma_2 S_T,$$

$$U_j(t; \mathbf{b}) = \int_0^t \Gamma_1 dF_C + \exp(-b_j) \int_0^t \Gamma_2 dF_T, \quad j = 1, 2,$$

$$\Lambda_j(t; \mathbf{b}) = \int_0^t \frac{dU_j(s; \mathbf{b})}{L(s)}, \quad j = 1, 2,$$

$$P(t; \mathbf{b}) = \exp\{-\Lambda_2(t; \mathbf{b})\}, \quad R(t; \mathbf{b}) = \frac{1}{P(t; \mathbf{b})} \int_0^t P(s; \mathbf{b}) d\Lambda_1(s; \mathbf{b}),$$

$$f_j^0(t; \mathbf{b}) = \frac{\exp(-b_j) R^{j-1}(t; \mathbf{b})}{\exp(-b_1) + \exp(-b_2) R(t; \mathbf{b})}, \quad j = 1, 2,$$

$$m_j(\mathbf{b}) = \left\{ \int_0^\tau f_j^0 \Gamma_2(t) dF_T(t) - \int_0^\tau \frac{f_j^0 \Gamma_2(t) S_T(t) dR(t; \mathbf{b})}{\exp(-b_1) + \exp(-b_2) R(t; \mathbf{b})} \right\}, \quad j = 1, 2,$$

and $m(\mathbf{b}) = (m_1(\mathbf{b}), m_2(\mathbf{b}))'$. We will also assume

Condition 4. The function $m(\mathbf{b})$ is non-zero for $b \in \mathcal{B} - \{\beta\}$, where \mathcal{B} is a compact neighborhood of β .

Theorem 1. *Suppose that Conditions 1–4 hold. Then, (i) the zero $\hat{\beta}$ of $Q(\mathbf{b})$ in \mathcal{B} is strongly consistent for β ; (ii) $\hat{\Phi}(t)$ is strongly consistent for $\Phi(t)$, uniformly for $t \in [0, \tau]$, and $\hat{\Psi}(t)$ is strongly consistent for $\Psi(t)$, uniformly on $t \in [0, \tau]$; (iii) $\hat{\Omega}$ converges almost surely to a limiting matrix Ω^* .*

Proof. Under Conditions 1–3, the limit of $\sum_{i=1}^n I(X_i \geq t)/n$ is bounded away from zero on $t \in [0, \tau]$. Thus, with probability 1,

$$\frac{\sum_{i=1}^n \delta_i e^{-b_j Z_i} I(X_i = t)}{\sum_{i=1}^n \delta_i I(X_i \geq t)} \rightarrow 0, \quad j = 1, 2, \quad (14)$$

uniformly for $t \in [0, \tau]$ and $b \in \mathcal{B}$. From this, one also has, with probability 1,

$$|\Delta \hat{P}(t; \mathbf{b})| \rightarrow 0, \quad |\Delta \hat{R}(t; \mathbf{b})| \rightarrow 0, \quad (15)$$

uniformly for $t \in [0, \tau]$ and $b \in \mathcal{B}$, where Δ indicates the jump of the function in t .

Define the martingale residuals

$$\hat{M}_i(t; \mathbf{b}) = \delta_i I(X_i \leq t) - \int_0^t I(X_i \geq s) \frac{\hat{R}(ds; \mathbf{b})}{e^{-b_1 Z_i} + e^{-b_2 Z_i} \hat{R}(s; \mathbf{b})}, \quad 1 \leq i \leq n.$$

From (12) and (13), and the fundamental theorem of calculus, it follows that, with probability 1,

$$Q(\mathbf{b}) = \sum_{i=1}^n \int_0^\tau \{f_i(t; \mathbf{b}) + o(1)\} \hat{M}_i(dt; \mathbf{b}), \quad (16)$$

uniformly in $t \leq \tau$, $b \in \mathcal{B}$ and $i \leq n$, where $f_i = (f_{1i}, f_{2i})^T$, with

$$f_{1i}(t; \mathbf{b}) = \frac{Z_i e^{-b_1 Z_i}}{e^{-b_1 Z_i} + e^{-b_2 Z_i} \hat{R}(t; \mathbf{b})}, \quad f_{2i}(t; \mathbf{b}) = \frac{Z_i e^{-b_2 Z_i} \hat{R}(t; \mathbf{b})}{e^{-b_1 Z_i} + e^{-b_2 Z_i} \hat{R}(t; \mathbf{b})}.$$

From the strong law of large numbers ([15], p. 41) and repeated use of Lemma A1 of Yang and Prentice [22], one obtain, with probability 1,

$$\hat{P}(t; \mathbf{b}) \rightarrow \hat{P}(t; \mathbf{b}), \quad \hat{R}(t; \mathbf{b}) \rightarrow R(t; \mathbf{b}), \quad Q(\mathbf{b})/n \rightarrow m(\mathbf{b}), \quad (17)$$

uniformly in $t \leq \tau$ and $\mathbf{b} \in \mathcal{B}$. From these results and Condition 4, one obtains the strong consistency of $\hat{\beta}$, $\hat{\Phi}(t)$ and $\hat{\Psi}(t)$, and almost sure convergence of $\hat{\Omega}$.

Appendix 2: Weak Convergence

For $C(t)$, $D(t)$, $\mu_1(t)$, $\mu_2(t)$, $v_1(t)$, $v_2(t)$, let $C^*(t)$, $D^*(t)$, etc. be their almost sure limit. In addition, let L_j be the almost sure limit of K_j/n , $j = 1, 2$. For $0 \leq s, t < \tau$, let

$$\begin{aligned} & \sigma_\Phi(s, t) \\ &= D^{*T}(s) \Omega^* \left(\int_0^\tau \frac{\mu_1^* \mu_1^{*T}}{1+R} L_1 dR + \int_0^\tau \frac{\mu_2^* \mu_2^{*T}}{e^{-\beta_1} + e^{-\beta_2 R}} L_2 dR \right) \Omega^{*T} D^*(t) \\ &+ C^*(s) C^*(t) \left(\int_0^s \frac{v_1^{*2}}{1+R} L_1 dR + \int_0^s \frac{v_2^{*2}}{e^{-\beta_1} + e^{-\beta_2 R}} L_2 dR \right) \\ &+ C^*(t) D^{*T}(s) \Omega^* \left(\int_0^t \frac{\mu_1^* v_1^*}{1+R} L_1 dR + \int_0^t \frac{\mu_2^* v_2^*}{e^{-\beta_1} + e^{-\beta_2 R}} L_2 dR \right) \\ &+ C^*(s) D^{*T}(t) \Omega^* \left(\int_0^s \frac{\mu_1^* v_1^*}{1+R} L_1 dR + \int_0^s \frac{\mu_2^* v_2^*}{e^{-\beta_1} + e^{-\beta_2 R}} L_2 dR \right), \quad (18) \end{aligned}$$

and

$$\begin{aligned}
& \sigma_{\Psi}(s, t) \\
= & \int_0^s D^{*T}(x) dx \Omega^* \left(\int_0^{\tau} \frac{\mu_1^*(w) \mu_1^{*T}(w)}{1+R(w)} L_1(w) dR(w) \right. \\
& + \int_0^{\tau} \frac{\mu_2^*(w) \mu_2^{*T}(w)}{e^{-\beta_1} + e^{-\beta_2} R(w)} L_2(w) dR(w) \left. \right) \Omega^{*T} \int_0^t D^{*T}(x) dx \\
& + \int_0^s \frac{v_1^{*2}(w)}{1+R(w)} \left(\int_w^s C^*(x) dx \right)^2 L_1(w) dR(w) \\
& + \int_0^s \frac{v_2^{*2}(w)}{e^{-\beta_1} + e^{-\beta_2} R(w)} \left(\int_w^s C^*(x) dx \right)^2 L_2(w) dR(w) \\
& + \int_0^s D^{*T}(x) dx \Omega^* \int_0^t \frac{\mu_1^*(w) v_1^*(w)}{1+R(w)} \left(\int_w^t C^*(x) dx \right) L_1(w) dR(w) \\
& + \int_0^s D^{*T}(x) dx \Omega^* \int_0^t \frac{\mu_2^*(w) v_2^*(w)}{e^{-\beta_1} + e^{-\beta_2} R(w)} \left(\int_w^t C^*(x) dx \right) L_2(w) dR(w) \\
& + \int_0^t D^{*T}(x) dx \Omega^* \int_0^s \frac{\mu_1^*(w) v_1^*(w)}{1+R(w)} \left(\int_w^s C^*(x) dx \right) L_1(w) dR(w) \\
& + \int_0^t D^{*T}(x) dx \Omega^* \int_0^s \frac{\mu_2^*(w) v_2^*(w)}{e^{-\beta_1} + e^{-\beta_2} R(w)} \left(\int_w^s C^*(x) dx \right) L_2(w) dR(w). \quad (19)
\end{aligned}$$

Theorem 2. *Suppose that Conditions 1–4 hold and that the matrix Ω^* is non-singular. Then, (i) U_n is asymptotically equivalent to the process \tilde{U}_n in (8) which converges weakly to a zero-mean Gaussian process U^* on $[0, \tau]$, with covariance function $\sigma_{\Phi}(s, t)$ in (18). In addition, $\hat{U}_n(s)$ given the data converges weakly to the same limiting process U^* . (ii) $V_n(t)$ is asymptotically equivalent to the process \tilde{V}_n in (11) which converges weakly to the zero-mean Gaussian process $\int_0^t U^*(s) ds$ on $t \in [0, \tau]$, with covariance function $\sigma_{\Psi}(s, t)$ in (19). The process $\int_0^t \hat{V}_n(s) ds$ given the data also converges weakly to the same limiting process $\int_0^t U^*(s) ds$.*

Proof. (i) As in the proof for Theorem A2 (ii) in Yang and Prentice [22], from the strong embedding theorem and (16), $Q(\beta)/\sqrt{n}$ can be shown to be asymptotically normal. Now Taylor series expansion of $Q(\mathbf{b})$ around β and the non-singularity of Ω^* imply that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal. From the \sqrt{n} -boundedness of $\hat{\beta}$,

$$\sqrt{n}(\hat{R}(t; \hat{\beta}) - \hat{R}(t; \beta)) = \frac{\partial R(t; \beta)}{\partial \beta} \sqrt{n}(\hat{\beta} - \beta) + o_p(1),$$

uniformly in $t \leq \tau$. These results, some algebra and Taylor series expansion together show that U_n is asymptotically equivalent to \tilde{U}_n . Similarly to the proof of the

asymptotic normality of $Q(\beta)/\sqrt{n}$, one can show that \tilde{U}_n converges weakly to a zero-mean Gaussian process. Denote the limiting process by U^* . From the martingale integral representation of \tilde{U}_n , it follows that the covariation process of U^* is given by $\sigma(s, t)$ in (18), which can be consistently estimated by $\hat{\sigma}(s, t)$ in (9). By checking the tightness condition and the convergence of the finite-dimensional distributions, it can be shown that $\hat{U}_n(s)$ given the data also converges weakly to U^* .

(ii) From the results in (i), the assertions on V_n and \tilde{V}_n follow.

References

1. Bie, O., Borgan, O., Liestøl, K.: Confidence intervals and confidence bands for the cumulative hazard rate function and their small-sample properties. *Scand. J. Stat.* **14**, 221–233 (1987)
2. Chen, P., Tsiatis, A.A.: Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57**, 1030–1038 (2001)
3. Cheng, S.C., Wei, L.J., Ying, Z.: Predicting survival probabilities with semiparametric transformation models. *J. Am. Stat. Assoc.* **92**, 227–235 (1997)
4. Cox, D.R.: Regression models and life-tables (with Discussion). *J. R. Stat. Soc. B* **34**, 187–220 (1972)
5. Dabrowska, D.M., Doksum, K.A., Song, J.: Graphical comparison of cumulative hazards for two populations. *Biometrika* **76**, 763–773 (1989)
6. Kalbfleisch, J.D., Prentice, R.L.: *The Statistical Analysis of Failure Time Data*, 2nd edn. Wiley, New York (2002)
7. Kaplan, E., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958)
8. Lin, D.Y., Wei, L.J., Ying, Z.: Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572 (1993)
9. Lin, D.Y., Fleming, T.R., Wei, L.J.: Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81 (1994)
10. Manson, J.E., Hsia, J., Johnson, K.C., Rossouw, J.E., Assaf, A.R., Lasser, N.L., Trevisan, M., Black, H.R., Heckbert, S.R., Detrano, R., Strickland, O.L., Wong, N.D., Crouse, J.R., Stein, E., Cushman, M., Women'S Health Initiative Investigators: Estrogen plus progestin and the risk of coronary heart disease. *N. Engl. J. Med.* **349**, 523–534 (2003)
11. McKeague, I.W., Zhao, Y.: Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Stat. Probab. Lett.* **60**, 405–415 (2002)
12. Nair, V.N.: Confidence bands for survival functions with censored data: a comparative study. *Technometrics* **26**, 265–275 (1984)
13. Parzen, M.I., Wei, L.J., Ying, Z.: Simultaneous confidence intervals for the difference of two survival functions. *Scand. J. Stat.* **24**, 309–314 (1997)
14. Peng, L., Huang, Y.: Survival analysis with temporal covariate effects. *Biometrika* **94**, 719–733 (2007)
15. Pollard, D.: *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward (1990)
16. Prentice, R.L., Langer, R., Stefanick, M.L., Howard, B.V., Pettinger, M., Anderson, G., Barad, D., Curb, J.D., Kotchen, J., Kuller, L., Limacher, M., Wactawski-Wende, J., Women'S Health Initiative Investigators: combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *Am. J. Epidemiol.* **162**, 404–414 (2005)

17. Royston, P., Parmar, M.K.: The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat. Med.* **19**, 2409–2421 (2011)
18. Schaubel, D.E., Wei, G.: Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics* **67**, 29–38 (2011)
19. Tian, L., Zucker, D., Wei, L.J.: On the Cox model with time-varying regression coefficients. *J. Am. Stat. Assoc.* **100**, 172–183 (2005)
20. Tong, X., Zhu, C., Sun, J.: Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments. *Can. J. Stat.* **35**, 575–584 (2007)
21. Writing Group for the Women’s Health Initiative Investigators: Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women’s health initiative randomized controlled trial. *J. Am. Med. Assoc.* **288**, 321–333 (2002)
22. Yang, S., Prentice, R.L.: Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika* **92**, 1–17 (2005)
23. Yang, S., Prentice, R.L.: Estimation of the 2-sample hazard ratio function using a semiparametric model. *Biostatistics* **12**, 354–368 (2011)
24. Zucker, D.M.: Restricted mean life with covariates: modification and extension of a useful survival analysis method. *J. Am. Stat. Assoc.* **93**, 702–709 (1998)

Connecting Threshold Regression and Accelerated Failure Time Models

Xin He and G.A. Whitmore

Abstract The accelerated failure time model is one of the most commonly used alternative methods to the Cox proportional hazards model when the proportional hazards assumption is violated. Threshold regression is a relatively new alternative model for analyzing time-to-event data with non-proportional hazards. It is based on first-hitting-time models, where the time-to-event data can be modeled as the time at which the stochastic process of interest first hits a boundary or threshold state. This paper compares threshold regression and accelerated failure time models and demonstrates the situations when the accelerated failure time model becomes a special case of the threshold regression model. Three illustrative examples from clinical studies are provided.

Introduction

Statistical models for analyzing time-to-event or survival data are important in diverse fields of scientific enquiry. Two broad classes of such models are the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model. Threshold regression (TR) is a relatively new model for analyzing time-to-event data. See Lee and Whitmore [20] for an overview of TR. Lee and Whitmore [21] undertook a study of the connection between TR and PH models. In this paper, we study how and to what extent AFT models are embedded within the TR family.

X. He (✉)

University of Maryland, College Park, MD, USA

e-mail: xinhe@umd.edu

G.A. Whitmore

McGill University, Montreal, QC, Canada

e-mail: george.whitmore@mcgill.ca

Our study reveals much about both types of models and integrates previously unconnected findings. Our discussion of the relation between TR and AFT will focus on the fields of medicine and health. The reader will see, however, that the theoretical ideas and results extend immediately to other disciplines that are concerned with duration data, such as engineering and economics. Several case examples from clinical studies are provided to illustrate many of the results.

Accelerated Failure Time (AFT) Model

The main idea of an accelerated failure time model is that characteristics of an individual or the individual's environment, as described by an individual covariate vector z , tend to accelerate or decelerate the progress of the individual toward a medical endpoint, such as death, initiation of a cancer, and the like. AFT is a topic covered by most reference books that deal with survival data and event history analysis [1, 14, 15, 18]. Let T denote the failure time, then a common formulation of the AFT model is as a log-linear regression model of the following form:

$$\ln(T) = z'\gamma + U, \quad (1)$$

where z is a vector of covariates, γ is the vector of corresponding regression coefficients, and U is a baseline random term that does not depend on z . For different individuals, the random terms U are typically assumed to be independent and identically distributed random variables. When the distribution of the baseline random term is known or specified, the AFT model is a parametric model. A number of conventional survival distributions, such as the Weibull, gamma, log-logistic and log-normal distribution families, are mathematically tractable under this logarithmic transformation.

An equivalent formulation to (1) postulates the existence of a baseline survival function $S_0(r)$ and an acceleration multiplier $A > 0$ for the survival time T of each individual such that random variable $R = T/A$ follows the baseline survival function; in other words,

$$\Pr(T > t|A) = S(t|A) = S_0(t/A). \quad (2)$$

The effect of the multiplier A is to amplify or shrink the time scale, according to whether A is greater or less than 1. Subsequently, we refer to t as calendar or clock time and refer to the transformed time $r = t/A$ as analytical, operational or running time.

The acceleration multiplier A is usually taken to be some regression function of the covariate vector z . A common choice is the log-linear form $\ln(A) = z'\gamma$ or, equivalently, $A = \exp(z'\gamma)$. If $z = 0$ then $A = 1$ and we have the reference case in which time progresses at the rate dictated by the baseline survival distribution.

The log-linear form $\ln(A) = z'\gamma$ immediately links the alternative AFT formulations in (1) and (2) because the survival function in (2) can be re-expressed as follows:

$$S_0(t/A) = S_0[\exp\{\ln(t) - z'\gamma\}] = \Pr(U > \ln(t) - z'\gamma).$$

Thus, the baseline random term U in the log-linear formulation and the running time variable R , with the survival function $S_0(r)$, are related by the identity $\ln(R) \equiv U$.

Threshold Regression (TR) Model

Threshold regression refers to a family of survival models in which the survival time or time to event is the first hitting time of a boundary (or threshold) by a stochastic process. Parameters of the stochastic process and boundary can be linked to covariates using regression functions that are suited to the application at hand. Hence, the word “regression” is used in the name. Following Lee and Whitmore [20], the first-hitting-time (FHT) model has two basic components, namely, (1) a parent stochastic process $\{Y(t), t \in \mathcal{T}, y \in \mathcal{Y}\}$ with initial value $Y(0) = y_0$, where \mathcal{T} is the time space and \mathcal{Y} is the state space of the process; and (2) a boundary set \mathcal{B} , where $\mathcal{B} \subset \mathcal{Y}$. The initial value of the process y_0 is assumed to lie outside the boundary set \mathcal{B} . The first hitting time of \mathcal{B} is the random variable $T = \inf\{t : Y(t) \in \mathcal{B}\}$, which is the time when the stochastic process first encounters set \mathcal{B} . The unknown parameters in the parent stochastic process $\{Y(t)\}$ and the boundary set \mathcal{B} may be connected to linear combinations of covariates using suitable regression link functions. For instance, a variance parameter σ^2 may employ a logarithmic link function, such as $\ln(\sigma^2) = z'\beta$, where z is the vector of covariates and β is the vector of corresponding regression coefficients.

In many TR applications, the natural time scale of the parent stochastic process is not calendar or clock time t but rather some alternative time scale r , where $r(t)$ is a non-decreasing transformation of calendar time t , with $r(0) = 0$. As with the AFT model, we refer to r as the running time. With this transformation, the parent process is defined in terms of running time r as $\{Y(r)\}$ and the subordinated process $\{Y[r(t)]\}$ defines the original process in terms of calendar time t .

Connecting TR and AFT Models

The preceding overview of the AFT and TR models contains the connection between the two types of models. The connection lies in the running time transformation $r(t|z)$. We extend the traditional notion of an AFT model by allowing the following more general formulation:

$$\Pr(T > t|z) = S(t|z) = S_0[r(t|z)]. \quad (3)$$

Here again $S_0(r)$ is a baseline survival function and $r(t|z)$ is a non-decreasing function of calendar time t that is dependent on the covariate vector z . We require $r(0|z) = 0$ for all z . The transformation $r(t|z)$ encapsulates what we mean by the acceleration and deceleration of time.

The generalized AFT model in (3) is also a TR model if $S_0(r)$ is a first-hitting-time distribution for some baseline process $\{Y_0(r)\}$, baseline boundary set \mathcal{B}_0 and running time $r(t|z)$. The AFT model will not be a TR model if $S_0(r)$ is not of the FHT variety. From our experience, it is difficult to conceive of an AFT model that is scientifically meaningful which lies outside the TR family. On the other hand, the AFT model in (3) is a proper subset of the TR family. TR models extend the AFT model (3) whenever the parameters of the baseline survival function $S_0(t)$ are made to depend on the covariate vector z . In this extension, we show this dependence by the notation $S_0(t|z)$.

A large variety of practical AFT models are created by appropriate choices for the baseline survival function $S_0(r)$ and the running time $r(t|z)$. The following examples of TR models that are also AFT models illustrate the range of possibilities:

1. **Poisson process.** Consider a Poisson process $\{N_0(r)\}$ with a baseline hazard rate $\lambda_0 > 0$. The time until occurrence of the first event in the baseline process has survival function $S_0(r) = \exp(-\lambda_0 r)$. Under the simplest acceleration multiplier $A = \exp(z'\gamma)$, the running time function becomes $r(t|z) = t/A$ and then the survival function of the AFT model takes the form:

$$S(t|z) = S_0[t/A] = \exp[-t\lambda_0 \exp(-z'\gamma)].$$

Observe that the baseline hazard rate λ_0 may be viewed as the intercept term in the covariate regression function through the correspondence $\lambda_0 = \exp(\gamma_0)$. Conventional methods of statistical inference for survival data can be used to estimate the vector of regression coefficients γ and the baseline hazard rate λ_0 . It is noteworthy that this AFT model is also a proportional hazards model with a family of constant hazard functions $\lambda_0 \exp(-z'\gamma)$. If the time to, say, the k th event in the Poisson process is of scientific interest, then the baseline survival function $S_0(r)$ is an Erlang distribution of order k with scale parameter λ_0 (a special gamma distribution). Substituting $r(t|z) = t \exp(-z'\gamma)$ for r in the baseline survival function produces an AFT family with a gamma error structure.

2. **Wiener process.** Consider the FHT for a Wiener diffusion process $\{Y(r)\}$ starting at $Y(0) = y_0 > 0$ and having a boundary at zero. Let the baseline case be defined by the mean parameter $\mu_0 < 0$ and a unit variance parameter. The baseline survival function $S_0(r)$ has an inverse Gaussian form that depends on parameters y_0 and μ_0 . Again, for simplicity, if the running time function $r(t|z)$ is taken as $t \exp(-z'\gamma)$, then the survival function of the corresponding AFT model is given by

$$S(t) = S_0[t \exp(-z'\gamma)].$$

In this scenario, the running time function $r(t|z)$ characterizes the same AFT and TR model. If, however, the boundary of the process were made to depend on the covariate vector z then the TR model becomes broader than an AFT model. Our discussion of practical issues and specific case illustrations later will draw out this important distinction.

The general AFT model in (3) is not new. There is a large literature dealing with survival models having collapsible, composite, and alternative time scales that are essentially of the form shown in (3). See, for example, Oakes [26], Kordonsky and Gertsbakh [16], Duchesne and Lawless [5], and Duchesne and Rosenthal [6]. What is new in our development is the placement of this general class of AFT models within the context of threshold regression and the elucidation of some practical variants of the model that may be valuable in medical applications.

Variants of AFT Model

To give a flavor of the variety of running time transformations that are available for AFT model (3), we present a few illustrations next.

1. **Multiplier AFT model.** The multiplier version of the AFT model in (2) is a special case of the general formation in (3) as may be seen if we define $r(t|z) = t/\exp(z'\gamma)$. The multiplier version is simple in that it postulates a constant rate of progression of illness or disease, with the rate varying with z .
2. **Change-point AFT model.** An important variant of the preceding model is one in which acceleration engages at a point in time or *change-point* c . A simple version of this model is:

$$r(t|z) = \begin{cases} t & \text{if } t \leq c(z), \\ c(z) + [t - c(z)]\exp(z'\gamma) & \text{if } t > c(z). \end{cases} \quad (4)$$

This version makes the change point c a function of the covariate vector z and is a special case of the *exposure AFT model* that follows.

3. **Exposure AFT model.** In many applications, an individual is exposed during different intervals to toxins or other harmful influences, in varying intensities, that can accelerate the onset of a medical endpoint. The following exposure version of AFT model (3) is useful in this context:

$$r(t|z) = \sum_{j=1}^J \alpha_j(z)t_j, \quad \text{where } t = \sum_{j=1}^J t_j, \alpha_j(z) \geq 0, \text{ and } \alpha_1(z) = 1. \quad (5)$$

Here t_j is the time an individual is exposed to toxin j during calendar interval $(0, t)$. Toxin 1 is taken as the reference exposure type. The reference type might be, for instance, a non-toxic environment. Coefficients $\alpha_j(z)$ determine the accelerator or decelerator effect associated with exposure to toxin j , relative to

the reference type. Covariates z modify the α_j parameters. The equation $t = \sum_j t_j$ is an accounting equation that ensures that every moment of calendar time t is spent in one of the J exposure types. See Lee et al. [22, 23] for an example of (5) in the context of the exposure of railroad workers to diesel exhaust and the onset of lung cancer. This exposure model is an extension of the simple multiplier model (2) because $\alpha_j(z)t_j$ allows for a different multiplier (α_j) for each type of exposure (and each vector z).

4. **Stochastic AFT model.** Running time in some applications will proceed like a stochastic process $\{R(t)\}$ that has non-decreasing sample paths. The function $r(t|z)$ in (3) would then be such a sample path. Refer to Lawless and Crowder [19] for an application to crack propagation in which the gamma process serves as a running time (and degradation process).

Further Adaptations of AFT Model

1. **Cure rate.** A *cure rate* version of the AFT model in (3) allows for the possibility that an individual will be cured of the disease that would bring on the medical endpoint or be immune to it. A cure rate is accommodated in the AFT model if the baseline survival function is given a probability mass p at infinity, with $0 < p < 1$, as follows:

$$S_0(r) = p + (1 - p)S_*(r). \quad (6)$$

Here $S_*(r)$ denotes the baseline survival function of those individuals that are susceptible to the medical endpoint. A careful look at this AFT formulation, however, shows that it may have limited practical application. As acceleration (or deceleration) of time affects only the running time r , formula (6) shows that all individuals in this formulation must have the same cure rate p . The basic issue is that acceleration, pure and simple, only modifies the time scale and an immune or cured individual would not be influenced by its effect. In contrast, TR models in general do not have this restriction. Thus, the cure-rate case is one type that distinguishes the AFT model from more general TR models.

2. **Initial disease progression.** In some investigations, individuals do not enter the study at the same stage of disease progression in the sense that each individual has already experienced some “wear and tear” at the outset. This initial progression may be interpreted as an initial running time $r_0(z)$, which varies with the covariate vector z . In this case, general AFT model (3) takes the form of the following conditional survival function:

$$\Pr(T > t|z) = S(t|z) = \frac{S_0[r_0(z) + r(t|z)]}{S_0[r_0(z)]}. \quad (7)$$

The conditioning in this model is necessary because the individual has experienced running time $r_0(z)$ without yet experiencing the medical endpoint (i.e., the survival distribution $S_0(\cdot)$ is left truncated at r_0).

Illustrative Examples

In this section, we illustrate the comparison between TR and AFT models using three previously published datasets. In each of these datasets, there is a single binary group indicator z . Let $Y(r)$ denote the health status of a patient, which first hits 0 at the event time. Assume that $Y(r)$ can be described by a Wiener diffusion process with the variance parameter $\sigma^2 = 1$ and that $\ln(y_0)$ and μ have the following forms

$$\ln(y_0) = \alpha_0 + \alpha_1 z, \quad \mu = \beta_0 + \beta_1 z,$$

where α_0 and β_0 denote the mean logarithm of the initial health status and the mean change of health status for patients in the reference group ($z = 0$), and α_1 and β_1 represent the group effects on the initial health status and the mean change of health status, respectively.

Kidney Dialysis Dataset

In a clinical study conducted by The Ohio State University from January 1988 to May 1990, the time to first cutaneous exit site infection (in months) was recorded for patients with renal insufficiency, where the cutaneous exit site infection was defined as a painful cutaneous exit site and positive cultures, or peritonitis, defined as the presence of clinical symptoms, elevated peritoneal dialytic fluid white blood cell count (100 white blood cells/ μ L with $>50\%$ neutrophils), and positive peritoneal dialytic fluid cultures [25]. Following Klein and Moeschberger [15], we restrict our attention to 43 patients who utilized a surgically placed catheter and 76 patients who utilized a percutaneous placement of their catheter.

To analyze the data, define z to be equal to 1 if the patient utilized a percutaneous placed catheter and 0 otherwise. An application of the TR model gave the results in Table 1. These results suggest that patients in the percutaneous group had a significantly worse initial health status than those in the surgical group and that the two groups seemed to have a significant difference in the drift of the health status. In particular, the health status for patients who utilized a surgically placed catheter tended to decline, but it improved over time for those who utilized a percutaneous placed catheter. Figure 1 displays the estimated survival functions of the time to first cutaneous exit site infection for the two groups. It can be seen that the obtained results are close to the Kaplan-Meier survival function estimates in Fig. 2 and indicate that the Wiener diffusion process assumption seems to be appropriate.

Table 1 Estimation results of the TR model for the kidney dialysis data

Parameter	Estimate	Standard error	p -value
α_0	1.4113	0.1435	<0.001
α_1	-1.0731	0.1891	<0.001
β_0	-0.0959	0.0765	0.210
β_1	0.6377	0.1280	<0.001

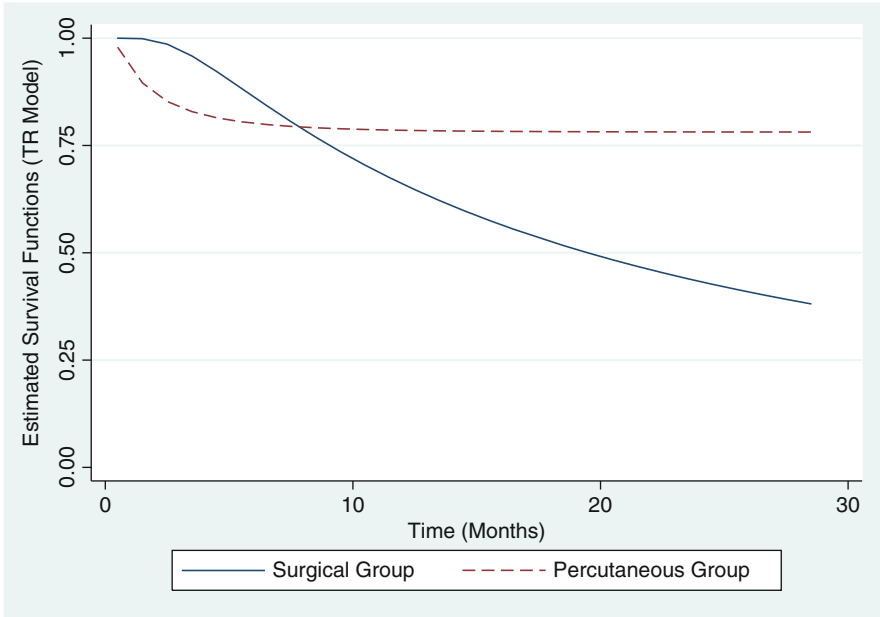


Fig. 1 Estimated survival functions by the TR model for the kidney dialysis data

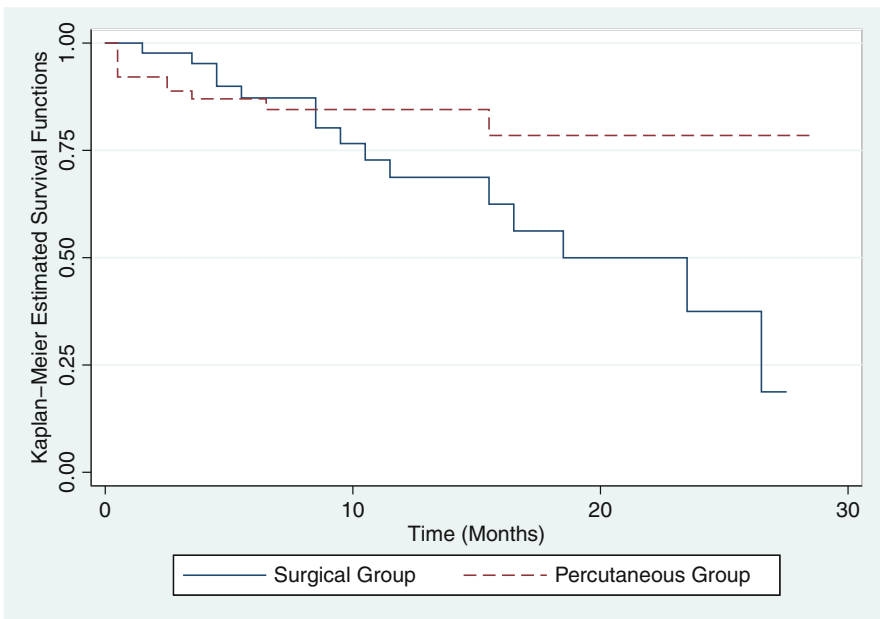


Fig. 2 Kaplan-Meier survival function estimates for the kidney dialysis data

Table 2 Estimation results of the TR model for the ovarian cancer data

Parameter	Estimate	Standard error	<i>p</i> -value
α_0	2.5252	0.2328	<0.001
α_1	0.4689	0.2916	0.108
β_0	0.0277	0.0193	0.151
β_1	-0.0725	0.0257	0.005

For comparison, AFT models with generalized gamma and log-logistic distributions were used to analyze the dataset. Note that the family of generalized gamma distributions includes exponential, Weibull, log-normal and gamma as its special cases and has considerable flexibility to characterize the shape of the underlying survival distribution [32]. The results suggest that none of the AFT models detected a significant group effect. Moreover, compared to the TR model, the AFT models cannot capture the cross-over pattern of the Kaplan-Meier survival function estimates.

Ovarian Cancer Dataset

In a study performed at the Mayo Clinic, a total of 35 patients with limited Stage II or IIIA ovarian carcinoma were divided into two groups based on grade of disease [3, 7, 9, 10]. Fifteen patients had low-grade or well-differentiated cancer, and 20 had high-grade or undifferentiated cancer. For each patient, the time to progression of disease (in days) was recorded. The main goal was to determine whether or not grade of disease was associated with time to progression of disease.

For the TR and AFT models, define z to be equal to 1 for patients with high-grade tumors and 0 for those with low-grade tumors. An application of the TR model gave the results in Table 2. These results suggest that patients with low-grade and high-grade tumors had a similar initial health status, but the health status for patients with high-grade tumors tended to decline more quickly than that for patients with low-grade tumors. The estimated survival functions based on the TR model, as shown in Fig. 3, indicate their close agreement with the Kaplan-Meier survival function estimates in Fig. 4. Although the AFT model with generalized gamma distribution gives a similar conclusion with respect to the significance of the group effect, none of the AFT models provide a good fit to the early crossing of survival curves.

Bile Duct Cancer Dataset

In a clinical trial performed at the Mayo Clinic, 47 patients with bile duct cancer were followed to determine whether a combination of radiation treatment (RöRx) and the drug 5-fluorouracil (5-FU) significantly prolonged patients' survival [3, 7, 10, 18]. The survival times (in days) were given for a group of 22 patients with the radiation-drug therapy and for a control group of 25 patients.

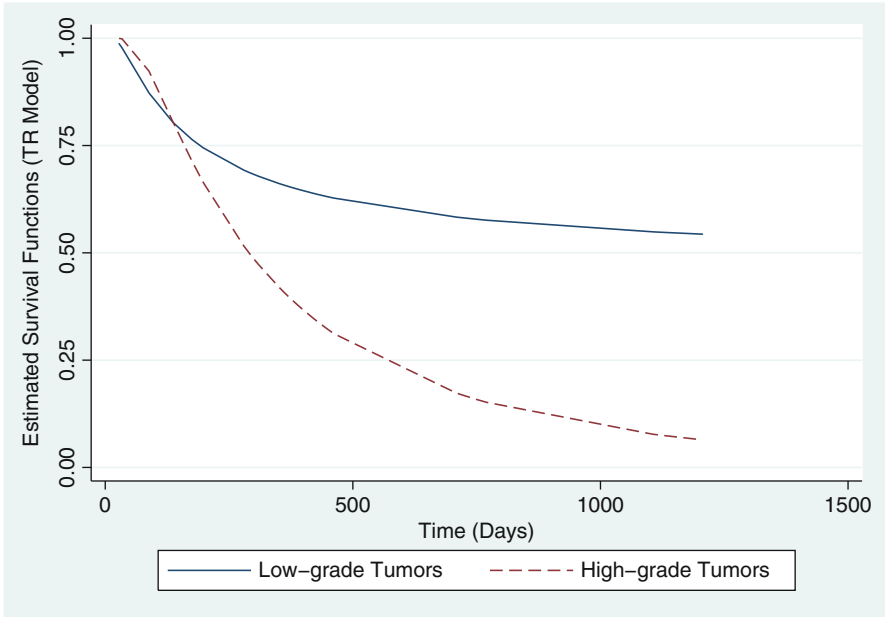


Fig. 3 Estimated survival functions by the TR model for the ovarian cancer data

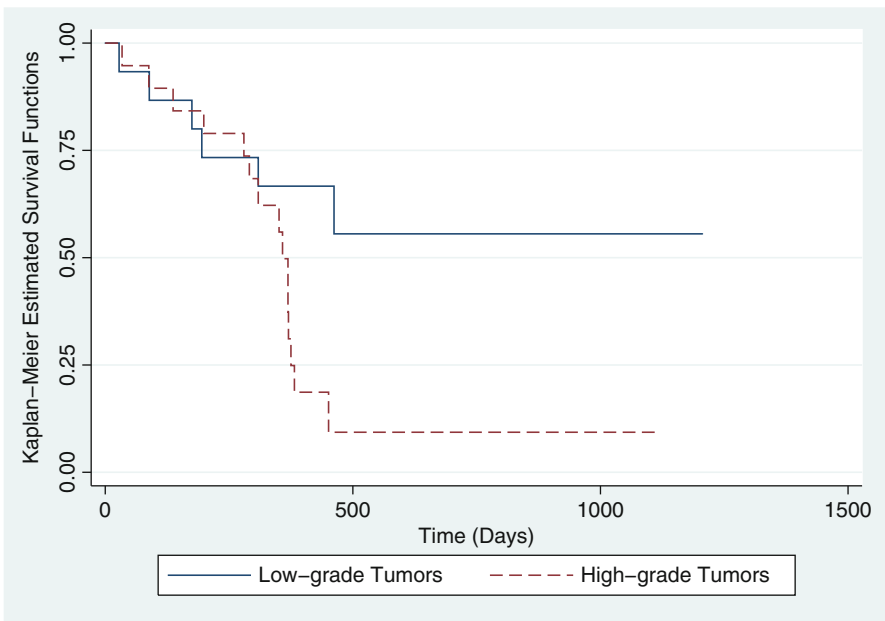


Fig. 4 Kaplan-Meier survival function estimates for the ovarian cancer data

Table 3 Estimation results of the TR model for the bile duct cancer data

Parameter	Estimate	Standard error	<i>p</i> -value
α_0	2.7219	0.1414	<0.001
α_1	0.2391	0.2106	0.256
β_0	-0.0415	0.0120	0.001
β_1	-0.0255	0.0215	0.235

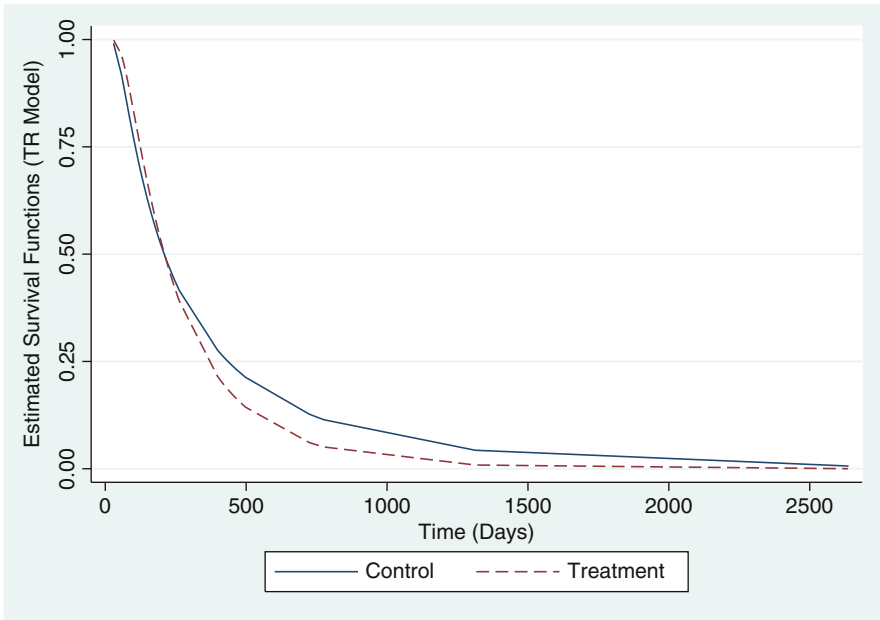


Fig. 5 Estimated survival functions by the TR model for the bile duct cancer data

To analyze the data, let z be 1 for treated and 0 for control patients. An application of the TR model gave the results in Table 3. These results suggest that there was no significant difference between treated and control patients in terms of the initial health status and its drift. A similar conclusion is given by fitting the corresponding AFT models with generalized gamma and log-logistic distributions. As shown in Fig. 5, the TR model successfully illustrates the crossing of the Kaplan-Meier estimated survival curves (Fig. 6). However, this pattern is not detected by any of the AFT models.

Discussion

In the preceding sections, we compared threshold regression and accelerated failure time models with respect to the underlying distribution of failure time. We showed that a large variety of AFT models can be derived from TR models by specifying appropriate baseline survival functions and running times.

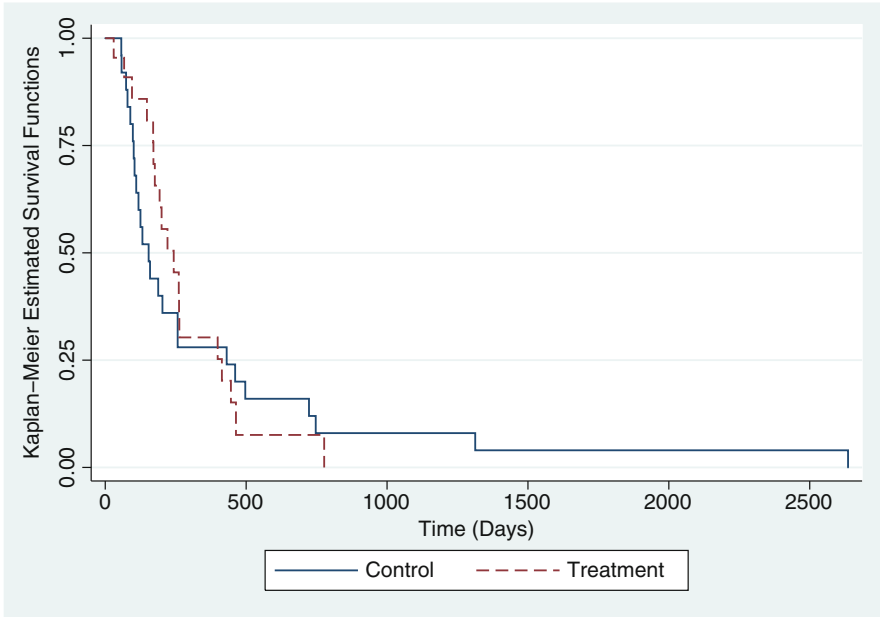


Fig. 6 Kaplan-Meier survival function estimates for the bile duct cancer data

In this article, we only focused on the comparison between threshold regression and accelerated failure time models in the parametric setting. The AFT model has been studied extensively in the literature for right censored data when the error distribution is completely unspecified. In general, there are two most commonly used semiparametric estimation procedures. One approach is the least squares based estimator [4, 11, 13, 17, 27], and the other is the rank based estimator [8, 12, 30, 31, 33]. Recently, some other approaches have been proposed for estimation and inference for the AFT model for current status and interval censored data [2, 24, 28, 29]. However, semiparametric extensions remain an open research issue for TR models.

Acknowledgements This research was supported in part by CDC/NIOSH grant OH008649 (Lee).

References

1. Aalen, O.O., Borgan, Ø., Gjessing, H.K.: *Survival and Event History Analysis: A Process Point of View*. Springer, New York (2008)
2. Betensky, R.A., Rabinowitz, D., Tsiatis, A.A.: Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88**, 703–711 (2001)
3. Breslow, N.E., Edler, L., Breger, J.: A two-sample censored-data rank test for acceleration. *Biometrics* **40**, 1049–1062 (1984)

4. Buckley, I.V., James, I.: Linear regression with censored data. *Biometrika* **66**, 429–436 (1979)
5. Duchesne, T., Lawless, J.: Alternative time scales and failure time models. *Lifetime Data Anal.* **6**, 157–179 (2000)
6. Duchesne, T., Rosenthal, J.S.: On the collapsibility of lifetime regression models. *Adv. Appl. Probab.* **35**, 755–772 (2003)
7. Fleming, T.R., O’Fallon, J.R., O’Brien, P.C., Harrington D.P.: Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36**, 607–625 (1980)
8. Fygensov, M., Ritov, Y.: Monotone estimating equations for censored data. *Ann. Stat.* **22**, 732–746 (1994)
9. Gill, R.D., Schumacher, M.: A simple test of the proportional Hazards assumption. *Biometrika* **74**, 289–300 (1987)
10. Hess, K.R.: Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat. Med.* **14**, 1707–1723 (1995)
11. Huang, J., Harrington, D.P.: Operating characteristics of partial least squares in right-censored data analysis and its application in predicting the change of HIV-I RNA. In: Nikulin, M., Commenges, D., Huber, C. (eds.) *Probability, Statistics, and Modelling in Public Health*, pp. 202–230. Springer, New York (2005)
12. Jin, Z., Lin, D.Y., Wei, L.J., Ying, Z.: Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–353 (2003)
13. Jin, Z., Lin, D.Y., Ying, Z.: On least-squares regression with censored data. *Biometrika* **93**, 147–161 (2006)
14. Kalbfleisch, J.D., Prentice, R.L.: *The Statistical Analysis of Failure Time Data*, 2nd edn. Wiley, New York (2002)
15. Klein, J.P., Moeschberger, M.L.: *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edn. Springer, New York (2003)
16. Kordonsky, K.B., Gertsbakh, I.: Multiple time scales and the lifetime coefficient of variation: engineering applications. *Lifetime Data Anal.* **3**, 139–156 (1997)
17. Lai, T.L., Ying, Z.: Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Stat.* **19**, 1370–1402 (1991)
18. Lawless, J.F.: *Statistical Models and Methods for Lifetime Data*, 2nd edn. Wiley, New York (2003)
19. Lawless, J., Crowder, M.: Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Anal.* **10**, 213–227 (2004)
20. Lee, M.-L.T., Whitmore, G.A.: Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Stat. Sci.* **21**, 501–513 (2006)
21. Lee, M.-L.T., Whitmore, G.A.: Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Anal.* **16**, 196–214 (2010)
22. Lee, M.-L.T., Whitmore, G.A., Laden, F., Hart, J.E., Garshick, E.: Assessing lung cancer risk in railroad workers using a first hitting time regression model. *Environmetrics* **15**, 501–512 (2004)
23. Lee, M.-L.T., Whitmore, G.A., Laden, F., Hart, J.E., Garshick, E.: A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. *J. Stat. Plan. Inferences* **139**, 1633–1642 (2009)
24. Murphy, S.A., Van der Vaart, A.W., Wellner, J.A.: Current status regression. *Math. Method. Stat.* **8**, 407–425 (1999)
25. Nahman, N.S., Middendorf, D.F., Bay, W.H., McElligott, R., Powell, S., Anderson, J.: Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *J. Am. Soc. Nephrol.* **3**, 103–107 (1992)
26. Oakes, D.: Multiple time scales in survival analysis. *Lifetime Data Anal.* **1**, 7–18 (1995)
27. Ritov, Y.: Estimation in linear regression model with censored data. *Ann. Stat.* **18**, 303–328 (1990)
28. Shen, X.: Linear regression with current status data. *J. Am. Stat. Assoc.* **95**, 842–852 (2000)

29. Tian, L., Cai, T.: On the accelerated failure time model for current status and interval censored data. *Biometrika* **93**, 329–342 (2006)
30. Tsiatis, A.A.: Estimating regression parameters using linear rank tests for censored data. *Ann. Stat.* **18**, 354–372 (1990)
31. Wei, L.J., Ying, Z., Lin, D.Y.: Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–851 (1990)
32. Yamaguchi, K.: Accelerate failure-time regression models with a regression model of surviving fraction: an application to the analysis of “permanent employment” in Japan. *J. Am. Stat. Assoc.* **87**, 284–292 (1992)
33. Ying, Z.: A large sample study of rank estimation for censored regression data. *Ann. Stat.* **21**, 76–99 (1993)

Residuals and Functional Form in Accelerated Life Regression Models

Stein Aaserud, Jan Terje Kvaløy, and Bo Henry Lindqvist

Abstract We study residuals of parametric accelerated failure time (AFT) models for censored data, with the main aim of inferring the correct functional form of possibly misspecified covariates.

Introduction

The accelerated failure time (AFT) regression model can be written

$$\log T = f(\mathbf{X}) + \sigma W, \quad (1)$$

where T is the event time; $\mathbf{X} = (X_1, \dots, X_p)$ is a vector of covariates; $f(\cdot)$ is some function determining the influence of the covariates; while σW is an “error” term. The parameter σ is here considered as a scale parameter, while W is assumed to have a fully specified ‘standardized’ distribution, such as the standard normal distribution; the standard Gumbel distribution for the smallest extreme (in which case T is Weibull-distributed); or the standard logistic distribution (see, e.g., Collett [2]).

The present short paper displays some main points from the preprint [5] and the master thesis [1], concerning the use of residuals to check model fit and to suggest functional form for covariates in AFT models. By the presented approach, this may alternatively be viewed as a search for a “best possible” additive model

S. Aaserud (✉) • B.H. Lindqvist
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: stein1618@gmail.com; bo@math.ntnu.no

J.T. Kvaløy
University of Stavanger, Stavanger, Norway
e-mail: jan.t.kvaloy@uis.no

of the form $\log T = f_1(X_1) + \dots + f_p(X_p) + \sigma W$ for functions $f_j(\cdot)$, $j = 1, \dots, p$, in the following called *covariate functions*. We hence seek to complement results and methods for the semiparametric Cox-model as earlier presented in Therneau et al. [6] and Grambsch et al. [3].

Residuals in AFT Models

Standardized residuals in AFT models are based on solving Eq. (1) for W . For given data $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$, where the δ_i are censoring indicators, we shall therefore define the standardized residuals by (\hat{s}_i, δ_i) , $i = 1, \dots, n$, where

$$\hat{s}_i = \frac{\log t_i - \hat{f}(\mathbf{x}_i)}{\hat{\sigma}}, \quad (2)$$

with $\hat{f}(\cdot)$, $\hat{\sigma}$ being appropriate estimators of the underlying f and σ , respectively. The Cox-Snell residuals are based on the fact that if T is a lifetime and $G(t) = P(T > t)$, then $-\log G(T)$ is unit exponentially distributed. Since for the AFT model, $P(T > t | \mathbf{X} = \mathbf{x}) = 1 - \Phi((\log t - f(\mathbf{x}))/\sigma)$, the Cox-Snell residuals are given as (\hat{r}_i, δ_i) , $i = 1, \dots, n$, with

$$\hat{r}_i = -\log(1 - \Phi(\hat{s}_i)), \quad (3)$$

where $\Phi(\cdot)$ is the distribution function of W . If the model is correctly specified, then the set of (\hat{r}_i, δ_i) is expected to behave similar to a censored sample of unit exponentially distributed variables.

When there are censored observations, a frequently used approach is to adjust the censored residuals by adding the expected residual “life” to the censored residuals and then proceed as if one has a complete set of uncensored observations. For Cox-Snell residuals one thus adds 1 to the censored residuals (see e.g. [2]), while for standardized residuals, adjusted residuals are obtained similarly by computations involving the distribution Φ (see [5]).

Let now X be a specific component of the covariate vector. One may want to plot the residuals versus this covariate. For censored survival data, such plots may, however, be misleading, and a possible remedy here is the use of the adjusted residuals which may work well when there are not too many censored values.

An alternative method which does not require the adjustment of censored residuals, is based on exponential regression smoothing, valid for continuous covariates. The idea is to consider a synthetic data set given as $(\hat{r}_1, \delta_1, x_1), \dots, (\hat{r}_n, \delta_n, x_n)$, where x_1, \dots, x_n are the values of the specific covariate X for the n observation units, respectively, where we impose the model for these data that \hat{r} given $X = x$ is exponentially distributed with hazard rate $\lambda(x)$. Then we use nonparametric exponential regression to estimate the function $\lambda(\cdot)$. Specific methods for nonparametric exponential regression are considered in, e.g., [4]. These include the so called covariate order method and local likelihood methods.

A residual plot versus X is now a plot of the points $(x_i, \log \hat{\lambda}(x_i))$, $i = 1, \dots, n$. The idea is that if the assumed model is correct, then the $\hat{\lambda}(x_i)$ should be close to 1, so $\log(\hat{\lambda}(x_i))$ should fluctuate around 0.

Functional Form for a Covariate

Suppose we want to conclude whether a specific covariate X is appropriately represented in our model. Assume that the correct model for the lifetime T is

$$\log T = \beta_0 + \beta' \mathbf{Z} + f(X) + \sigma W. \quad (4)$$

Based on data $\{(t_i, \delta_i, \mathbf{z}_i, x_i); i = 1, \dots, n\}$ we want to derive the appropriate form for $f(X)$ for the specific covariate X .

Suppose we fit by maximum likelihood the simpler linear model where $f(x) = \gamma x$. From this possibly misspecified model we use (2) to compute the standardized residuals

$$\hat{s}_i = \frac{\log t_i - \hat{\beta}_0 - \hat{\beta}' \mathbf{z}_i - \hat{\gamma} x_i}{\hat{\sigma}}. \quad (5)$$

and (3) to obtain the corresponding Cox-Snell residuals \hat{r}_i . In the following we show how these residuals can be used to infer the true form of $f(X)$.

From White [7] it follows that there are parameter values $(\beta_0^*, \beta^*, \gamma^*, \sigma^*)$ of the fitted model which are the limits (*a.s.*) of the estimators $(\hat{\beta}_0, \hat{\beta}, \hat{\gamma}, \hat{\sigma})$ as $n \rightarrow \infty$. In the model defined by $(\beta_0^*, \beta^*, \gamma^*, \sigma^*)$ we would compute the “theoretical” standardized residual as $S^* = (\log T - \beta_0^* - \beta^{*'} \mathbf{Z} - \gamma^* X) / \sigma^*$, which by inserting the true model (4) can be written

$$S^* = \frac{\sigma}{\sigma^*} W + \frac{(\beta_0 - \beta_0^*) + (\beta - \beta^*)' \mathbf{Z} + f(X) - \gamma^* X}{\sigma^*}. \quad (6)$$

Solving (6) for $f(X)$, taking the conditional expectation given $X = x$, and assuming that \mathbf{Z} and X are independent, gives that $f(x)$ is of the form

$$f(x) = \text{constant} + \gamma^* x + \sigma^* E(S^* | X = x).$$

Thus, modulo an unknown additive constant, we can estimate $f(x)$ by $\hat{f}(x) = \hat{\gamma} x + \hat{\sigma} \hat{H}(x)$, where $\hat{H}(x)$ is an estimate of $H(x) \equiv E(S^* | X = x)$.

If there are no censorings, we can use the standardized residuals \hat{s}_i from (5) and estimate the function $H(x)$ by smoothing the points (x_i, \hat{s}_i) ; $i = 1, \dots, n$. This can also be done with censored data if we adjust the residuals corresponding to censored observations in the way explained in section “Residuals in AFT Models”.

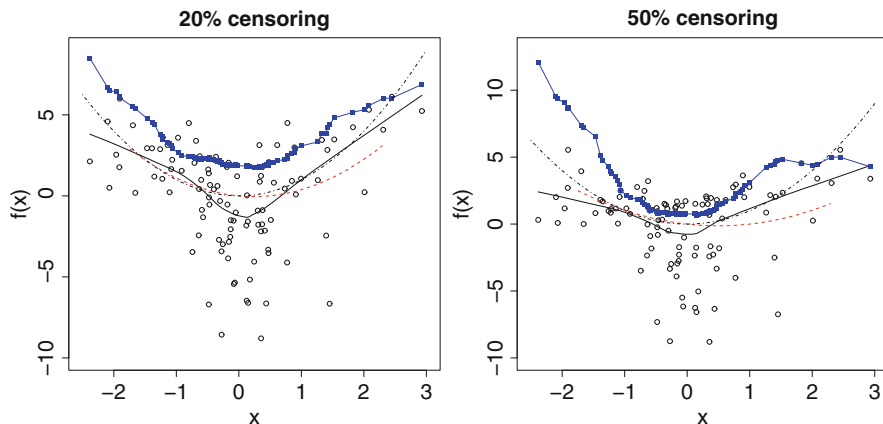


Fig. 1 Simulated Weibull distributed data. *Circles* are $(x_i, \hat{\gamma}x_i + \hat{\sigma}\delta_i)$ using the adjusted residuals; *solid line* is loess smooth from these points; *squares* are from censored exponential regression and the use of $\hat{H}(x)$; *dash-dot line* is the true quadratic function

Alternatively, we may use the Cox-Snell residuals \hat{r}_i to obtain smoothed estimates $\hat{\lambda}(x)$ as in section “Residuals in AFT Models”, in order to estimate the function $H(x)$. Note that by (3) we have $\hat{\delta}_i = \Phi^{-1}(1 - e^{-\hat{r}_i})$, where the Cox-Snell residuals \hat{r}_i are supposed to behave like exponentials with expected value $1/\hat{\lambda}(x_i)$. Thus we may estimate $H(x)$ by $\hat{H}(x) = \Phi^{-1}(1 - \exp(-1/\hat{\lambda}(x)))$. This avoids the use of adjusted residuals for censored observations.

Example (Simulated data from Weibull-distribution). We simulated $n = 100$ observations from the Weibull-distribution using the model

$$\log T_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + f(X_i) + \sigma W_i; \quad i = 1, \dots, 100,$$

where $\beta_0 = 0$, $\beta_1 = 5$, $\beta_2 = 0.2$, $f(x) = x^2$, $\sigma = 2$; the W_i were drawn from the Gumbel distribution of the smallest extreme, while the Z_{i1}, Z_{i2}, X_i were independently drawn from standard normal distributions. We imposed two different censoring scenarios by drawing independent censoring times C_i giving approximately 20 and 50% censoring, respectively.

Figure 1 shows the resulting estimates of the covariate function $f(x) = x^2$, using both a loess smoothing on the adjusted residuals, and a censored nonparametric exponential regression using the nonadjusted residuals. A possible conclusion from this and similar datasets is that there are no large differences in the estimates of the covariate function $f(X)$ for low censoring, while for more heavy censoring the nonparametric exponential regression method seemingly performs slightly better.

References

1. Aaserud, S.: Residuals and functional form in accelerated life regression models. Master thesis. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim (2011)
2. Collett, D.: Modelling Survival Data in Medical Research. Chapman & Hall/CRC, Boca Raton (2003)
3. Grambsch, P.M., Therneau, T.M., Fleming, T.R.: Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51**, 1469–1482 (1995)
4. Kvaløy, J.T., Lindqvist, B.H.: Estimation and inference in nonparametric Cox models: time transformation methods. *Comput. Stat.* **18**, 205–221 (2003)
5. Lindqvist, B.H., Aaserud, S., Kvaløy, J.T.: Residual plots for model checking and for revealing functional form of covariates in accelerated life regression models. Statistics preprint. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim (2012)
6. Therneau, T.M., Grambsch, P.M., Fleming, T.R.: Martingale-based residuals for survival models. *Biometrika* **77**, 147–160 (1990)
7. White, H.: Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25 (1982)

Neyman, Markov Processes and Survival Analysis

Grace Yang

Abstract J. Neyman used stochastic processes extensively in his applied work. One example is the Fix and Neyman (F-N) competing risks model (Fix and Neyman, Hum Biol 23(30):205–241, 1951) that uses finite homogeneous Markov processes to analyse clinical trials with breast cancer patients. We revisit the F-N model, and compare it with the Kaplan-Meier (K-M) formulation for right censored data. The comparison offers a way to generalize the K-M formulation to include risks of recovery and relapses in the calculation of a patient’s survival probability. The generalization is to extend the F-N model to a nonhomogeneous Markov process. Closed-form solutions of the survival probability are available in special cases of the nonhomogeneous processes, like the popular multiple decrement model (including the K-M model) and Chiang’s staging model, but these models do not consider recovery and relapses while the F-N model does. An analysis of sero-epidemiology current status data with recurrent events is illustrated. Fix and Neyman used Neyman’s RBAN (regular best asymptotic normal) estimates for the risks, and provided a numerical example showing the importance of considering both the survival probability and the length of time of a patient living a normal life in the evaluation of clinical trials. The said extension would result in a complicated model and it is unlikely to find analytical closed-form solutions for survival analysis. With ever increasing computing power, numerical methods offer a viable way of investigating the problem.

G. Yang (✉)

Department of Mathematics, University of Maryland, College Park, MD 20742, USA
e-mail: gly@math.umd.edu

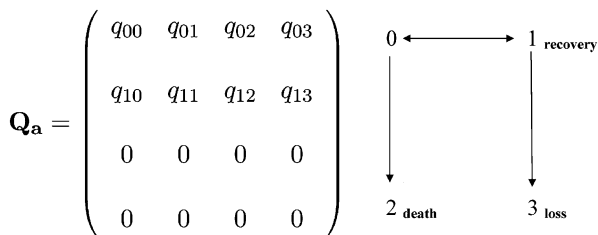
Introduction

J. Neyman used stochastic processes in his applied work extensively, particularly Markov processes since the late forties. Examples include the use of birth and death processes to study tumour growth, Markov branching processes to study radiation effects on single cells, a discrete time branching process with a disperse function to model an epidemic, and others. When doing applied work, Neyman typically would first construct stochastic models for the data and then develop inference procedures for data analysis. In this presentation, we shall revisit the Fix-Neyman (F-N) competing risks model which was introduced in a joint publication of Fix and Neyman [11] titled *A simple stochastic model of recovery, relapse, death and loss of patients*. The paper models the disease development of a patient in a clinical trial using a Markov process. The model is constructed for comparing survival times and quality of life of patients who have undergone different treatments of breast cancer. We shall compare the calculation of a patient's survival probability in the F-N competing risks model with that of the Kaplan-Meier (K-M) formulation [16]. By way of comparison, it is seen that the F-N competing risks model offers a very general mathematical model system for tackling many of the problems that arise in survival analysis. An example of the analysis of sero-epidemiology survey data or current status data will be presented. Markov models, of course, have been used in various contexts in survival analysis, for example the popular multiple decrement model of competing risks. What distinguishes the F-N model is the introduction of the relapse and recovery of a breast cancer patient in the calculation of her survival probability. Fix and Neyman used a system of Kolmogorov equations of transition probabilities as the basic tool. This presentation is focused on the F-N model and its extension. Except for citing a few references in the concluding section, no attempt is made to review the current work on recurrent events in survival analysis.

The Fix-Neyman Competing Risks Model

The F-N model is a 4-state Markov process $\{\xi_t : t \geq 0\}$, where ξ_t describes the status of a patient at time t , and the four states are S_0 (original state of the patient being under treatment for cancer), S_1 apparent recovery from cancer, S_2 death from the treatment of cancer, and S_3 lost to follow-up. It is worth noting that in applications, the original state S_0 may be variously defined according to the way a clinical trial is analyzed. It could be the state of the time of diagnosis of cancer or time of entering the clinical trial or others. Clearly the selection of the initial state, S_0 , is important. It will affect the interpretation of other states. Over time (t), a patient moves back and forth between the states of recovery (S_1) and relapse (S_0) until she is either lost to follow-up or enters the absorbing state of death (Fig. 1).

Fig. 1 The Fix-Neyman model with transition paths



For ease of notation, we shall denote the states S_0, S_1, S_2, S_3 by 0, 1, 2, 3 respectively. An individual who is in state i at time s and will be in state j at time t for $s < t$ is governed by the transition probabilities

$$P_{ij}(s, t) = P[\xi_t = j | \xi_s = i], \quad \text{for } 0 \leq s < t, i, j = 0, 1, 2, 3$$

and it is assumed that as $t \rightarrow s, P_{ij}(s, t) \rightarrow 1$ if $i = j$ and $P_{ij}(s, s) \rightarrow 0$ if $i \neq j$.

The transition probabilities in the F-N model are generated by the matrix \mathbf{Q}_a of constant risks q_{ij} , where $q_{ii} = -\sum_{j \neq i} q_{ij}$, for $i, j = 0, \dots, 3$.

In \mathbf{Q}_a there are two transient states 0, 1 and two absorbing states 2, 3. Regarding the selection of the initial state 0, ([11], p. 210) states “initial state with some specific definition as visualized by Berkson, for example, the state of being under treatment for cancer”. Throughout their paper, Fix and Neyman acknowledged the consultations with Joe Berkson on medical questions and the acquisition of some of their clinical trial data. Berkson was then the Chief of the Division of Biometry and Medical Statistics of the Mayo Clinic. According to Berkson, there was very little loss of patients from state 0 (the state of being under treatment for cancer). Fix and Neyman therefore set the transition rates $q_{03} = 0$ and also $q_{12} = 0$. Mathematically, there is no difficulty in deducing a patient’s survival probability if q_{03} and q_{12} were positive because the F-N model is a homogeneous Markov process. Further discussion of the transition probabilities is in section “Extension of the Fix-Neyman Competing Risks Model”.

Note that we have interchanged the labels of the states 1 and 2 in the F-N paper for convenience of matrix presentation. That is, our states 1 and 2 correspond respectively to states 2 and 1 in the F-N paper.

A distinct feature of the F-N model is the inclusion of the possibility of recovery and relapse of patients in the calculation of a patient’s survival probability. Fix and Neyman used breast cancer data from some clinical trials to estimate the risks in the 4-state Markov process. To emphasize the presence of all these competing risks in \mathbf{Q}_a , Fix and Neyman called anything that has to do with this model *crude*. Therefore the transition probabilities $P_{ij}(s, t)$ derived from \mathbf{Q}_a are crude probabilities. Likewise, a patient’s survival probability or its estimate calculated directly under model \mathbf{Q}_a is crude.

Fix and Neyman were after the *net* survival probability of a patient when the risk of loss to follow-up (q_{13}) is *eliminated*. This will be discussed in the next section.

Comparison of the Fix-Neyman Model with the Kaplan-Meier Formulation

Kaplan and Meier did not use Markov processes to compute patient’s survival probability, but their underlying model can be described as a three-state Markov process generated by the risk matrix $\mathbf{Q}_c(t)$ in Fig. 2. The matrix $\mathbf{Q}_c(t)$ can be obtained by eliminating State 1 (no recovery) from \mathbf{Q}_a and assuming that the transition intensities depend on t .

Risk matrix $\mathbf{Q}_c(t)$ has two absorbing states; death and loss to follow-up or other causes. There are no risks of recovery and relapse in model $\mathbf{Q}_c(t)$. In this case, the state of an individual at any time t , ξ_t , will be in one of the three states $\{0, 1, 2\}$.

Throughout the article we shall denote the state of an individual at time t by ξ_t and her/his lifetime by X . The probability distributions of ξ_t and X change with the underlying model. We shall denote the survival probability of X calculated from model \mathbf{Q}_a by $P_a[X > t]$, similarly a subscript indicating a specific model employed is attached to all other probability calculations. However, to reduce cumbersome notations, we shall not attach the model symbol to each component of a risk matrix. What risks we are referring to will be made clear in the context of the discussion. For example, indication of model \mathbf{Q}_c in the following Eq. (2) suffices.

Model $\mathbf{Q}_c(t)$ is identical to the usual Kaplan-Meier formulation in that the survival time of a patient, X , is subject to right censoring by an independent positive random variable, C . The observation on a patient is a pair (Z, δ) where $Z = \min(X, C)$, $\delta = I[X \leq C]$. Suppose that X has hazard rate $q_{02}(t)$ and C has hazard rate $q_{03}(t)$. To be precise, we shall assume $q_{02}(t)$ and $q_{03}(t)$ are continuous functions in t and satisfy the hazard function requirement of $\int_0^\infty q_{02}(t)dt = \infty$ and same for $q_{03}(t)$.

Under the K-M formulation the joint distribution of Z and δ is given by

$$P[Z \leq t, \delta = 1] = P[X \leq t, X \leq C] = \int_0^t q_{02}(u) e^{-\int_0^u (q_{02}(v)+q_{03}(v))dv} du \quad (1)$$

where $P[Z \leq t, \delta = 0]$ is similarly calculated.

The probability in (1) is exactly the following transition probability calculated from model $\mathbf{Q}_c(t)$,

$$P_{02,c}(0, t) = P_c[\xi_t = 2 | \xi_0 = 0] = \int_0^t q_{02}(u) e^{-\int_0^u (q_{02}(v)+q_{03}(v))dv} du \quad (2)$$

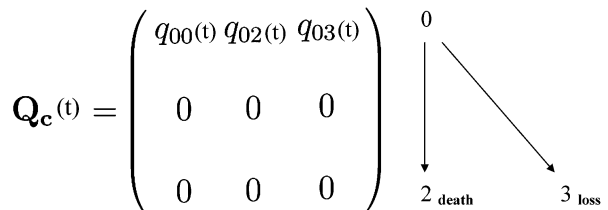


Fig. 2 The Kaplan-Meier model with transition paths

where the initial state $\xi_0 = 0$ at time 0 must match with that of the K-M formulation under which the survival time X in (1) is measured. For example, if the problem is to study the survival times of a group of patients after surgical removal of tumours, we can set state 0 as the state of post-surgery of a living patient and the initial time 0 is the instance immediately after the surgery. The meaning of states 2 and 3 are indicated in the path diagram in Fig. 2.

From either model $\mathbf{Q}_c(t)$ or the distribution (1), the survival probability of X can be derived. We can also use matrix $\mathbf{Q}_d(t)$ which is obtained by eliminating state 3 from $\mathbf{Q}_c(t)$ (See Fig. 4). For the purpose of comparison, it is convenient to call $\mathbf{Q}_c(t)$ the K-M model. Then the survival probability of the K-M model is

$$P_d[X > t] = P_d[\xi_t = 0 \mid \xi_0 = 0] = P_{00,d}(0,t) = e^{-\int_0^t q_{02}(v)dv}. \tag{3}$$

This appears to be a roundabout way of doing things. However, due to right censoring, the distribution of X cannot be modelled directly by $\mathbf{Q}_c(t)$. Deleting the risk of loss to follow-up, $q_{03}(t)$, provides an easier way to calculate the survival probability especially when the number of states in the Markov model gets larger as we shall see in the F-N model.

It is important to remember that when eliminating a state from a risk matrix, the values on the diagonal of the new matrix change so that each row sum of the new risk matrix remains to be zero. For example, in $\mathbf{Q}_c(t)$, $q_{00}(t) = -(q_{02}(t) + q_{03}(t))$. But in $\mathbf{Q}_d(t)$, $q_{00}(t) = -q_{02}(t)$.

The F-N model is a 4-state homogeneous Markov process in which the risks, q_{ij} , are independent of time, and q_{03} and q_{12} are assumed to be zero. The K-M formulation, on the other hand, corresponds to a 3-state nonhomogeneous Markov process with unspecified time-dependent risks $q_{ij}(t)$. The K-M formulation is a special case of the multiple decrement model with two competing risks. With respect to statistical inference, the F-N model is used in parametric analyses while the K-M estimator is for nonparametric analyses.

Both Fix and Neyman and Kaplan and Meier were after the elimination of the *loss to follow up or other causes* in the estimation of a patient's survival probability. Without elimination, the estimated survival probability would be biased and the treatment comparisons would be inappropriate because patterns of loss to follow-up may vary from one clinical trial to another.

Then if state 3 (loss to follow up) is eliminated, the F-N model \mathbf{Q}_a reduces to \mathbf{Q}_b in Fig. 3 and the K-M model $\mathbf{Q}_c(t)$ reduces to $\mathbf{Q}_d(t)$ in Fig. 4. While Kaplan and Meier used the pair $(\mathbf{Q}_c(t), \mathbf{Q}_d(t))$, Neyman and Fix used the pair $(\mathbf{Q}_a, \mathbf{Q}_b)$ to deduce and estimate the survival probability.

The (net) survival probability in the F-N model is given by

$$P_b[X > t] = 1 - P_b[\xi_t = 2 \mid \xi_0 = 0] = 1 - P_{02,b}(0,t), \tag{4}$$

where to be definitive, we assume that the initial state is 0.

$$\mathbf{Q}_b = \begin{pmatrix} q_{00} & q_{01} & q_{02} \\ q_{10} & q_{11} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{array}{c} 0 \xleftrightarrow{\hspace{1cm}} 1 \text{ recovery} \\ \downarrow \\ 2 \text{ death} \end{array}$$

Fig. 3 Elimination of loss to follow up in the Fix-Neyman model

Fig. 4 Elimination of loss to follow up in the Kaplan-Meier model

$$\mathbf{Q}_b^{(t)} = \begin{pmatrix} q_{00(t)} & q_{02(t)} \\ 0 & 0 \end{pmatrix} \quad \begin{array}{c} 0 \\ \downarrow \\ 2 \text{ death} \end{array}$$

This is parallel to the survival probability (3) computed from $\mathbf{Q}_a(t)$.

Although one can calculate the net survival probability directly from \mathbf{Q}_b , the estimation of the net survival probability, (4), requires the estimation of q_{01}, q_{02} and q_{10} in \mathbf{Q}_b which cannot be done directly. Their estimates are taken to be the corresponding estimates of q_{ij} in \mathbf{Q}_a .

Fix and Neyman deduced an explicit formula for the survival probability (4) by way of calculating the transition probability $P_{02,a}(0, t)$ in \mathbf{Q}_a . It is given by

$$P_{02,a}(0, t) = q_{02} \left(\alpha_1 + \frac{q_{10} + q_{13}}{\lambda_1 \lambda_2} \left[1 - \left(\frac{1}{2} \right) (q_{01} + q_{02} + q_{10} + q_{13}) \alpha_1 - \alpha_2 \right] \right) \quad (5)$$

where $-\lambda_1$ and $-\lambda_2$ are the two non zero eigenvalues of the matrix \mathbf{Q}_a ,

$$\lambda_1 = \frac{1}{2} \left(q_{01} + q_{02} + q_{10} + q_{13} - \sqrt{(q_{01} + q_{02} - q_{10} - q_{13})^2 + 4q_{01}q_{10}} \right),$$

$$\lambda_2 = \frac{1}{2} \left(q_{01} + q_{02} + q_{10} + q_{13} + \sqrt{(q_{01} + q_{02} - q_{10} - q_{13})^2 + 4q_{01}q_{10}} \right)$$

and

$$\alpha_1 = \frac{e^{-\lambda_1 t} - e^{-\lambda_2 t}}{\lambda_2 - \lambda_1}, \quad \alpha_2 = \frac{1}{2} \left(e^{-\lambda_1 t} + e^{-\lambda_2 t} \right).$$

The eigenvalues $(-\lambda_{1,b}, -\lambda_{2,b})$ of \mathbf{Q}_b can be obtained by setting $q_{13} = 0$ in λ_1 , and λ_2 , likewise in α_1 and α_2 . They are denoted by $\lambda_{1,b}, \lambda_{2,b}, \alpha_{1,b}$ and $\alpha_{2,b}$ respectively. Setting $q_{13} = 0$ in (5) and replacing $\lambda_1, \lambda_2, \alpha_1$ and α_2 by $\lambda_{1,b}, \lambda_{2,b}, \alpha_{1,b}$

and $\alpha_{2,b}$ respectively, an explicit formula for the probability $P_{02,b}(0,t)$ in (4) is obtained, and hence the (net) survival probability,

$$\begin{aligned}
 P_b[X > t] &= 1 - P_{02,b}(0,t) \\
 &= 1 - q_{02} \left(\alpha_{1,b} + \frac{q_{10}}{\lambda_{1,b}\lambda_{2,b}} \left[1 - \left(\frac{1}{2} \right) (q_{01} + q_{02} + q_{10})\alpha_{1,b} - \alpha_{2,b} \right] \right). \quad (6)
 \end{aligned}$$

Neyman’s RBAN (regular best asymptotic normal) estimates (1949) were used for estimating the risks in \mathbf{Q}_a with the breast cancer data. This will be discussed in section “Neyman’s Method of Minimum Modified χ^2 ”.

In complete parallel, Kaplan and Meier estimated $\mathbf{Q}_c(t)$ with a sample of n independent right-censored survival times, where a right-censored survival time is defined previously (above Eq. (1)). The estimate of $q_{02}(t)$, $\hat{q}_{02}(t)$ obtained in the $\mathbf{Q}_c(t)$ model is employed to construct an estimate of the survival probability in (3). One estimator could be $\exp(-\int_0^t \hat{q}_{02}(v)dv)$. Another is the K-M product-limit estimator.

In the Kaplan-Meier formulation the failure rate $q_{02}(t)$ is unspecified. If $q_{02}(t) = \lambda_0(t)e^{\beta z}$, then the Cox-regression (or proportional hazard rates) model with right censoring is a 3-state Markov chain with two absorbing states as specified by $\mathbf{Q}_c(t)$.

There is no particular advantage of using the Markov model to obtain the K-M estimator of the survival probability, because $\mathbf{Q}_c(t)$ and $\mathbf{Q}_a(t)$ are very simple risk matrices. Reformulating the Kaplan-Meier model as a Markov process is to show that one could extend in the direction of the popular K-M model to include relapses and recovery events as in the Fix-Neyman model \mathbf{Q}_a , but with some time-dependent $q_{ij}(t)$. As the number of states increases and recurrences are allowed, the theory of Markov processes provides important analytical tools for survival analysis. For many diseases, recovery and relapse could be significant occurrences in the course of disease development and treatment. Breast cancer is one example as studied by Fix and Neyman. Other examples abound. A leukemia patient might recover from a bone marrow transplant and later experience a relapse leading to the need for further transplants. A patient with aplastic anemia (an auto-immune disease) is usually treated with an immune-suppressant (IST). Some patients will respond to IST and relapse several months later. The same patient may receive a 2nd IST and so on and so forth until death or loss to follow up. It is likely that a patient’s survival time could be affected by such recurring recovery – relapse events. It is desirable to include the available data on recovery and relapse in the survival analysis.

Extension of the Fix-Neyman Competing Risks Model

The F-N model assumes constant risks for q_{ij} . While the assumption leads to a closed-form solution for the survival probability, it would be more realistic to include some time-dependent risks, $q_{ij}(t)$, as pointed out by Fix and Neyman.

However that could pose mathematical challenges in solving a finite system of Kolmogorov equations of transition probabilities given below:

$$\frac{dP_{ij}(s,t)}{dt} = \sum_{l \neq j} P_{il}(s,t)q_{l,j}(t) + P_{ij}(s,t)q_{j,j}(t), \quad \text{for all states } i, j \quad (7)$$

with initial conditions $P_{ij}(s,t) = 1$ if $i = j$, 0 otherwise.

The derivation can be found in Feller ([10], Vol. I, Chap. 17, 2nd edn.) under the conditions that for $0 \leq s < t$, (i) $P_{ii}(s,t) \rightarrow 1$ as $t \rightarrow s$; (ii) for each j , there is a non negative continuous function $-q_{jj}(t)$ such that

$$\lim_{h \rightarrow 0} \frac{1 - P_{jj}(t, t+h)}{h} = -q_{jj}(t);$$

(iii) for each pair of i, j with $i \neq j$ there is a non negative continuous function $q_{ij}(t)$ such that

$$\lim_{h \rightarrow 0} \frac{P_{ij}(t, t+h)}{h} = q_{ij}(t).$$

Only special cases of (7) have been solved explicitly. Otherwise, we rely on numerical solutions. This is a system of forward equations. We shall not dwell on the details of the Eq. (7) and refer the reader to a standard reference, Feller ([10], Vol. I, 2nd edn.), and Feller [9] for the existence and uniqueness of the solution.

If q_{ij} are independent of t , the transition probabilities $P_{ij}(s,t)$ simplify to $P_{ij}(t) = P[\xi_t = j \mid \xi_0 = i]$ with s setting equal to 0. The solution is given by

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \quad \text{for } t \geq 0. \quad (8)$$

where $\mathbf{P}(t)$ is a matrix with components $P_{ij}(t)$ and initial condition $\mathbf{P}(0) = \mathbf{I}$, an identity matrix, and \mathbf{Q} is the corresponding risk matrix with components q_{ij} .

Several of Neyman's students continued the work in this direction. In particular, B. Altshuler [3] and C. L. Chiang [8]. Altshuler considered time-dependent $q_{ij}(t)$ in the multiple decrement model and obtained nonparametric estimates of survival probabilities which were later studied by Aalen in a seminal paper [1] based on his PhD thesis (1975). Aalen used an entirely different approach based on counting processes and Le Cam's LAN theory. Chiang (Sect. 7 of Chap. 11, [8]) proposed a staging (or illness-death) model for analyzing survival times of a patient with a chronic disease. It is assumed that the disease progresses from a mild stage S_0 to a severe stage through intermediate stages $\{S_1, \dots, S_{k-1}\}$ and the patient may enter the death state S_k from each of these stages as shown in the transition paths in Fig. 5. The transition rates from S_i to S_j are given in the matrix \mathbf{Q}_g .

This staging model has been used for studying HIV and other chronic diseases. Chang et al. [7] developed a statistical test of goodness of fit for a 3-state ($k = 2$) staging models. These authors formulated the problem in terms of counting processes and developed an asymptotic test of a Markov staging model versus a semi-Markov model with power calculations.

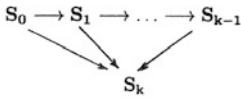
$$\mathbf{Q}_g = \begin{pmatrix} q_{00} & q_{01} & 0 & \cdots & q_{0k} \\ 0 & q_{11} & q_{12} & \cdots & q_{1k} \\ 0 & 0 & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & q_{k-1,k-1} & q_{k-1,k} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$


Fig. 5 Chiang’s staging model

In Chiang’s staging model the transition rates of the transient states are constant and *one-directional* where $q_{ij} = 0$ if $i > j$. Thus the components of \mathbf{Q}_g below the diagonal are zero. This special feature permits a closed-form solution to the corresponding Kolmogorov equations (7) with time-dependent $q_{ij}(t)$. There are no risks of recovery and relapses in the model. At the request of a reviewer, we shall provide the solution for the nonhomogeneous \mathbf{Q}_g . We shall assume that for $j = 0, \dots, k - 1$, $q_{j,j+1}(t) > 0$ and $q_{jk}(t) > 0$. Also $\int_0^\infty q_{j,j+1}(v)dv = \infty$ and $\int_0^\infty q_{jk}(v)dv = \infty$ as required of a hazard function of a random variable.

We shall set the initial time $s = 0$ and assume the initial state of a patient at time 0 is $\xi(0) = 0$. We write the transition rates as $q_{ij}(t)$ to state explicitly their dependence on time. The Kolmogorov equations (7) for model \mathbf{Q}_g are

$$\frac{dP_{0j,g}(0,t)}{dt} = P_{0,j-1,g}(0,t)q_{j-1,j}(t) + P_{0j,g}(0,t)q_{jj}(t)$$

for $0 \leq j \leq k - 1$. (9)

with initial conditions

$$\begin{aligned}
 P_{0j,g}(0,0) &= 1 \quad \text{if } j = 0 \\
 &= 0 \quad \text{if } j \neq 0,
 \end{aligned}$$

and $q_{jj}(t) = -(q_{j,j+1}(t) + q_{jk}(t))$ for $j = 0, \dots, k - 1$.

For consistency we use in (9) the symbol $P_{ij,g}(0,t)$ to denote the transition probability under model \mathbf{Q}_g .

A patient can enter the death state k from any one of the states $\{0, 1, \dots, k - 1\}$. Therefore the survival time X of a patient is

$$X = \inf\{t > 0 : \xi(t) = k\}.$$

(10)

The patients survival probability is given by

$$\begin{aligned}
 P_g[X > t] &= P_g[\xi(t) \in \{0, 1, \dots, k - 1\} \mid \xi(0) = 0] \\
 &= \sum_{j=0}^{k-1} P_{0j,g}(0,t) \quad \text{for } t > 0.
 \end{aligned}$$

(11)

Use Eq. (9) to solve for $P_{0j,g}(0,t)$. Starting with $P_{00,g}(0,t)$, $P_{0j,g}(0,t)$ can be solved recursively for $j = 0, 1, \dots, k-1$. For $j = 0$,

$$P_{00,g}(0,t) = e^{\int_0^t q_{00}(u) du} \quad (12)$$

In what follows we put

$$\mu_j(t) = e^{-\int_0^t q_{jj}(u) du}, \quad \text{for } j = 0, 1, \dots, k-1.$$

The equation for $P_{01,g}(0,t)$ is a first order linear equation,

$$\frac{dP_{01,g}(0,t)}{dt} = P_{00,g}(0,t)q_{01}(t) + P_{01,g}(0,t)q_{11}(t). \quad (13)$$

The solution of $P_{01,g}(0,t)$ is given by

$$P_{01,g}(0,t) = \frac{\int_0^t \mu_1(v)P_{00,g}(v)q_{01}(v)dv}{\mu_1(t)}. \quad (14)$$

Substituting (12) for $P_{00,g}(0,v)$ in (14), we obtain an explicit solution for $P_{01,g}(0,t)$. By the same token, for $j = 1, \dots, k-1$, the solution of $P_{0j,g}(0,t)$ is given by

$$P_{0j,g}(0,t) = \frac{\int_0^t \mu_j(v)P_{0,j-1,g}(0,v)q_{j-1,j}(v)dv}{\mu_j(t)}, \quad (15)$$

with the initial conditions stated in (9).

Chiang ([8], Chap. 11.7) derives the survival probability for constant q_{ij} . The above is a generalization to the nonhomogeneous case.

A different line of attack is to use product integrals. Aalen and Johansen [2] express the transition probabilities for finite nonhomogeneous Markov processes in terms of product integrals. Andersen et al. ([4], p. 312) is one of the very few publications that put recovery – relapse of a disease in a nonhomogeneous Markov model. A three-state nonhomogeneous Markov model is used to study survival time of patients with liver cirrhosis where loss to follow up is not considered in the model (which may not be needed for this particular study). An individual is at any time t in one of the three states: having normal prothrombin level (0), having abnormal prothrombin level (1) and death (2). The transition rates $q_{01}(t)$, $q_{10}(t)$, $q_{12}(t)$, and $q_{02}(t)$ are assumed to be positive where $q_{21}(t)$ and $q_{20}(t)$ are of course zero. The product integral representation facilitates the estimation of transition rates nonparametrically but the estimation requires the data on the exact time of the direct transition of each patient from one state to another. Many of these direct transitions are difficult to observe if possible at all. To overcome the data problem, Andersen et al. defined changes of states to take place at the time when patients are examined at follow-up visits to the hospital and obtained the estimates $\hat{q}_{ij}(t)$

of $q_{ij}(t)$. These estimates are used to estimate the survival probability $P[X > t]$ of a patient which in their notation is equal to $1 - P_{02}(0, t)$. No explicit analytical solution for $P_{02}(0, t)$ is provided. Andersen et al. used numerical solutions to obtain an estimate $\hat{P}_{02}(0, t)$, known as the Aalen-Johansen (A-J) estimate. Figures IV.4.15 and IV.4.16 on pages 315–316 [4] show discrepancies in the estimated survival curves and standard deviations between the A-J estimate and the K-M estimate. In our interpretation, the discrepancies could be attributed to the fact that these probabilities were estimated using two different models. Although both are Markov models, the K-M model $\mathbf{Q}_c(t)$ allows no recovery – relapse while the model for A-J estimate does. It is interesting to note that in the treated group, the A-J estimate of the survival curve is larger up to the 4th year, then the K-M estimate is larger. For the placebo group, the A-J estimate appears to be uniformly worse than that of the K-M estimate. The estimated standard deviations of the A-J estimates of the survival probabilities are nearly always smaller than that of the K-M estimates. The recovery rate $q_{10}(t)$ seems to have played a role in the treatment effect. The statistical significance of the result is not known.

Yang and Chang [21] used recurrent events in a parametric analysis to study prevalence of hepatitis A antibodies. This will be discussed in the next section.

An Example of a Nonhomogeneous Competing Risks Model with Application to Cross-Sectional Surveys of Hepatitis A Antibody

Cross-sectional surveys are conducted in many areas of science, including epidemiology, demography and construction of *current life tables*. The sampling design of such surveys is to collect current status data of individuals at a fixed point in time (in actual practice, the collection is often carried out in a very short period in time). Thus the exact time of a direct transition of a sampled subject from one state to another is not observable. In our case of using a Markov process $\{\xi_t, t \geq 0\}$ to model the status of an individual over time, survey data are severely censored. Current status data appear in a variety of forms depending on applications and models employed. A brief comparison of survey data with case 2 interval-censored data studied in the literature will be made at the end of the section after the discussion of the hepatitis A example.

Our example is taken from sero-epidemiology surveys in populations for studying age-specific prevalence of antibody to hepatitis A virus (anti-HAV) [19]. This information is useful for understanding the spread of infection like infection rate and age-dependent characteristics in a population. Parametric models such as logistic distributions were used but did not fit the age-specific prevalence data of anti-HAV [19]. Yang and Chang [21] used a 3-state nonhomogeneous Markov competing risks model to derive and estimate the age-specific prevalence of anti-HAV. The model offers an explanation of a well-known phenomenon of the decline

$$\mathbf{Q}_e(t) = \begin{pmatrix} q_{00}(t) & q_{01}(t) & q_{02}(t) \\ q_{10}(t) & q_{11}(t) & q_{12}(t) \\ 0 & 0 & 0 \end{pmatrix}$$

Fig. 6 Age-specific prevalence model with loss and regain of immunity

in anti-HAV in older ages of individuals. The model successfully isolated the confounding factors of mortality and diminished immunity of individuals. Without elimination of the competing risks of death and diminished immunity, the estimates of the prevalence of HAV infection would be biased.

The 3-state nonhomogeneous Markov process $\{\xi_t, t \geq 0\}$ describes the process of an individual acquiring and loss of antibody of hepatitis A over his or her lifetime, where ξ_t represents the status of an individual at age t . This is a conceptual model in that the exact age (t) of transitions are not observable. The process $\{\xi_t, t \geq 0\}$ however is partially observable from the survey data which amounts to taking a snapshot of the process at some fixed t . The precise connection between the survey data and the underlying Markov model will be given below. For the purpose of model building, *imaging* that an individual can be monitored from birth to death and the age at which he acquires anti-HAV could be recorded if he ever acquires it in his lifetime. With respect to detectability of antibodies, an individual, ξ_t , at any age t is in one of the three possible states $\{0, 1, 2\}$, where $0 =$ alive with no detectable anti-HAV, $1 =$ alive with anti-HAV, $2 =$ deceased. Transitions with recurrences between state 0 and 1 in the model allow the possibility that an individual acquires the antibody, then at a later age the titer falls below a detectable level, and subsequently through reinfection or boost, the titer rises above a detectable level prior to his death. The risk matrix of the process $\{\xi_t, t \geq 0\}$ is given in Fig. 6.

Model $\mathbf{Q}_e(t)$ is similar to \mathbf{Q}_b but with different interpretations of the competing risks.

Cross-sectional sampling is to sample from a living population. The age-specific prevalence $\theta(t)$ of an individual at age t is therefore the conditional probability that $\xi_t = 1$ given that he is alive at age t . (The death of an individual censors the observation of prevalence!)

Under model \mathbf{Q}_e with recurrences of loss and regain immunity, the age-specific prevalence at age t is given by

$$\theta(t) = P[\xi_t = 1 \mid \xi_t = 0 \text{ or } 1] = \frac{P_{01,e}(0,t)}{P_{00,e}(0,t) + P_{01,e}(0,t)}. \tag{16}$$

To obtain an explicit expression for $\theta(t)$ one must first solve the differential equation (7) for the transition probabilities $P_{ij,e}(s,t) = P[\xi_t = j \mid \xi_s = i]$ for $i, j = 0, 1, 2$, assuming risk matrix to be \mathbf{Q}_e .

Equations in (7) corresponding to \mathbf{Q}_e are second-order differential equations with variable coefficients $q_{ij}(t)$. Under a mild condition specific to the characteristics of HAV, analytical solutions for $P_{01,e}(s,t)$ and $P_{00,e}(s,t)$ can be obtained and are given by

$$P_{01,e}(s,t) = \exp\left(\int_s^t [q_{11}(v) - q_{01}(v)]dv\right) \times \int_s^t q_{01}(v) \exp\left(\int_s^v [q_{10}(y) + q_{01}(y)]dy\right) dv, \quad (17)$$

$$P_{00,e}(s,t) = \exp\left(\int_s^t [q_{00}(v) + q_{01}(v)]dv\right) - \exp\left(\int_s^t [q_{11}(v) - q_{01}(v)]dv\right) \times \int_s^t q_{01}(v) \exp\left(\int_s^v [q_{10}(y) + q_{01}(y)]dy\right) dv. \quad (18)$$

Substituting the solutions of $P_{01,e}(s,t)$ and $P_{00,e}(s,t)$ in $\theta(t)$ in (16), the age-specific prevalence at any age t can be calculated explicitly.

$\theta(t)$ is used to fit the age-specific prevalence survey data from seven European countries.

The age-specific prevalence model $\theta(t)$ has several interesting features.

1. If there is no decline in antibody, i.e., $q_{10}(t) = 0$, then the model $\theta(t)$ in (16) reduces to a distribution function $F(t)$,

$$\theta(t) = F(t) = 1 - \exp\left(-\int_0^t q_{01}(v)dv\right).$$

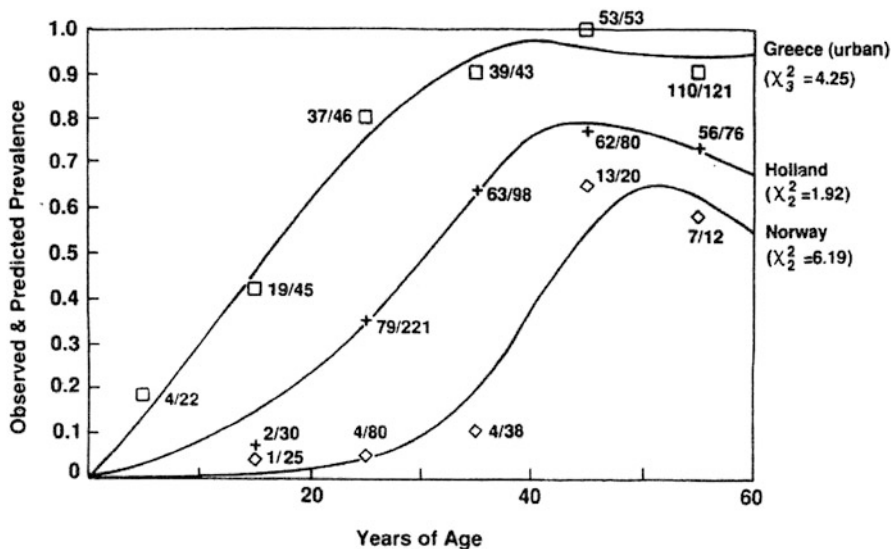
Then whatever be the chosen model for the risk $q_{01}(t)$, the age-specific prevalence model $F(t)$ will always be nondecreasing and will not produce ups and downs as in the observed prevalence data. In particular, the logistic model, i.e.,

$$q_{01}(t) = \frac{1}{1 + \exp(\alpha + \beta t)}$$

would not fit the observed prevalence data.

2. Examining $\theta(t)$ shows that $\mathbf{Q}_e(t)$ cannot be a homogeneous Markov process, for it will result in an increasing function for $\theta(t)$ in t . An increasing $\theta(t)$ is inconsistent with the observed prevalence curves. This rules out constant values for competing risks q_{ij} . Or at least either q_{01} or q_{10} should be age-dependent in some manner.
3. The survey data (shown in Fig. 7 as proportions) were taken from Frösner et al. [12]. The survey was conducted by cross-sectional sampling of healthy individuals (that is free of liver disease) in the population some time in 1976 which is set equal to zero in the calculation for Fig. 7. Serum specimens collected from selected individual were tested for the presence of HAV antibody. The data

A STOCHASTIC MODEL FOR PREVALENCE SURVEYS



Observed values y_k/n_k of anti-HAV and model estimates of $\theta(t)$ in urban Greece, Holland, and Norway.

Fig. 7 Age-specific prevalence model with loss and regain of immunity. Symbols in the figure are explained in item 3 (The author thanks *Mathematical Biosciences* for the permission to reproduce the figure from Yang and Chang [21])

were grouped into K age intervals with varying width. The available data are proportions $\frac{y_k}{n_k}$ of individuals that are anti-HAV positive in the age interval k . Each age interval spans a number of years and the mid-point of each age interval, t_k , represents (approximates) the age of individuals in the k th interval. So y_k is the number of anti-HAV positives in a sample of size n_k taken from the subpopulation of age t_k . With this simplification, y_k is a binomial random variable $B(n_k, \theta_k)$ where θ_k , computed from $\theta(t)$ in (16) as

$$\theta_k = \frac{P_{01,e}(-t_k, 0)}{P_{00,e}(-t_k, 0) + P_{01,e}(-t_k, 0)}. \tag{19}$$

A shift of t in the above formula is necessary because the time of the survey in 1976 is set equal to zero. To evaluate θ_k , Yang and Chang [21] use the logistic model specified in item 1 for $q_{01}(t)$ with two unknown parameters α and β , and $q_{10}(t)$ is modelled by a constant risk starting at age 40 which is known to be about the time when an individual begins to lose immunity. Therefore,

$$q_{10}(t) = \lambda \quad \text{for } t \geq 40 - t_k$$

$$= 0 \quad \text{for } t < 40 - t_k$$

where λ is an unknown parameter which is denoted by ξ in Eq.(3.4) of Yang and Chang [21]. The maximum likelihood estimates of the three parameters, α , β and λ were obtained by using the likelihood function:

$$L = \prod_{k=1}^K \theta(t_k)^{y_k} [1 - \theta(t_k)]^{n_k - y_k} . \tag{20}$$

(Note that the y_k were sampled from subpopulations of age t_k and hence are independent.)

Replacing the unknown parameters in $\theta(t)$ by their respective maximum likelihood estimates computed from (20) yields a predicted age-specific prevalence curve as shown in Fig. 7, computed for each of the three countries. The observed proportions are denoted respectively by the squares, crosses and diamonds for Greece, Holland and Norway. The goodness of fit of the model was measured by the chi-squared values on the right margin. The estimated values of the parameters and their covariances matrices are given in Table 1 of Yang and Chang [21].

4. Allowing the recurrences of regain and diminishing immunity (i.e. $q_{10}(t) > 0$) into the model greatly increases the complexity of the model $\theta(t)$. However, the complexity is unavoidable for producing a more realistic model.
5. It is important to obtain explicit analytical solution for $\theta(t)$. Otherwise, it would be difficult to see how different risks are affecting the age-specific prevalence in a global manner as noted in (1), (2), and (3).
6. Remark on interval censoring and current status data: An important problem in survival analysis is to determine the time to a specific event. A typical example is the time, Y , of a patient with a progressive chronic disease to enter a particular state, say α . The information on Y is collected at a patient's checkup times say, $\tau_1 < \tau_2, \dots$. From such information one cannot determine the exact value of Y except for knowing that it will be either in a time interval $(\tau_i, \tau_{i+1}]$, for some $i = 1, 2, \dots$ or before τ_1 or after the last checkup time. Such interval-censored data on Y are referred to as current status data. In particular Y is called case 2 interval-censored if there are only two checkup times $\tau_1 < \tau_2$ (observable and possibly random). Markov multistate models have been widely employed in the analysis of interval-censored data, see, e.g., Andersen and Keiding [5], Banerjee [6], and books by Kalbfleish and Prentice [15] and Sun [20]. The cross-sectional data in the hepatitis example are collected only at one point in time which corresponds to τ_1 and the sampling is conditioned on individuals being alive at the time of survey. Let b denote the birth time (calendar time) of an individual whose lifetime is X . Suppose that the survey time is τ_1 (a calendar time). Then the current age and the residual lifetime of this individual at time τ_1 are respectively $\tau_1 - b$ and $b + X - \tau_1$. An individual will be included in the population for survey if and only if

$$-(b - \tau_1) \leq 0 \leq b + X - \tau_1. \quad (21)$$

This is equivalent to the event $[\xi_{\tau_1-b} = 0, 1]$, where the current age $\tau_1 - b$ is the age of an individual at the time of survey, see (16). Because of recurrences in model \mathbf{Q}_e , an individual can be in and out of state 1 a random number of times. Let U be the last time (calendar time) before τ_1 that an individual enters into state 1 and v be his sojourn time in state 1 since U . We then have $\xi_{\tau_1-b} = 1$ if and only if (21) holds and τ_1 is contained in the random interval

$$U \leq \tau_1 \leq \min(U + v, b + X)$$

where U , τ_1 and b are measured in calendar time. Or equivalently

$$U - \tau_1 \leq 0 \leq \min(U + v - \tau_1, b + X - \tau_1).$$

We set $U = \infty$ if this individual is not infected at the survey time. In this sense, we might say that U is conditionally interval-censored but it is not exactly the case 2 (discussed above). Here U is neither a stopping time nor observable. Furthermore, U is age-dependent. The conditioning event $[-(b - \tau_1) \leq 0 \leq b + X - \tau_1]$ reminds us of the random truncation model.

Neyman's Method of Minimum Modified χ^2

Fix and Neyman provided a method of parameter estimation and some qualitative discussion of the data, but did not carry out the actual data analysis in the paper. Neyman's RBAN (regular best asymptotic normal) estimates were proposed for the risks in \mathbf{Q}_a . The RBAN estimation was introduced by Neyman as BAN estimation for multinomial distributions in a paper presented at the first Berkeley Symposium in 1945 and appeared in the Proceedings of the Symposium in 1949. The paper shows that under certain restrictions, estimates obtained by methods such as minimum χ^2 , modified minimum χ^2 and maximum likelihood are all RBAN. Having a rather sparse data set, Fix and Neyman felt that the method of minimum modified χ^2 (23) appears to be the only feasible one for their estimation of risks q_{ij} for all i and j . Several sections of the F-N paper are devoted to the discussion of RBAN estimates which we shall outline below.

Fix and Neyman considered two estimation problems. One is a patient's (net) survival probability in (4) and the other is the expected length of time a patient lives a normal life (in state 2) under model \mathbf{Q}_b . The expected length of normal life during an observation period $[0, T]$ is defined by

$$e_{01} = \int_0^T P_{01,b}(0, u) du \quad (22)$$

where $P_{01,b}(0,u)$ is defined in (4) and computed in (6). Here for illustration we arbitrarily choose the initial state of a patient to be state 0.

For both estimation problems it is necessary to estimate the four transition rates q_{ij} , for $i \neq j$, in the model \mathbf{Q}_a . The authors considered two different approaches: (i) the development of an estimation method with the follow-up data they have, and (ii) the development of an estimation method with the follow-up data they would like but do not have. The authors knew it is impracticable to collect the ideal data on the exact time of direct transitions between any pair of states. What they would like to have is “the information about the duration of some specified phases in the fate of at least a part of the individuals considered”. For instance the number of individuals originally in state 0 who at the conclusion of the period of observation are still in state 0 without having had any period of normal life. Approach (ii) serves as a suggestion for designing future follow-up studies. Both (i) and (ii) are about finding RBAN estimates. But the computation in (ii) is considerably more complex. We shall use approach (i) to convey the general idea of their estimation method.

Fix and Neyman ([11], pp. 230–231) organized their follow-up data as follows. Suppose that the clinical trial started at time, say 0 and the period of the observation is of length T . Initially, there are N_0 persons in state 0 and N_1 persons in state 1. The available data are N_0, N_1 and the number N_{ij} of individuals who initially are in state i and are found in state j at time T , for $i = 0, 1$ and $j = 0, 1, 2, 3$. Let $\phi_{ij} = N_{ij}/N_i$. The model counter part of the relative frequency ϕ_{ij} in (23) is the transition probability $P_{ij,a}(0, T)$. The RBAN estimates of the four risks q_{01}, q_{10}, q_{13} , and q_{02} in \mathbf{Q}_a are obtained by minimizing the modified χ^2 as given by

$$\chi^2 = N_0 \sum_{j=0}^3 \frac{(P_{0j}(0, T) - \phi_{0j})^2}{\phi_{0j}} + N_1 \sum_{j=0}^3 \frac{(P_{1j}(0, T) - \phi_{1j})^2}{\phi_{1j}} \quad (23)$$

under appropriate side conditions.

Note that ϕ_{ij} is not the relative frequency of the number of direct transitions from i to j as nowadays often used in nonparametric survival analysis. In Fix and Neyman’s data, the direct transitions between any pair of distinct states i and j are not available.

To emphasize the importance of a patient living a normal life, Fix and Neyman produced a numerical example (p. 222 of the F-N paper) showing that a larger survival probability results in a shorter duration of a patient living a normal life. This example illustrates that not only the survival probability but also the duration of living a normal life should be considered together for evaluation of a clinical trial.

A question arises as to why Neyman introduced the notion of RBAN when the MLE is believed to be asymptotically efficient without any restrictions. Many authors tried to prove this property but succeeded for restricted classes of estimates only, and Neyman proved it for the RBAN class. Around 1950, the issue was settled by the counterexamples produced by J. Hodges (unpublished), which by now are well-known. The reader is referred to Le Cam [17] for an interesting historical note on the development of the asymptotic theory of estimation in that period.

In choosing the method of minimum modified χ^2 , Fix and Neyman reworded the definition of the RBAN in the context of their competing risks model. It is easier for us to define the RBAN estimates by a brief sketch of Hodges's counterexample first as given in the F-N paper. Let \hat{p}_n be the relative frequency of the number of successes in n iid Bernoulli trials. We know that \hat{p}_n is the MLE of the probability of success, p , in one trial, and that for any $0 < p < 1$ the sequence, $\sqrt{n}(\hat{p}_n - p)$, tends to the normal distribution $N(0, \sigma^2(p) = p(1 - p))$ as n tends to infinity. By modifying the sequence \hat{p}_n for some values of n , Hodges constructed a competing sequence of estimates \tilde{p}_n of p such that $\sqrt{n}(\tilde{p}_n - p)$ converges to a normal distribution $N(0, \tau^2(p))$, where the asymptotic variance $\tau^2(p)$ is strictly smaller than $p(1 - p)$ for some p , otherwise equals to $p(1 - p)$. Therefore, contrary to popular beliefs, the sequence of MLE \hat{p}_n is not asymptotic efficient. In keeping with the usual definition of asymptotic efficiency, the estimating sequence such as \tilde{p}_n is called a super efficient estimating sequence.

Neyman's conditions for RBAN estimates are stated in two parts. Consider a class C of estimates that are functions of the observed relative frequencies ϕ_{ij} for all i, j having the following properties:

- (i) (Exclusion of super efficient estimates) every estimate in C does not explicitly depend on the sample size n ; (Hodges' example makes it clear.)
- (ii) (Linearity) every estimate in C (as a function of ϕ_{ij}) has continuous partial derivatives of the first order with respect to ϕ_{ij} for all i, j ;
- (iii) (Consistency) as n tends to infinity, every estimate in C converges in probability to the true parameter for which it estimates.

Under conditions (i), (ii) and (iii), the following property (iv) holds:

- (iv) (Asymptotic normality) every estimate $\hat{\theta}$ of parameter in class C has asymptotic normal distribution, i.e. $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \sigma^2)$, for some variance σ^2 .

Neyman called the estimates in class C "regular asymptotic normal estimates" (RAN). A "best regular asymptotic normal estimate" (RBAN) is an estimate whose asymptotic variance does not exceed any other RAN estimate in class C . Conditions for existence of RBAN estimates are given in Neyman [18]. The asymptotic properties of RBAN estimates are the same and can be obtained by several methods including the MLE, minimum χ^2 and minimum modified χ^2 . Finding MLE becomes increasingly challenging when the model gets more complex. Fix and Neyman felt the minimum modified χ^2 is the only feasible method of finding RBAN estimates with their limited data. It would be interesting in survival analysis to compare properties of different estimation methods for finite samples.

Concluding Remarks

We have revisited the Fix-Neyman competing risks model from the perspective of the Kaplan-Meier estimator with a focus on recovery and relapses of a disease. We made no attempt to review the literature on recurrences in survival analysis except for a brief remark and citing a few publications. Many of the early publications assume that recurrent events are not terminated by death or failure. This assumption is different from that of the F-N model. Effort has been made in more recent years to include death in modelling recurrent events. For instance, Y. Huang and Wang [13] and C-Y Huang and Wang [14] introduced various bivariate models to analyze the number of recurrent events up to the time of failure. The problems studied by these authors seem different from that of Fix and Neyman. Fix and Neyman investigated how a patient's survival probability is affected by relapses and recoveries. The interested reader is referred to Y. Huang and Wang [13], C-Y Huang and Wang [14] and references therein.

The F-N framework offers a unified and general model system for analyzing survival data in parametric as well as nonparametric analysis. The F-N model itself needs to be extended to include time-dependent risks and thereby it generalizes the fundamental K-M estimator to include recoveries and relapses in the calculation of a patient's survival probability. One such extension exists in the product-integral representation of Aalen and Johansen [2] of the transition probabilities. However the application of the product-integral representation requires the data on the exact times of direct transitions between states which are not always available, as seen in the example by Andersen et al. discussed in section "Extension of the Fix-Neyman Competing Risks Model". The Fix and Neyman approach of solving a system of Kolmogorov equations can accommodate the situation when the data are relatively sparse and not as informative as the direct transition counts. Clinical trial data are often of this kind and so are the survey data. The extension we have in mind is to follow Fix and Neyman's approach to develop parametric methods. Many clinical trials have data on recovery and relapses. Account for these data in the survival analysis would hopefully strengthen the findings in the treatment comparisons in clinical trials.

Finding an explicit form of the survival probability (or the probability of first passage time to death) requires solving a system of Kolmogorov equation (7) for nonhomogeneous Markov processes. In general closed-form solutions are not available except in special cases. For more complicated patterns of competing risks, numerical solutions offer a viable method for investigating the problem.

Acknowledgements G. Yang was supported in part by NSF grant DMS 1209111. The author thanks the referees and the Associate Editor for their helpful comments and suggestions that led to an improved presentation and an extension of the original manuscript. The original version of this article has previously been published in *Lifetime Data Analysis* in 2013.

References

1. Aalen, O.O.: Non-parametric inference for a family of counting processes. *Ann. Stat.* **6**, 701–726 (1978)
2. Aalen, O.O., Johansen, S.: An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Stat.* **5**, 141–150 (1978)
3. Altshuler, B.: Theory for the measurement of competing risks in animal experiments. *Math. Biosci.* **6**, 1–11 (1970)
4. Andersen, P.K., Borgan, O., Gill, R.R., Keiding, N.: *Stat Models Based on Counting Processes*. Springer, New York (1993)
5. Andersen, P.K., Keiding, N.: Multi-state models for event history analysis. *Stat. Methods. Med. Res.* **11**, 91–115 (2002)
6. Banerjee, M.: Current status data in the twenty first century: some interesting developments. In: Chen, D-G., Sun, J., Peace, K.E. (eds.) *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman and Hall/CRC Biostatistics Series, London (2011)
7. Chang, I-S., Chuang, Y-C., Hsiung, C.A.: Goodness-of-fit tests for semi-Markov and Markov survival models with one intermediate state. *Scand. J. Stat.* **28**, 505–520 (2001)
8. Chiang, C.L.: *Introduction to Stochastic Processes in Biostatistics*. R.E. Krieger, New York (1980)
9. Feller, W.: On the integrodifferential equations of purely discontinuous Markoff processes. *Trans. Am. Math. Soc.* **48**, 488–515 (1940)
10. Feller, W.: *An Introduction to Probability Theory and Its Applications*, vol. I, 2nd edn. Wiley, New York (1966)
11. Fix, E., Neyman, J.: A simple stochastic model of recovery, relapse, death and loss of patients. *Hum. Biol.* **23**(3), 205–241 (1951)
12. Frösner, G.G., Papaevangelou, G., Butler, R., Iwarson, S., Lindhol, A., Courouze-Pauty, A., Haas, H., Deinhardt, F.: Comparison of prevalence data in different age groups. *Am. J. Epidemiol.* **110**, 63–69 (1979)
13. Huang, Y., Wang, M-C.: Frequency of recurrent events at failure time: modeling and inference. *J. Am. Stat. Assoc.* **98**, 663–670 (2003)
14. Huang, C-Y., Wang, M-C.: Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics* **61**(2), 392–402 (2005)
15. Kalbfleisch, J.D., Prentice, R.L.: *The Statistical Analysis of Failure Time Data*, 2nd edn. Wiley, New York (2002)
16. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958)
17. Le Cam, L.: On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates, vol. 1, no. 11, pp. 277–330. University of California Publication in Statistics, California (1953)
18. Neyman, J.: Contribution to the theory of the χ^2 test. In: *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 239–273. University of California Press, Berkeley (1949)
19. Schenzle, D., Dietz, K., Frösner, G.G.: Antibody against hepatitis A in seven European countries. *Am. J. Epidemiol.* **110**, 63–69 (1979)
20. Sun, J.: *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York (2006)
21. Yang, G.L., Chang, M.: A stochastic model for prevalence of hepatitis A antibody. *Math. Biosci.* **98**, 157–169 (1990)

Quantiles of Residual Survival

Christopher Cox, Michael F. Schneider, and Alvaro Muñoz

Abstract In reliability theory, the lifetime remaining in a network of components after an initial run-in period is an important property of the system. Similarly, for medical interventions residual survival characterizes the subsequent experience of patients who survive beyond the beginning of follow-up. Here we show how quantiles of the residual survival distribution can be used to provide such a characterization. We first discuss properties of the residual quantile function and its close relationship to the hazard function. We then consider parametric estimation of the residual quantile function, focusing on the generalized gamma distribution. Finally, we describe an application of quantiles of residual survival to help describe the effects at the population level of the introduction and sustained use of highly active antiretroviral therapy for the treatment of HIV/AIDS.

Introduction

In many applications, the comparison of two survival distributions is summarized by the estimation of a single relative hazard, under the standard proportional hazards assumption. We have previously argued [4], along with others, that this

C. Cox (✉)

Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St E7642,
Baltimore, MD 21205, USA

e-mail: ccox@jhsph.edu

M.F. Schneider

Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St E7012,
Baltimore, MD 21205, USA

e-mail: mschneid@jhsph.edu

A. Muñoz

Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St E7648,
Baltimore, MD 21205, USA

e-mail: amunoz@jhsph.edu

assumption is frequently violated, and that estimation of selected quantiles of the two distributions and their comparison by relative times (relative quantiles) can be not only more appropriate but also more informative.

Patients returning for a follow-up visit after an initial diagnosis or treatment often want to know what to expect in the future. This information, contained in the conditional distribution of residual survival times, is an important metric for evaluating the long term effects of interventions. In reliability theory, mean residual life has been extensively studied, and is sometimes used instead of the hazard function to characterize families of survival distributions. Here, we argue that the use of quantiles of the residual survival distribution is a useful alternative.

We first discuss properties of the residual quantile function, including its close relationship to the hazard function. We show that, like mean residual life, the residual quantile function has the opposite shape from the hazard. More importantly, the residual quantiles provide useful information about residual survival that is not apparent in the behavior of the hazard. An advantage of the residual quantiles is that they are expressed in units of time, which is the natural metric of survival. Using the generalized gamma (GG) family, which we have previously advocated as a platform for parametric survival analysis [4], we then discuss estimation of the residual quantile function from the parametric perspective. In this case, estimation of the residual quantile function for selected quantiles such as the median and quartiles is relatively straightforward, and can be accomplished using standard statistical software. Finally, using data from two multicenter cohort studies, we consider an application in which absolute and relative residual twenty-fifth percentiles are used to assess the effect at the population level of the introduction and continued use of highly active antiretroviral therapy (HAART) for the treatment of HIV/AIDS.

Residual Survival

Basic Properties

Consider a random lifetime T with survival function $S(t)$, distribution function $F(t)$, density $f(t)$ and hazard $h(t)$. For simplicity we assume that the survival function is strictly decreasing and always positive, so that it has a well-defined inverse with percentile function $t(p) = F^{-1}(p) = S^{-1}(1 - p)$. Residual life is the lifetime remaining given survival to the present, i.e., to a particular time $w > 0$. The residual survival distribution is most naturally defined by its survival function.

$$S(t|w) = P(T - w > t | T > w) = \frac{S(w+t)}{S(w)} \quad t \geq 0 \quad (1)$$

The residual hazard function can be simply described in terms of the original hazard, $h(t|w) = h(w + t)$.

The mean residual life function, whose properties have been extensively studied, is the integral of $S(t|w)$.

$$m(w) = \int_0^\infty \frac{S(w+t)}{S(w)} dt = \frac{1}{S(w)} \int_w^\infty S(t) dt = E(T - w | T > w) \tag{2}$$

Its derivative is related to the hazard function by $m'(w) = h(w)m(w) - 1$, and $m(w)$ therefore determines the underlying survival distribution ([14], Section 1.B.g). It also follows that $m'(w) \geq -1$, i.e., the function $m(w) + w$ is increasing.

Since survival times are typically right-skewed, the mean is perhaps not the most useful summary measure. Instead, quantiles such as the median have been considered more relevant for understanding the distribution of survival times. The p th quantile of the residual survival distribution is

$$t(p, w) = S^{-1}[(1-p)S(w)] - w = t[p + (1-p)F(w)] - w = t[1 - (1-p)S(w)] - w \tag{3}$$

This relationship is illustrated in Fig. 1, which shows how the residual quantiles are determined by the $1 - S(w) + pS(w)$ quantiles of the underlying survival distribution. In particular $t(p,0) = t(p)$ is the p th quantile of this distribution, and for any fixed $w > 0$, $t(p,w)$ is a strictly increasing function of p with $t(0,w) = 0$ and $t(1^-,w) = \infty$. It follows that the function $t(w|p) + w$ is also increasing. As with mean residual life, we consider $t(p,w)$ as a function of $w \geq 0$, in this case for fixed values of p , and we therefore denote the residual p th quantile function by $t(w|p)$, for example, $t(w|0.5)$ is the residual median function.

The residual p th quantile function, particularly the residual median, has been the subject of a number of studies. It is well known for example that for any given $0 < p < 1$, $t(w|p)$ determines $S(w)$ only up to a periodic function with period $-\log(1-p)$ [10]. In addition a number of papers have focused on nonparametric estimation of $t(w|p)$ [1, 5, 7, 9, 12].

The behavior of the residual quantile function can be characterized by computing its derivative. Using the well-known result that the derivative of the p th quantile function is the reciprocal of the density evaluated at the p th quantile, it follows that

$$t'(w|p) = \frac{(1-p)f(w)}{f\{t[1-(1-p)S(w)]\}} - 1 = \frac{h(w)}{h\{t[1-(1-p)S(w)]\}} - 1 = \frac{h(w)}{h[t(w|p) + w]} - 1 \tag{4}$$

Thus, as with mean residual life, the derivative of the residual quantile function is closely related to the underlying hazard. Equation 4 characterizes the reciprocity of $t(w|p)$ and $h(w)$. Specifically, if the hazard function is increasing (respectively, decreasing) then the residual quantile function is decreasing (increasing). These results were derived for the residual median by Lillo [13] but they clearly hold for any $0 < p < 1$.

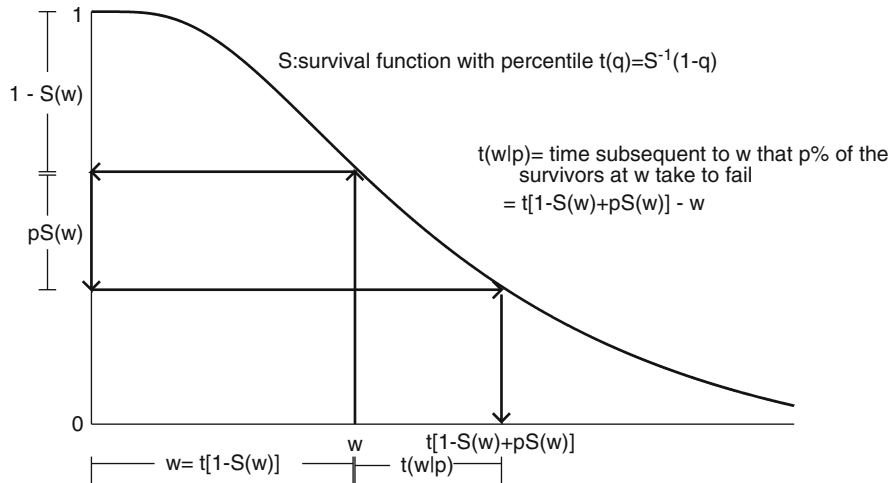


Fig. 1 Definition of the residual p th percentile after w in terms of the percentile function of the underlying distribution. Following the flow determined by the arrows starts at w and ends at $t(w|p) + w = t[1 - S(w) + pS(w)]$

In addition, in the [appendix](#) we show that for arch-shaped hazards, the residual quantile function either has a bathtub shape or is always increasing (Lemma 4). Similarly, for bathtub-shaped hazards the residual quantile function either has an arch shape or is always decreasing. In particular, for arch-shaped hazards satisfying $h(0) = 0$ the residual quantile function always has a bathtub shape, while for a bathtub-shaped hazard satisfying $h(0) = \infty$, the residual quantile function must have an arch shape. These two conditions are satisfied by the parametric families discussed in the next section. Of course if the hazard has multiple local extreme points then the situation may be more complicated; our interest is in parametric families where this does not occur. It is worth noting that the mean residual life also reflects the behavior of the hazard in a similar way as the residual quantiles [8, 15, 17].

Another interesting property of the residual p th quantile function involves the comparison of two distributions. We have $t_0(w|p) \geq t_1(w|p)$ for all $w \geq 0$ and $0 \leq p < 1$ if and only if $S_0(t|w) \geq S_1(t|w)$ for all $t \geq 0$ and $w \geq 0$ if and only if $S_0(w)/S_1(w)$ is increasing if and only if $h_1(w) \geq h_0(w)$ for all $w \geq 0$ ([14], Section 2.A), which is a stronger condition than simply $S_0(w) \geq S_1(w)$. Conversely if the two hazard functions cross, then the residual quantile functions must cross as well, for at least some quantiles.

Differentiating both sides of Eq. 4 shows that the sign of $t''(w|p)$ is also determined by the hazard, through the function $h'(w)/h^2(w)$, so that when this function is increasing (decreasing), we have $t''(w|p) < 0$ (> 0) and the residual quantile function is concave (convex).

Examples: Parametric Survival Distributions

A very flexible family of parametric models is provided by the generalized gamma (GG) distribution. As discussed in Cox et al. [4], this is a three-parameter ($\beta, \sigma > 0, \kappa$) distribution with survival function

$$S_{GG}(t) = 1 - \Gamma(\kappa^{-2} \exp\{\kappa[\log(t) - \beta]/\sigma\}; \kappa^{-2}) = 1 - \Gamma\left[\kappa^{-2} \left(e^{-\beta t}\right)^{\kappa/\sigma}; \kappa^{-2}\right] \quad \text{if } \kappa > 0$$

$$S_{GG}(t) = \Gamma\left[\kappa^{-2} \left(e^{-\beta t}\right)^{\kappa/\sigma}; \kappa^{-2}\right] \quad \text{if } \kappa < 0$$

where $\Gamma(t; \gamma) = \int_0^t x^{\gamma-1} e^{-x} dx / \Gamma(\gamma)$ is the cumulative distribution function for the gamma distribution with mean and variance equal to $\gamma > 0$. Since the incomplete gamma function is supplied in standard statistical software packages, the GG survival distribution effectively has a closed form representation, so that computation of either the density or the survival function is straightforward. The limiting case $\kappa = 0$ is the log normal distribution, and $\kappa = 1$ is the Weibull. In addition, the hazard functions of the GG family include all four of the basic shapes [4]: (1) increasing hazard for $0 < \sigma < 1$ and $\sigma \leq \kappa \leq 1/\sigma$; (2) decreasing hazard for $\sigma > 1$ and $1/\sigma \leq \kappa \leq \sigma$; (3) bathtub hazard for $\kappa > \max(\sigma, 1/\sigma)$; and (4) arch-shaped hazard for $\kappa < \min(\sigma, 1/\sigma)$.

The GG family also has the property that $h(0) = 0$ for arch-shaped hazards, and $h(0) = \infty$ when the hazard has a bathtub shape [4]. Therefore the shape of the residual p th quantile function is determined by the shape of the hazard function: (1) decreasing for $0 < \sigma < 1$ and $\sigma \leq \kappa \leq 1/\sigma$; (2) increasing for $\sigma > 1$ and $1/\sigma \leq \kappa \leq \sigma$; (3) arch-shaped for $\kappa > \max(\sigma, 1/\sigma)$; (4) bathtub-shaped for $\kappa < \min(\sigma, 1/\sigma)$.

For example, the residual p th quantile function for the Weibull distribution ($\kappa = 1$) is increasing if $\sigma > 1$, and decreasing if $\sigma < 1$. The particular case of the exponential distribution ($\kappa = \sigma = 1$) has $t(w|p) \equiv t(p)$. The log normal distribution always has a bathtub-shaped residual quantile function since $\kappa = 0 < \min(\sigma, 1/\sigma)$.

The GG quantile function is given by [4]

$$\log [t_{GG(\beta, \sigma, \kappa)}(p)] = \beta + \sigma \log [t_{GG(0, 1, \kappa)}(p)] = \beta + \frac{\sigma}{\kappa} \log [\kappa^2 \Gamma^{-1}(p; \kappa^{-2})] \quad \text{if } \kappa > 0$$

$$\log [t_{GG(\beta, \sigma, \kappa)}(p)] = \beta + \sigma \log [t_{GG(0, 1, \kappa)}(p)] = \beta + \frac{\sigma}{\kappa} \log [\kappa^2 \Gamma^{-1}(1 - p; \kappa^{-2})] \quad \text{if } \kappa < 0$$

The quantile function can be combined with the survival function to obtain the residual quantile function for $GG(\beta, \sigma, \kappa)$.

$$\log [t(w|p) + w] = \beta + \frac{\sigma}{\kappa} \log \left(\kappa^2 \Gamma^{-1} \left[1 - (1 - p) \left(1 - \Gamma \left[\kappa^{-2} \left(e^{-\beta w} \right)^{\kappa/\sigma}; \kappa^{-2} \right] \right); \kappa^{-2} \right] \right) \quad \kappa > 0$$

$$\log [t(w|p) + w] = \beta + \frac{\sigma}{\kappa} \log \left[\kappa^2 \Gamma^{-1} \left((1 - p) \Gamma \left[\kappa^{-2} \left(e^{-\beta w} \right)^{\kappa/\sigma}; \kappa^{-2} \right]; \kappa^{-2} \right) \right] \quad \kappa < 0$$

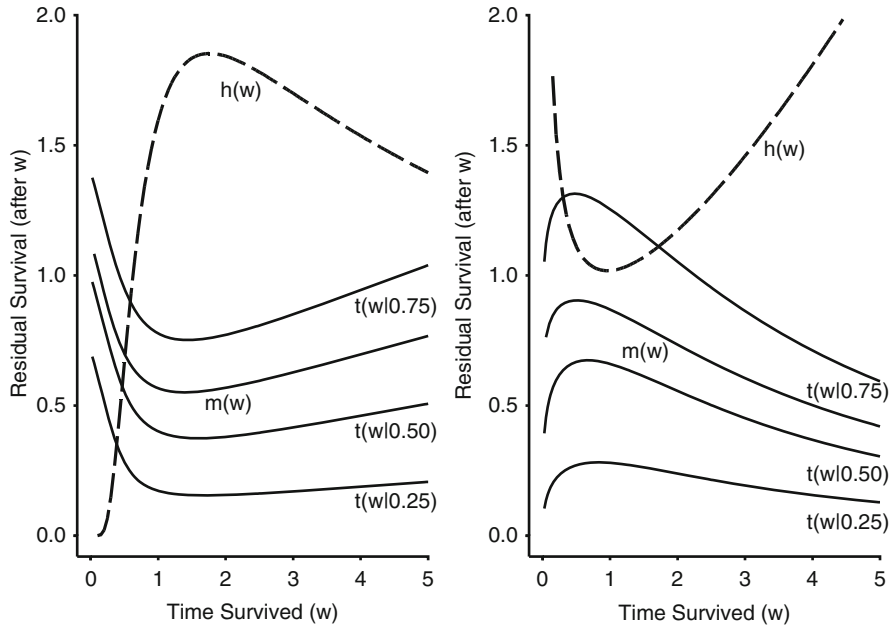


Fig. 2 Hazard $[h(w)]$, residual 25th percentile $[t(w|0.25)]$, residual 50th percentile $[t(w|0.50)]$, residual 75th percentile $[t(w|0.75)]$, and mean residual life $[m(w)]$ functions for GG $(\beta, \sigma, \kappa) = (0, 0.5, 0)$ (left panel) and GG $(\beta, \sigma, \kappa) = (0, 4/3, 3)$ (right panel) distributions. For display purposes the hazard function was multiplied by 100

Note that in this expression $e^{-\beta}$ appears as a scale factor for w and e^β is a scale factor for $t(w|p)$. Therefore, with this parameterization the shape of the hazard is independent of the location parameter β , and so the shape of the residual quantile function is independent of β as well.

Examples of residual quantile functions for two different GG distributions are shown in Fig. 2, illustrating the reciprocal tendencies of the hazard and residual quantile functions. The left hand panel shows the arch-shaped hazard for the log normal distribution GG $(0, 0.5, 0)$. The figure also shows the bathtub-shaped residual quantile functions for $p = 0.25, 0.5, 0.75$. For comparison we have included the mean residual life, $m(w)$, computed using numerical integration. In the right hand panel of Fig. 2 we have GG $(0, 4/3, 3)$, which is an example of a bathtub-shaped hazard since $\kappa > \max(\sigma, 1/\sigma)$. The three residual quantile functions each have the required arch shape, as does the mean residual life.

Programs to estimate the parameters of GG regression models by maximum likelihood, allowing right censoring and in some cases late entry, are available in standard statistical software packages such as SAS, Stata and R. (For sample programs see <http://www.statepi.jhsph.edu/software/generalgamma/generalgamma.html>.) As described in Cox et al. [4], the residual quantile function can be estimated for the GG family using a general purpose program for maximum likelihood

estimation, also available in standard statistical packages. We used SAS PROC NLMIXED, which has the additional capability of estimating nonlinear functions of the (estimated) parameters and data, with standard errors computed by the delta method. This feature was used to compute the residual quantiles for a series of different time points ($w > 0$) for plotting. As described in the next section, bootstrap methods were used to calculate standard errors.

Application

Study Goals and Design

HIV therapy has evolved over time, and it is possible to define sequential calendar periods corresponding to distinct therapeutic eras. Our current analysis includes four calendar periods: an initial period of no therapy or only monotherapy (July 1984–December 1989), followed by a second period of monotherapy or combination therapy (January 1990–December 1994), then by the introduction of highly active antiretroviral therapy (HAART) in the third period (January 1995–December 1999), and finally the era of stable HAART (January 2000–June 2009) in the fourth. The goal of the analysis was to study the effect of HAART on residual survival at the population level.

Our analysis used data from two large, multicenter cohort studies. The first is the Multicenter AIDS Cohort Study (MACS), an ongoing study of homosexual and bisexual men begun in 1983 [11]. The second is the Women's Interagency HIV Study (WIHS), a multicenter cohort study in women begun in 1994 [2]. Both studies have used similar methods and the same data coordinating center. Both studies conduct semi-annual interviews, which include detailed questions about the use of antiretroviral therapy. Data from both cohorts show that a high percentage of participants with clinical AIDS were actually using the indicated therapy during the corresponding period [16]. Deaths are ascertained using both active and passive methods, including abstraction of death certificates and searches of national death registries.

The analysis included 2216 individuals, 1559 (70%) men and 657 (30%) women with follow-up after an incident diagnosis of clinical AIDS, defined using the 1993 CDC surveillance criteria [3] based on clinical conditions (i.e., excluding the laboratory criterion of low CD4 count). MACS men with an incident AIDS diagnosis in either of the first two calendar periods who survived to the end of the period were treated as censored at the end of the period in which the diagnosis occurred, and did not contribute time to subsequent periods. This was done to ensure comparability between the men and women in the final two periods [16], since the WIHS began in 1994, approximately 11 years after the MACS. Individuals with an incident AIDS diagnosis in the third period who were alive at the end of the period were censored at that point and treated as late entries in the fourth period.

Furthermore, for WIHS participants, the date of the AIDS diagnosis could only be determined as the midpoint between two consecutive visits, so half the length of this interval was included in the analysis as left truncation.

If an individual enters observation in a given period at v years after an AIDS diagnosis at age a (with $v=0$ if the individual develops AIDS within the given period), the status at the end of the follow-up within the period at $t > v$ years from the AIDS diagnosis could be either deceased or alive. If $S_{GG}(t)$ denotes the survival function of the GG distribution that describes the time to death in a given period and $f_{GG}(t)$ denotes the corresponding density function, the contributions to the likelihood of those observed to die in the period were $f_{GG}(t)/S_{GG}(v)$. Since essentially all individuals are expected to die by age 100 years (i.e., $t + a < 100$), and in order to impart an anticipated degree of realism into the analysis [18], those alive at the end of follow-up were handled as interval censored observations so that their contributions to the likelihood were $[S_{GG}(t) - S_{GG}(100 - a)]/S_{GG}(v)$.

Given the dramatically improved survival among individuals with AIDS after the introduction of HAART, subsequent prognosis given survival to the present is of great relevance for treated patients. We describe the subsequent experience of individuals who survived for various amounts of time after an initial AIDS diagnosis, using parametric GG models to estimate the residual 25th percentile function, $t(w|0.25)$, in each of the four therapy eras. Since relative times (percentiles) are a useful metric for the comparison of two survival distributions, we also estimated relative (to period one) residual 25th percentiles for selected survival times ($w = 0.5, 1.5, 3$ and 4 years), together with appropriate measures of precision.

For estimation of the precision of the residual percentiles, the delta method performed well in periods one and two, allowing the construction of confidence bands for the residual survival function. For example, for period one, the standard errors of the logs of the residual 25th percentiles for $w = 0.5, 1.5, 3$, and 4 years from the delta method were 0.058, 0.067, 0.148, 0.210 and for the bootstrap 0.052, 0.073, 0.172, 0.243, indicating reasonably good agreement between the large sample and resampling approaches, with slightly less agreement at later times when less data were available. However, in periods three and four the standard errors were clearly inappropriate. We believe this is because, first, compared to the first two periods, in periods three and four many of the participants remained alive at the end of the period; and second, the assumption of interval censoring introduced nonlinear constraints on the three parameters. While this did not affect estimation of the parameters themselves, it did cause problems for estimation of standard errors. We therefore used bootstrap resampling with 500 samples to obtain information concerning the variability of parameter estimates and both absolute and relative residual quantiles estimated from the data. Bootstrap samples were selected at the individual level, sampling individuals and then including all observations for each individual selected for a given sample. This approach has the additional advantage of addressing the finite sample properties of the estimates. The bootstrap distributions of the relative residual 25th percentiles for $w = 0.5, 1.5, 3$, and 4 years are summarized using box-percentile plots [6]; the ends of the box are at the 5th and 95th percentiles so that the box can be interpreted as an approximate 90% bootstrap confidence interval based on the percentile method.

Finally, in the fourth period we estimated the residual 50th and 95th percentiles of residual life, after adjusting for each 10-year increase in age at AIDS diagnosis, with a separate adjustment for age at AIDS greater than 40 years. We first modeled $T \sim S_{GG}[\beta_0 + \beta_1(\text{age at AIDS} - 40)/10 + \beta_2(\text{age at AIDS} - 40)I(\text{age at AIDS} > 40)/10, \sigma_0 + \sigma_1 I(\text{age at AIDS} > 40), \kappa_0 + \kappa_1 I(\text{age at AIDS} > 40)]$, where $I(\cdot)$ is the indicator function. Because σ_1 and κ_1 were not significantly different from zero, we used a conventional GG regression model (i.e., the reduced model with $\sigma_1 = \kappa_1 = 0$), including the linear spline with a change point at 40 years.

Results

Descriptive statistics for the four therapy eras are provided in Table 1. The data show a dramatic decline in mortality from greater than 50% in period 1 before the introduction of HAART to 11% with the introduction in period 3 and to 5% with sustained use of HAART in the fourth period. The table also includes the parameter estimates and bootstrap-based standard errors for the GG models fit separately for each period. In each of the last two periods the estimates of the location parameter β are much larger than for the first two periods, indicating considerably improved survival after the introduction of HAART, as also shown by the medians and 95th percentiles at the bottom of Table 1. The 5th percentiles, however, show that the improvement in survival was not uniform. In addition, differences in the shape (κ) parameters are indicative of very different hazard behavior in the pre-HAART and post-HAART eras.

The estimated GG survival functions for each period are shown in the left-hand panel of Fig. 3, with selected percentiles of each distribution marked on the curves. The corresponding Kaplan-Meier curves are included to show goodness of fit, and indicate that the GG models fit the data well. The survival functions clearly illustrate the dramatic effect of HAART in the last two periods compared to the pre-HAART eras. For example, the 25th percentile increased from 0.58 and 0.73 years in periods one and two to 2.14 and 3.68 in periods three and four, respectively. The corresponding hazard functions are shown in the right-hand panel of Fig. 3, and also reflect improved survival in the HAART eras. The shapes of the hazards for the first two periods are in marked contrast to those of the two periods after the introduction of HAART. HAART induces a decreasing hazard in the first 10 years after an AIDS diagnosis before age becomes dominant and produces a steep increase in the hazard after 60 years of age, which corresponds to 20 years in the x-scale of the figure, as the median age at AIDS diagnosis was 40 years (see Table 1). It is clear that the period of stable use of more potent HAART was associated with a lower hazard than in the era when HAART was introduced.

Figure 4 shows the residual 25th percentile functions $t_{GG}(w|0.25) = t_{GG}(1 - 0.75S(w)) - w$ for the four periods based on the four GG models. Consistent with the bathtub-shaped hazard functions in periods three and four shown in

Table 1 Descriptive statistics for survival after an initial diagnosis of clinical AIDS in four calendar periods from July 1984 to June 2009, and parameter estimates, estimated percentiles and standard errors for GG models fitted to the data from each calendar period

Characteristic	Calendar period			
	July 1984–December 1989	January 1990–December 1994	January 1995–December 1999	January 2000–June 2009
Therapy era	No therapy/monotherapy	Monotherapy/combination therapy	HAART introduction	Stable HAART
No.	633	661	546	763
No. (%) incident AIDS	633 (100)	661 (100)	546 (100)	376 (49)
Median (IQR) date of AIDS diagnosis	October 1987 (July 1986–December 1988)	April 1992 (February 1991–June 1993)	May 1996 (July 1995–April 1997)	September 1999 (August 1996–August 2003)
Median (IQR) age at AIDS diagnosis, years	36.5 (32.1–41.1)	39.6 (35.3–43.9)	39.9 (35.1–44.8)	40.6 (35.4–45.9)
No. person-years in period	685	912	1,342	4,282
No. deaths in period (% of person-years)	388 (57%)	446 (49%)	144 (11%)	235 (5%)
GG model estimates				
$\hat{\beta} \pm SE^a$	0.674 \pm 0.062	0.650 \pm 0.054	2.928 \pm 0.160	3.297 \pm 0.041
$\hat{\sigma} \pm SE^a$	0.773 \pm 0.071	0.849 \pm 0.049	0.615 \pm 0.055	0.549 \pm 0.031
$\hat{\tau} \pm SE^a$	1.369 \pm 0.185	0.857 \pm 0.123	2.952 \pm 0.249	3.031 \pm 0.158
GG estimated percentiles, years				
5th $\pm SE^a$	0.10 \pm 0.02	0.19 \pm 0.02	0.12 \pm 0.04	0.25 \pm 0.08
50th $\pm SE^a$	1.30 \pm 0.06	1.48 \pm 0.06	7.53 \pm 0.85	11.66 \pm 0.84
95th $\pm SE^a$	4.14 \pm 0.35	5.11 \pm 0.38	26.89 \pm 3.73	37.13 \pm 1.48

HAART highly active antiretroviral therapy, IQR inter-quartile range, GG generalized gamma, SE standard error
^aStandard errors based on individual level bootstrap re-sampling with 500 samples

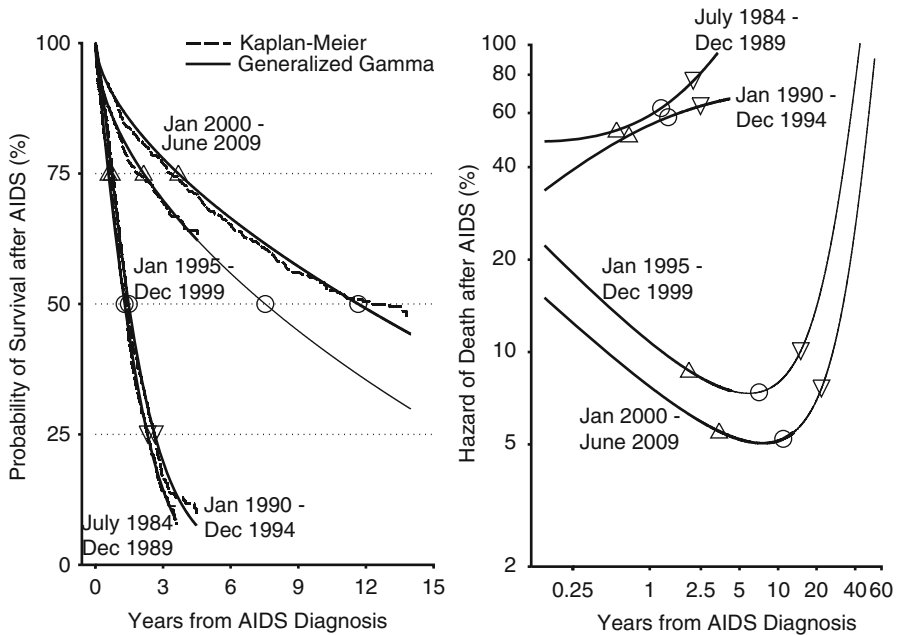


Fig. 3 Survival after a diagnosis of AIDS. Survival functions and appropriateness of the GG model with separate parameters for each of the four therapy eras, as judged by Kaplan-Meier curves (*left panel*); corresponding hazards of death (*right panel*). Parameter estimates $(\hat{\beta}, \hat{\sigma}, \hat{\kappa})$ for each period are given in Table 1. The symbols Δ , \circ , and ∇ correspond to the 25th percentile, 50th percentile, and 75th percentile, respectively, of each distribution

Fig. 3, the residual 25th percentile functions for periods three and four are both arch-shaped. In the final two periods, we estimate that 75% of participants who survive for five years after a diagnosis of AIDS will survive for more than 3.9 and 5.7 additional years, respectively.

To more directly compare the residual survival times in the four periods we turn to relative residual times, that is, to ratios of residual quantiles from periods 2, 3 and 4 to period one. For periods 2–4, Fig. 5 displays relative residual 25th percentiles with box-percentile plots based on 500 bootstrap samples at four different points in time after a diagnosis of AIDS ($w = 0.5, 1.5, 3$ and 4 years). In addition, for each period we have indicated the percentage of participants surviving to each of these times at the top of the figure. The graph shows a slight improvement in residual survival in period two relative to period one, in particular the 90% confidence interval for the relative residual 25th percentile excludes one for $w = 3, 4$ years. The relative residual quantiles increase in the later periods, indicating better residual survival with increasing time survived (w) in periods three and four compared to period one. For $w = 0.5$ and 1.5 years there is virtually no overlap between the bootstrap distributions of relative 25th percentiles for periods three and four, but

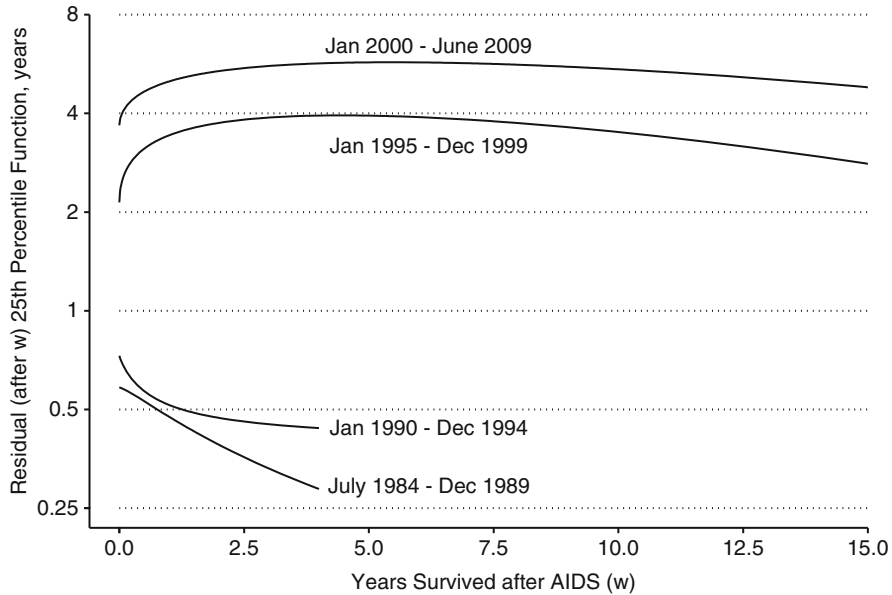


Fig. 4 Residual 25th percentile functions for survival after AIDS for four calendar periods based on the GG model with different parameters for each period. Parameter estimates $(\hat{\beta}, \hat{\sigma}, \hat{\kappa})$ for each period are given in Table 1

by $w = 4$ years there is substantial overlap. This reflects the time required before HAART was used universally in the two cohorts. As one would expect the precision of the estimates decreases with the time survived (w), since only 14% and 6% of participants in period one survive to 3.0 and 4.0 years after an AIDS diagnosis, respectively.

The parameter estimates (SE_{boot}) for the linear spline model for period 4 $\{S_{GG}[\beta_0 + \beta_1(age\ at\ AIDS - 40)/10 + \beta_2(age\ at\ AIDS - 40)I(age\ at\ AIDS > 40)/10, \sigma_0, \kappa_0]\}$ were $\hat{\beta}_0 = 3.492$ (0.075), $\hat{\beta}_1 = -0.119$ (0.144), $\hat{\beta}_2 = -0.301$ (0.189), $\hat{\sigma}_0 = 0.520$ (0.022), $\hat{\kappa}_0 = 3.220$ (0.121). Figure 6 shows the residual 50th percentiles (left panel) and residual 95th percentiles (right panel) for each of six different ages at AIDS diagnosis (30, 35, 40, 45, 50, and 55 years) in period four, based on the conventional GG model. The arch-shaped residual 50th percentile functions for ages 30, 35, and 40 are similar; the curves only start to separate for ages greater than 40. In contrast, for each age group the residual 95th percentile functions are almost entirely monotonically decreasing after a very slight initial increase.

The left panel of the figure shows that, at the time of AIDS diagnosis ($w = 0$), 50% of those who were 30 years of age when they developed AIDS will live an additional 16 years. Interestingly, 50% of individuals who were 30 years of age when they developed AIDS and survived an additional 15 years also lived an additional 16 years. This is in contrast to what is observed for those who were 55

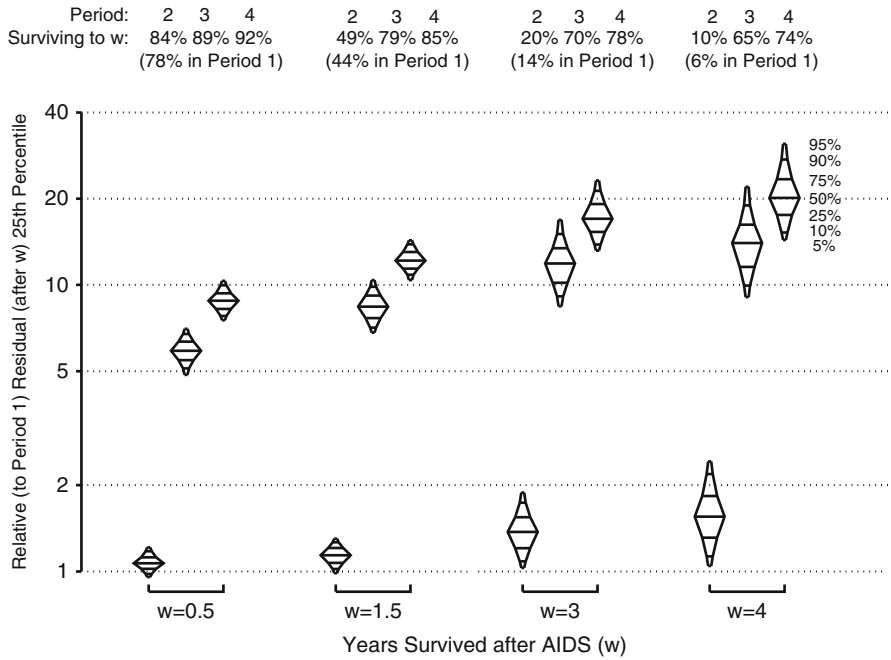


Fig. 5 Relative residual 25th percentile of survival for $w = 0.5, 1.5, 3,$ and 4 years after AIDS diagnosis in periods 2, 3 and 4 compared to period one. Box-percentile plots based on 500 bootstrap samples; the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of each distribution of relative residual percentiles are depicted. Thus each box corresponds to a 90% bootstrap confidence interval. Residual 25th percentiles (SE_{boot}) at 0.5, 1.5, 3 and 4 years for period one were 0.53 (0.03), 0.43 (0.03), 0.33 (0.06), and 0.29 (0.07). At the top are proportions surviving to each value of w for each period

years of age when they developed AIDS; the residual 50th percentile at $w = 0$ is 7.5 years, compared to the residual 50th percentile of 4.5 years at $w = 15$ years. These results are logical; $t(0|0.50) \approx t(15|0.50)$ among individuals who develop AIDS at age 30 because we are comparing relatively young 30 and 45 year-olds but, $t(0|0.50) > t(15|0.50)$ among individuals who develop AIDS at age 55 because we are now comparing 55 to 70 year olds.

Discussion

Here we have shown that residual survival is useful for the evaluation of interventions from both the patient and scientific perspectives. Residual survival provides additional information concerning longer term effects, which should be considered as part of any assessment of the effect of an intervention on a time-to-event outcome. The conditional metric is appropriate for the point(s) at which these

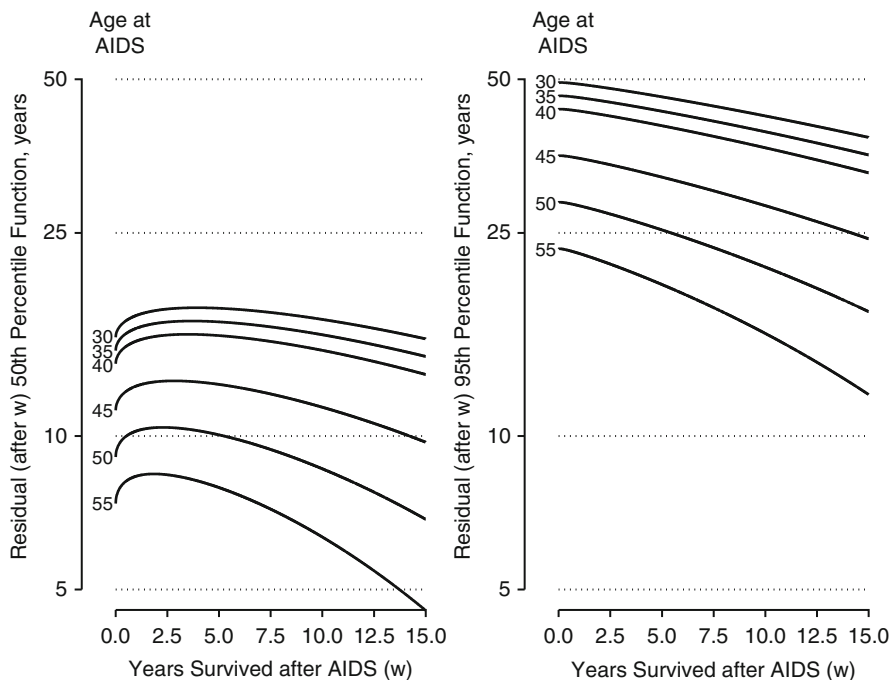


Fig. 6 Residual 50th percentile functions (*left panel*) and residual 95th percentile functions (*right panel*) for survival after AIDS in period 4 (January 2000 through June 2009) for different ages at time of AIDS diagnosis, based on a conventional GG regression model with a linear spline for age at AIDS diagnosis greater than 40

effects become most relevant. We have previously emphasized the importance of the quantile function and the use of relative times in addition to relative hazards, proportional or not, for providing a more complete comparison of the underlying survival distributions. These ideas work equally well in the context of residual survival.

The application considered four different periods of therapy for HIV infected individuals with a diagnosis of AIDS. Our results provide additional information concerning the beneficial effects of HAART. In particular we saw that, in the final two periods, 75% of those who survived for five years after an initial AIDS diagnosis were estimated to survive an additional 3.9 and 5.7 years. In this application, residual survival provided useful information about the longer term effects of HAART at the population level.

As recently reported by Wada and colleagues [18], when a substantial proportion of individuals in the study population remain alive at the end of follow-up, handling the observations as censored in the interval with upper bound determined by 100 years of age complements the observed data and imparts an anticipated degree of realism to the analysis. Indeed, for the study population of period four with 40 years

as the median age at AIDS diagnosis, 99.9% of them are expected to die by age 90.5 years (= 40 + 99.9th percentile of GG(3.297, 0.549, 3.031)). In contrast, if those remaining alive in period four were treated as right censored observations, 18% (= survival of GG(2.854, 1.730, 0.517) at 60) would be expected to survive more than 60 years after the diagnosis of AIDS at the median age of 40 years, which is an unrealistic estimate.

The GG family has again provided a useful platform for parametric survival analysis. Using these models, it was straightforward to obtain expressions for the residual quantile functions that were easily implemented using standard statistical software [4]. Although the delta method did not consistently provide useful estimates of variation, the bootstrap was not difficult to implement and worked quite well. We believe that the parametric GG approach has been shown to be very useful once again and deserves to be more widely employed.

Acknowledgements Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) with centers (Principal Investigators) at The Johns Hopkins Bloomberg School of Public Health (Joseph B. Margolick, Lisa P. Jacobson), Howard Brown Health Center, Feinberg School of Medicine, Northwestern University, and Cook County Bureau of Health Services (John P. Phair), University of California, Los Angeles (Roger Detels), and University of Pittsburgh (Charles Rinaldo). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with supplemental funding from the National Cancer Institute. U01-AI-35042, UL1-RR025005, U01-AI-35043, U01-AI-35039, U01-AI-35040, U01-AI-35041.

Data in this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group with centers (Principal Investigators) at New York City/Bronx Consortium (Kathryn Anastos); Brooklyn, NY (Howard Minkoff); Washington, DC Metropolitan Consortium (Mary Young); The Connie Wofsy Study Consortium of Northern California (Ruth Greenblatt); Los Angeles County/Southern California Consortium (Alexandra Levine); Chicago Consortium (Mardge Cohen); Data Coordinating Center (Stephen Gange). The WIHS is funded by the National Institute of Allergy and Infectious Diseases (U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-AI-34993, and U01-AI-42590) and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U01-HD-32632). The study is co-funded by the National Cancer Institute, the National Institute on Drug Abuse, and the National Institute on Deafness and Other Communication Disorders. Funding is also provided by the National Center for Research Resources (UCSF-CTSI Grant Number UL1 RR024131).

Appendix

Quantiles of Residual Survival

We derive the shape of the residual quantile function for continuous, arch-shaped hazards satisfying $h(w) > 0$ for $w > 0$. Statements and proofs for bathtub-shaped hazards are similar; the condition corresponding to $h(0) = 0$ in Lemma 4 is $h(0) = \infty$.

Lemma 1. Suppose $h(w)$ is continuous and decreasing on the interval (w_m, ∞) . Then $t'(w|p) > 0$ for $w > w_m$ and therefore $t(w|p)$ is increasing on (w_m, ∞) .

The proof is immediate from Eq. 4.

Lemma 2. Assume that $h(w)$ is arch-shaped, increasing on the interval $(0, w_m)$, and decreasing on (w_m, ∞) , with a maximum at w_m . Suppose that for some $w_p < w_m$ we have $h(w_p) \leq h[t(w_p|p) + w_p]$. Then $h(w) < h[t(w|p) + w]$ for every $w < w_p$, and therefore $t(w|p)$ is decreasing on $(0, w_p)$.

Since $h(w)$ is increasing on $(0, w_m)$ and $t(w|p) + w$ is increasing, the proof is immediate if $t(w_p|p) + w_p \leq w_m$, since in this case, $t(w|p) + w < t(w_p|p) + w_p \leq w_m$ for $w < w_p$ and therefore $h(w) < h[t(w|p) + w]$. Alternatively, if $w_m < t(w_p|p) + w_p$ then for $w < w_p$, either $t(w|p) + w \leq w_m$, which implies $h(w) < h[t(w|p) + w]$; or $w_m < t(w|p) + w < t(w_p|p) + w_p$, which implies $h(w) < h(w_p) \leq h[t(w_p|p) + w_p] < h[t(w|p) + w]$ since $h(w)$ is decreasing on (w_m, ∞) .

Lemma 3. Assume that $h(w)$ is arch-shaped as in Lemma 2. For a given $0 < p < 1$, suppose that for some $w_p > 0$, we have $t'(w_p|p) = 0$; equivalently, $h(w_p) = h[t(w_p|p) + w_p]$. Then $w_p < w_m$ and $t(w|p)$ is decreasing on $(0, w_p)$ and increasing on (w_p, ∞) , and therefore has a bathtub shape with a minimum at w_p . If there is no such value w_p then $t(w|p)$ is increasing for all $w > 0$.

Since $h(w)$ is arch-shaped we must have $w_p < w_m < t(w_p|p) + w_p$. By the first lemma, $t(w|p)$ is increasing on (w_m, ∞) , and by the second lemma, $t(w|p)$ is decreasing on $(0, w_p)$. For $w_p < w < w_m$, we must have $h(w) > h(w_p) = h[t(w_p|p) + w_p] > h[t(w|p) + w]$ since $w_m < t(w_p|p) + w_p < t(w|p) + w$. Therefore $t(w|p)$ is increasing on (w_p, w_m) . The second statement follows from Lemma 1 and the fact that $t(w|p)$ has no critical point.

Lemma 4. Suppose that $h(w)$ is arch-shaped as in Lemma 2. If $h(0) < h(w)$ for all $w > 0$, in particular if $h(0) = 0$, then $t(w|p)$ has a bathtub shape for all $0 < p < 1$. If $h(0) > 0$ and $h(0) > h(w^*)$ for some $0 < w^*$, then there is a percentile $0 < p^* < 1$, such that $t(w|p)$ is increasing for all $p^* < p < 1$.

Since $h(w)$ is decreasing on (w_m, ∞) , we have $h(w) > h[t(w|p) + w]$ for $w_m < w$ and all $0 < p < 1$. If $h(0) < h(w)$ for $w > 0$, then for $0 < w < w_m$ sufficiently small we must have $h(w) < h[t(w|p) + w]$ by continuity and $h(0) < h[t(0|p)] = h[t(0|p) + 0]$. Since $h(t)$ is continuous, there is a point $0 < w_p < w_m$ such that $h(w_p) = h[t(w_p|p) + w_p]$. By Lemma 3, the residual quantile function has a bathtub shape, with a minimum at w_p . Now suppose $h(0) > 0$ and $h(0) > h(w^*)$. Then $w_m < w^*$, and therefore $h(w)$ is decreasing on (w^*, ∞) and $h(w) > h(w^*)$ for $w < w^*$. By Lemma 1, $t(w|p)$ is increasing on (w_m, ∞) for all $0 < p < 1$. Now choose $0 < p^* < 1$ large enough that $w^* < t(p^*)$ and therefore $h(w^*) > h[t(p^*)]$. Then for all $p^* < p$ and $w < w_m$ it follows that $h(w) > h(w^*) > h\{t(w|p) + w\}$. Therefore $t(w|p)$ is also increasing on $(0, w_m)$.

Note, however, that $t(w|p)$ cannot be increasing for all $0 < p < 1$. For either an arch-shaped or bathtub-shaped hazard, there must be at least one pair of values $0 < w^* < w^{**}$ such that $h(w^*) = h(w^{**})$, and then $t'(w^*|p^*) = 0$ for $p^* = 1 - S(w^{**})/S(w^*)$ since $t(w^*|p^*) + w^* = w^{**}$. An example of an arch-shaped hazard satisfying $h(0) > 0$ and $h(0) > h(w^*)$ for some $0 < w^*$ is given by $h(t) = (2 + t - t^2)1_{0 \leq t \leq 1} + \{1 + \exp[-(t - 1)]\} 1_{1 \leq t}$. It is not difficult to verify graphically that $t(w|p)$ is increasing for $p \geq 0.90$.

References

1. Alam, K., Kulasekera, K.B.: Estimation of the quantile function of residual life time distribution. *J. Stat. Plann. Inference* **37**, 327–337 (1993)
2. Barkan, S.E., Melnick, S.L., Preston-Martin, S., Weber, K., Kalish, L.A., Miotti, P.: The Women's interagency HIV study. WIHS Collaborative Study Group. *Epidemiology* **9**, 117–125 (1998)
3. Centers for Disease Control and Prevention: 1993 Revised Classification System for HIV Infection and Expanded Surveillance Case Definition for AIDS Among Adolescents and Adults. *MMWR* **41**, 1–19 (1992)
4. Cox, C., Chu, H., Schneider, M.F., Muñoz, A.: Tutorial in biostatistics: parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat. Med.* **26**, 4352–4374 (2007)
5. Csörgö, S., Viharos, L.: Confidence bands for percentile residual lifetimes. *J. Stat. Plann. Inference* **30**, 327–337 (1992)
6. Esty, W., Banfield, J.: The box-percentile plot. *J. Stat. Softw.* **8**, 1–14 (2003)
7. Franco-Pereira, A.M., Lillo, R.E., Romo, J.: Comparing quantile residual life functions by confidence bands. *Lifetime Data Anal.* **18**, 195–214 (2011)
8. Ghai, G.L., Mi, J.: Mean residual life and its association with failure rate. *IEEE Trans. Reliab.* **48**, 262–266 (1999)
9. Jeong, J.-H., Jung, S.-H., Costantino, J.P.: Nonparametric inference on median residual life function. *Biometrics* **64**, 157–163 (2008)
10. Joe, H.: Characterizations of life distributions from percentile residual lifetimes. *Ann. Inst. Stat. Math.* **37**, 165–172 (1985)
11. Kaslow, R.A., Ostrow, D.G., Detels, R., Phair, J.P., Polk, B.F., Rinaldo Jr., C.R.: The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am. J. Epidemiol.* **126**, 310–318 (1987)
12. Kim, M.-O., Zhou, M., Jeong, J.-H.: Censored quantile regression for residual lifetimes. *Lifetime Data Anal.* **18**, 177–194 (2011)
13. Lillo, R.E.: On the median residual lifetime and its aging properties: a characterization theorem and applications. *Nav. Res. Logist.* **52**, 370–380 (2005)
14. Marshall, A.W., Olkin, I.: *Life Distributions*. Springer-Verlag, New York (2007)
15. Mi, J.: Bathtub failure rate and upside-down bathtub mean residual life. *IEEE Trans. Reliab.* **44**, 388–391 (1995)
16. Schneider, M.F., Gange, S.J., Williams, C.M., Anastos, K., Greenblatt, R.M., Kingsley, L., Detels, R., Muñoz, A.: Patterns of the hazard of death after AIDS through the evolution of antiretroviral therapy: 1984–2004. *AIDS* **19**, 2009–2018 (2005)
17. Tang, L.C., Lu, Y., Chew, E.P.: Mean residual life of lifetime distributions. *IEEE Trans. Reliab.* **48**, 73–78 (1999)
18. Wada, N., Jacobson, L.P., Cohen, M., French, A., Phair, J., Muñoz, A.: Cause-specific life expectancies after 35 years of age for human immunodeficiency syndrome-infected and human immunodeficiency syndrome-negative individuals followed simultaneously in long-term cohort studies: 1984–2008. *Am. J. Epidemiol.* **15**, 116–125 (2013). doi:[10.1093/aje/kws321](https://doi.org/10.1093/aje/kws321)

Part II

Evaluation of Predictions

Methods for Evaluating Prediction Performance of Biomarkers and Tests

Margaret Pepe and Holly Janes

Abstract This chapter covers material presented in a short course at the 2011 International Conference on Risk Assessment and Evaluation of Predictions. Methods for evaluating the performance of markers to predict risk of a current or future clinical outcome are reviewed. Specifically, we discuss criteria for evaluating a risk model including: calibration, accurate classification and benefit for decision making using the model. Measures for making comparisons between models are described. The role of risk reclassification techniques is discussed. We present a detailed example.

Introduction

Background

Predicting an individual's risk of a particular outcome of interest is a key component of medical decision making. For example, the Framingham Risk Calculator (www.framinghamheartstudy.org) provides 10 year risks of cardiovascular event outcomes such as coronary heart disease and myocardial infarction events as functions of risk factors [6]. Risk factors for hard coronary heart disease (HCHD), defined as myocardial infarction or coronary death, include age, smoking status, treatment for hypertension and levels of total cholesterol, high density lipoproteins and systolic blood pressure. If the 10-year risk exceeds 20 %, long term treatment

M. Pepe (✉)

Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

University of Washington, Seattle, Washington, USA

e-mail: mspepe@u.washington.edu

H. Janes

Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

e-mail: hjanes@fhcrc.org

with cholesterol lowering drug therapy is recommended. Another risk calculator routinely used in clinical practice is the Breast Cancer Risk Assessment Tool (BCRAT) [32]. The predicted outcome may be a future event such as a cardiovascular event for the Framingham Risk Calculator or breast cancer diagnosis for BCRAT. However, a current condition can also constitute the predicted outcome. For example, presence of acute kidney injury is the outcome predicted in Parikh et al. [16] and presence of critical illness requiring hospitalization is the outcome predicted by Seymour et al. [26]. New molecular biology techniques for measuring biomarkers and new imaging technologies portend the availability of excellent predictors of risk in the future. Moreover, easy dissemination of risk calculators over the internet will increase the impact of risk prediction models on clinical practice.

In this chapter we discuss methods for evaluating risk prediction models. Since the goal is to use risk calculators in medical decision making, our perspective is that the crucial evaluations are about determining whether or not good decisions can be made with use of the risk prediction model. For much of the chapter we define a good decision rule as one that recommends treatment for people who would get the outcome of interest in the absence of treatment (called *cases* here) and does not recommend treatment for those who would not have the outcome (called *controls* here). The rationale is that the cases could possibly benefit from treatment while the controls would not benefit but would be subjected to toxicities, expenses and other costs associated with treatment. Therefore, we consider that a prediction model is good if it leads to a large proportion of cases being classified into a treatment category and a large proportion of controls into a no-treatment category. However, prediction models are not just classifiers. A prediction model is an algorithm that people use to calculate their risk and as such it has real meaning and interpretation for individuals. This must be accounted for in evaluating a prediction model and sets it apart from evaluations of other classifiers such as diagnostic tests where a numerical score itself may not have meaning.

Notation and Assumptions

We write D for the outcome of interest. Without loss of generality we assume that D is a negative outcome and we refer to it as the “bad outcome”. We assume that the outcome is binary, $D = 1$ for a case and $D = 0$ for a control. If the outcome is an event occurring within a specific time period, for example a cardiovascular event within 10 years, the cases may be called *events* and the controls may be called *nonevents*. The prevalence or event rate in the population is denoted by ρ :

$$\rho = P(D = 1).$$

The predictors are denoted by X and Y , both of which may be multidimensional. In section “Measuring Prediction Performance of a Single Model”, we consider a single risk model and use X for the predictors in the model. In section “Comparing

Two Risk Models”, we consider two nested risk models, one with the baseline predictors denoted by X and one expanded model that includes the predictors Y in addition to X .

To focus the presentation we assume that in the absence of predictor information subjects do not receive treatment. The purpose of the risk model is to identify subjects for treatment. One may be interested in the opposite scenario in some settings. That is, standard of practice may be to receive treatment and the purpose of the model is to identify subjects at low risk who may forego treatment. This setting can be dealt with using methods analogous to those we describe here and is mentioned later, but for the most part, and to keep the discussion focused, we consider the default no-treatment scenario.

We assume that data are available for a cohort of N independent untreated subjects. We write the data as $\{(D_i, Y_i, X_i); i = 1, \dots, N\}$. Most of this chapter concerns conceptual formulations of measures to quantify and compare the prediction performance of risk models. As such, sampling variability and statistical inference is not a major focus, at least in sections “Measuring Prediction Performance of a Single Model” and “Comparing Two Risk Models”. In other words we assume N is very large.

Illustrative Data

For illustrative purposes we use a simulated data set. The simulated data are available on the DABS website (<http://labs.fhcr.org/pepe/dabs/index.html>) and were previously used in a publication [19]. The data reflect the prevalence and risk ranges that have been reported for cohort studies of cardiovascular disease. A total of 10,000 observations are included, of which 1017 are case subjects and 8983 are control subjects. Predictors X and Y are one-dimensional and continuous. In practice the predictors X and/or Y may be scores derived from multiple risk factors or biomedical measurements such as blood levels of lipids or C-reactive protein. In section “Measuring Prediction Performance of a Single Model” we focus on the predictor X only, while in section “Comparing Two Risk Models” we consider X and Y together as predictors. Figure 1 shows the joint and marginal distributor of X and Y among cases and controls. Table 1 shows fitted logistic regression models for the baseline risk model including X only and the expanded risk model including X and Y .

Chapter Outline

In section “Validity of the Risk Calculator” we discuss the concept of risk and validity of a risk model. The performance of a risk model is discussed in section “Measuring Prediction Performance of a Single Model”. A plethora of

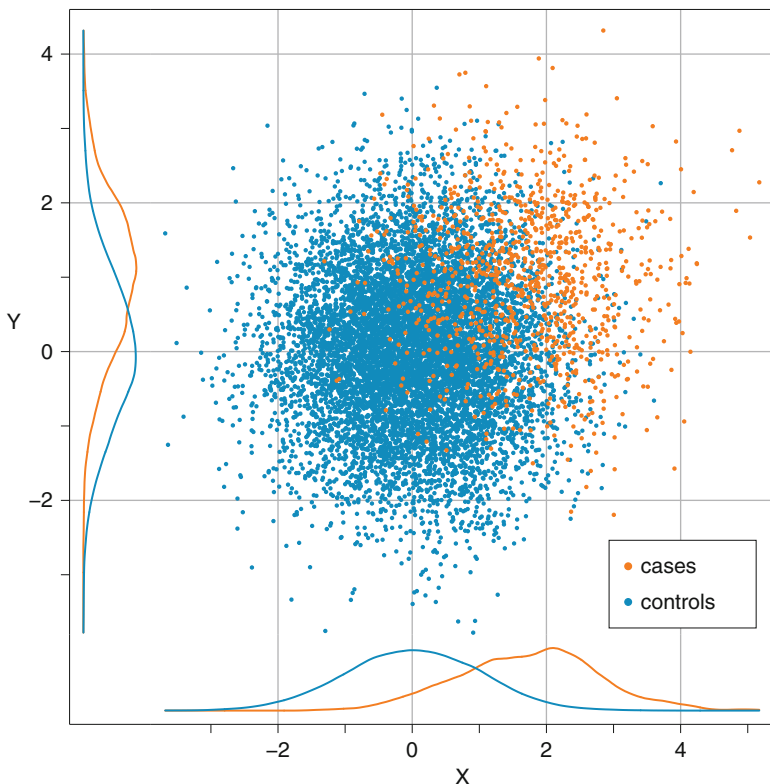


Fig. 1 Joint and marginal distributions of X and Y among cases and controls

Table 1 Estimated coefficients for baseline and expanded risk models

Factor	Baseline model		Expanded model	
	Coefficient	SE	Coefficient	SE
Intercept	-3.67	0.07	-4.23	0.09
X	1.72	0.05	1.77	0.05
Y		1.01	0.05	

performance measures are used to assess prediction performance and we describe the main ones. Insights and relationships amongst the measures are provided. In section “Comparing Two Risk Models” we consider the comparison of two risk prediction models focusing especially on the comparison of two nested models. This is a somewhat controversial area of statistical methodology where public debate and discourse is needed. We hope that this chapter will add constructively to the discourse.

Validity of the Risk Calculator

What Is risk(X)?

The function $\text{risk}(x) = P(D = 1|X = x)$ is the *frequency* of events among subjects with predictor values $X = x$. It is important to remember that statistical analysis delivers information about population level entities, such as averages and frequencies and distributions. We emphasize this point because the term ‘individual level risk’ is often used in this era of ‘personalized medicine’. But the risk value calculated from the risk calculator for a subject with predictors $X = x$, $\text{risk}(x)$, is not the probability of a random event for that subject. Rather it is the frequency of events in the group of subjects with the same predictors as that subject.

To make the distinction concrete, suppose that $\text{risk}(x) = 0.20$ and consider the (large) group of subjects with predictors $X = x$. The following scenarios are all consistent with $\text{risk}(x) = 0.20$: (i) 20% of the subjects are destined to have the event with probability 1 while 80% are destined not to have the event; (ii) for each subject i there is a stochastic mechanism giving rise to an event $D = 1$ with individual level probability $\pi_i = 0.20$; (iii) 10% of subjects are destined to have the event ($\pi_i = 1.00$ for them), 50% are destined not to have the event ($\pi_i = 0.00$ for them) and for 40% of subjects the outcome is stochastic with $\pi_i = 0.25$. Gail and Pfeiffer [7] discuss the distinction between the unobservable individual level probabilities denoted by π_i and $\text{risk}(x) = P(D = 1|X = x)$. They note the equality: $\text{risk}(x) = E\{\pi_i|X_i = x\}$. A simulated example that illustrates the distinction can be found in Pepe [19].

Unless repeated observations of the outcome were available for an individual, one cannot make inference about individual level risks. In this sense, the individual level risks, i.e. the π_i 's, are not observable. It is not clear that they are even well defined when a subject can only experience one event. We regard the concept of individual level risk π as a distraction. Individualized risk will not be discussed further in this chapter.

Risk is a function of the predictors modeled. It is important to remember that an individual with two sets of non-overlapping predictors $X = x$ and $Y = y$ has at least three risk values, $\text{risk}(x) = P(D = 1|X = x)$, $\text{risk}(y) = P(D = 1|Y = y)$ and $\text{risk}(x, y) = P(D = 1|X = x, Y = y)$. Each is his ‘true risk’. Each is a frequency of events but calculated amongst different groups of subjects: those with $X = x$, those with $Y = y$ and those with $X = x$ and $Y = y$, respectively.

The Meaning of Calibration

The traditional definition of calibration is that a well calibrated risk calculator $\text{risk}^*(\cdot)$, is one for which the frequency of events among subjects with $X = x$ is equal to $\text{risk}^*(x) : P(D = 1|X = x) = \text{risk}^*(x)$. When X is multidimensional it can be difficult to assess calibration defined in this strong sense. A weaker definition of

calibration is typically used in practice: the criterion is that $P(D = 1 | \text{risk}^*(X) = r) \approx r$ for all r . In words, if the frequency of events is r among subjects whose calculated risks are equal to r , then the model $\text{risk}^*(\cdot)$ is considered well calibrated in the weak sense. This level of validity seems like a minimal requirement to justify use of the risk model in practice.

We note that strong calibration implies weak calibration because under strong calibration, where $P(D = 1 | X = x) = \text{risk}^*(x)$ we have $P(D = 1 | \text{risk}^*(X) = r) = E\{P(D = 1 | X = x) | \text{risk}^*(X) = r\} = E\{\text{risk}^*(X) | \text{risk}^*(X) = r\} = r$. However, weak calibration does not imply strong calibration and must not be interpreted as such.

Calibration is an attribute that may not transport from one population to another. For example, if there are predictors (known or unknown) that are not included in the model and that have different distributions in two populations, the true risk models in the two populations will likely be different, $P(D = 1 | X, \text{populationA}) \neq P(D = 1 | X, \text{populationB})$. Strong calibration may not transport for this reason. However, if all relevant predictors are included in the model, strong calibration will not be affected by a change in the distribution of those predictors. On the other hand, a model that is weakly calibrated but not strongly calibrated may not transport its weak calibration to other populations where distributions of modeled covariates differ.

Assessing Calibration

To evaluate calibration one compares the observed event rates within subgroups of subjects defined by the modeled predictors to the average estimated risk values among those subjects. The subgroups are usually selected as having estimated risks in a narrow range. A visual aid to this comparison is the predictiveness curve [22]. The plot orders subjects from lowest to highest estimated risk, plotting each risk value vs its quantile, i.e. the proportion of subjects with risks less than or equal to that value. The x-axis allows one to identify subgroups similar in regards to estimated risk. For example, groups may be defined by decile of estimated risk. The observed event rate in each subgroup is superimposed on the plot using circles as shown in Fig. 2. If the circles follow the predictiveness curve, we conclude that estimated risks are close to observed risks and the model is well calibrated (in the weak sense) in the dataset. The predictiveness curve in Fig. 2 shows that the fitted model is extremely well calibrated. An alternative but related visual display is the calibration plot (Fig. 3) [26]. This also uses intervals of estimated risk (e.g. deciles) but plots observed event rates versus average estimated risk producing points that should lie along the 45° line if the model is well calibrated. A disadvantage of the calibration plot versus the predictiveness curve is that points can be more clumped. Moreover the variability in estimated risk within an interval risk category is not evident from the calibration plot. If substantial variation exists in a category one might choose to use smaller subdivisions of that category in comparing observed event rates with estimated rates.

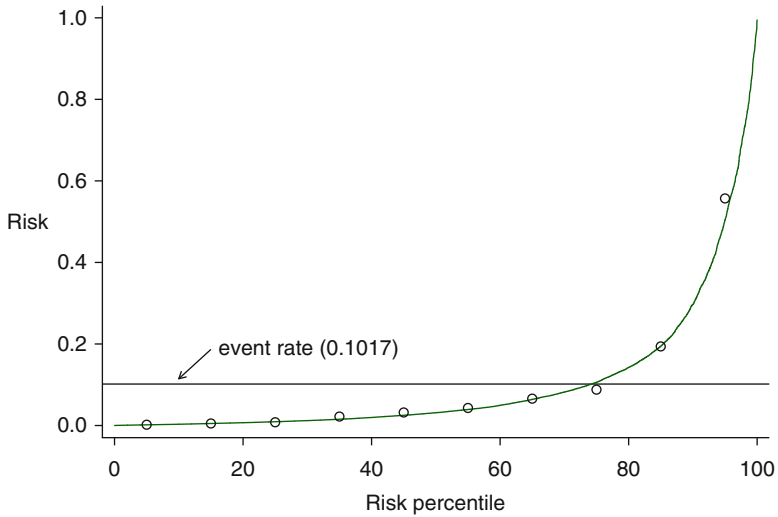


Fig. 2 Visual assessment of calibration with the predictiveness curve

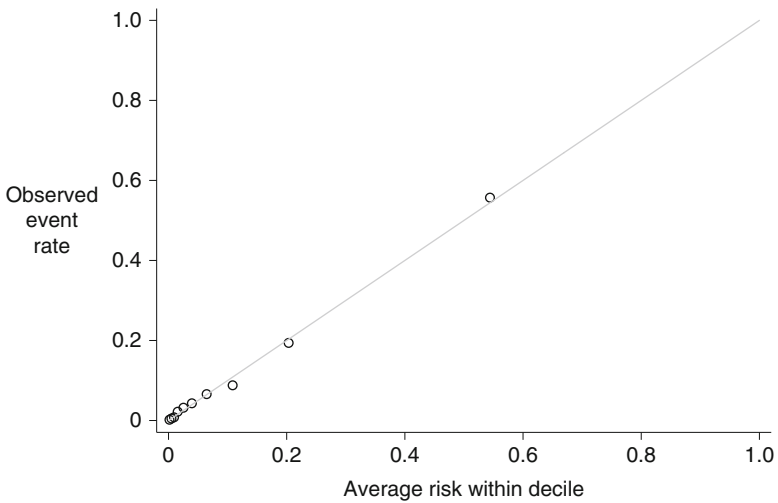


Fig. 3 Calibration assessed using the calibration plot. The model is well calibrated if points lie on the 45° line shown

The Hosmer-Lemeshow test is often reported as a test for calibration [15]. It also uses subsets defined by estimated risk, typically deciles, and calculates

$$H \equiv \sum_{k=1}^{10} N_k \frac{(O_k - \bar{r}_k(X))^2}{\bar{r}_k(X)(1 - \bar{r}_k(X))}$$

where N_k = the number of subjects in the k th group, O_k = the observed event rate and $\widehat{r}_k(X)$ is the average estimated risk. Under the null hypothesis of good model calibration, the statistic H has a chi-squared distribution with degrees of freedom equal to the number of groups minus 2. The statistic corresponds closely with the predictiveness and calibration plots in that it compares O_k and \widehat{r}_k . However, the statistic has been criticized for many reasons including that it is highly dependent on sample size (almost certainly significant if N is large enough and non-significant if N is small enough). In of itself, it does not convey the extent to which the model is well calibrated to the data. But it can serve as a descriptive adjunct to the visual display of calibration manifested in the predictiveness or calibration plots. Although deciles of risk are typically used for the predictiveness plot, calibration plot and Hosmer-Lemeshow statistic, as mentioned above there is no reason that other subgroupings could not be used.

Achieving a Well Calibrated Model

This chapter does not cover procedures to estimate risk(X). We refer to textbooks that cover the topic in depth [9, 27]. When only a few predictors are included, the task of fitting a model that is well calibrated to the data and not over-fit is relatively straightforward. Although sampling variability in the fitted model remains, assuming good study design practices have been followed and in the absence of additional data, one will propose the fitted model for use in practice. The next task will be to evaluate its performance for prediction.

Sometimes an externally fitted model will be proposed for validation on a new dataset. If the model is not found to be well calibrated on the new dataset, it must be regarded as not having been validated. Nevertheless, some investigators proceed to evaluate its classification performance. In our opinion this is inappropriate. Individuals will want to use the prediction model not just as an aid in classification but to calculate their risk as a function of predictors. A poorly calibrated model is known to be invalid for this purpose and should be abandoned.

A better strategy perhaps is to derive a revised risk prediction model with the new dataset. This may be done by using the original estimated risk value as a sole predictor and fitting a model with that single predictor to the data. This is called *recalibration* [27]. Because only one predictor is involved it should be easy to arrive at a revised model that is well calibrated to the new dataset and therefore worthy of evaluation for its predictive performance.

Another strategy is to begin anew and fit a model with each predictor in the original model included as a candidate predictor for the new model. If many predictors are involved, issues pertaining to overfitting arise and one will need to use techniques such as ‘shrinkage’ in order to arrive at a believable model [9, 28]. Fruitfully applying such techniques requires considerable skill and experience. An advantage of starting over with the new dataset is that a combination of predictors that is closer to optimal for application in the target population may be

arrived at. Recalibrating an existing model is easier and subject to less sampling variability, but one maintains the same predictor combination of the original model that may be suboptimal if it was derived from a population that is not the one of interest.

Why might an externally derived model fail to be well calibrated? Certainly if it was derived in a different population with different predictor effects (or distributions, see section “The Meaning of Calibration”) it may not be valid for use in the target population. Another common cause is overfitting the model in the original population.

For the remainder of this chapter we assume that a model that is well calibrated is to be evaluated for its performance.

Measuring Prediction Performance of a Single Model

Context

The focus of this section is on describing conceptual approaches to evaluating a risk prediction model. We suppose that we have an extremely large population and the true risk function, $\text{risk}(X) = P(D = 1|X)$, is available. How do we measure the performance of this risk function for use in the population?

It is important to remember that the purpose of calculating $\text{risk}(X)$ is to affect medical decisions. Recall that we assume that in the absence of knowledge of $\text{risk}(X)$ no treatment will be offered, but that if $\text{risk}(X)$ is found to be large enough, treatment will be offered. Implicitly we assume that treatment must have associated with it some costs, e.g. toxicity, monetary costs, inconvenience. Otherwise all subjects, regardless of their risk value, could be treated even if the benefit was minimal.

Case and Control Risk Distributions

All metrics to gauge the performance of a risk model are derived from the distribution of $\text{risk}(X)$ in cases and in controls. Having a visual display of the distributions is often helpful. Although probability density functions (pdfs) (Fig. 4) give a sense of the separation between case and control distributions, cumulative distribution functions (cdfs), or 1 minus cdfs shown in Fig. 5 and denoted by $\text{HR}_D(r) = P(\text{risk}(X) > r|D = 1)$ and $\text{HR}_{\bar{D}}(r) \equiv P(\text{risk}(X) > r|D = 0)$, are more useful because they explicitly show the proportions of cases and controls above any threshold value used to define ‘high risk.’ Gauging this using the pdfs in Fig. 4 is difficult. Since decisions to opt for treatment will be based on ‘high risk’

Fig. 4 Case and control risk distributions on the logit scale

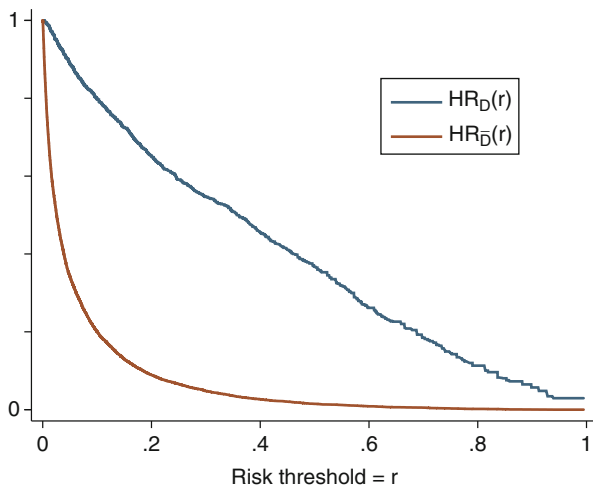
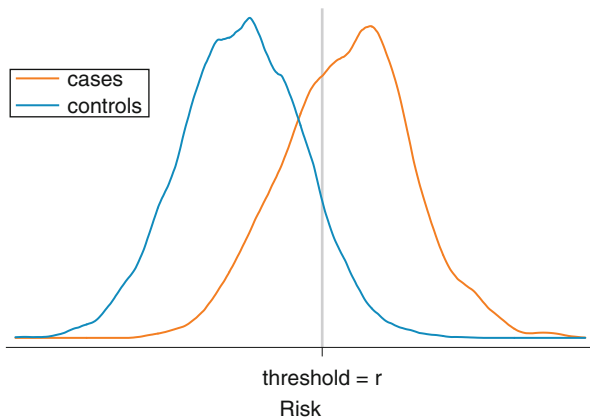


Fig. 5 Case and control distributions of risk shown with 1 minus cdfs, $HR_D(t) = P(\text{risk}(X) > t | D = 1)$ and $HR_{\bar{D}}(t) = P(\text{risk}(X) > t | D = 0)$

designation, it is of interest and can be seen directly from the cdfs how many cases and controls are recommended for treatment using the risk model. Note that $HR_D(r)$ is the true positive rate (TPR) or sensitivity and $HR_{\bar{D}}(r)$ is the false positive rate (FPR) or 1 minus specificity of the risk model using risk threshold r . True and false positive rates are common measures of the accuracy of general classification rules. We prefer to use the HR notation to emphasize what the classification rule is in this setting – it represents high risk designation.

Gail and Pfeiffer [7] make the point that the cdf of risk in the population as a whole, cases and controls together, is sufficient to calculate performance measures. This is true because the case distribution and the control distribution can both be

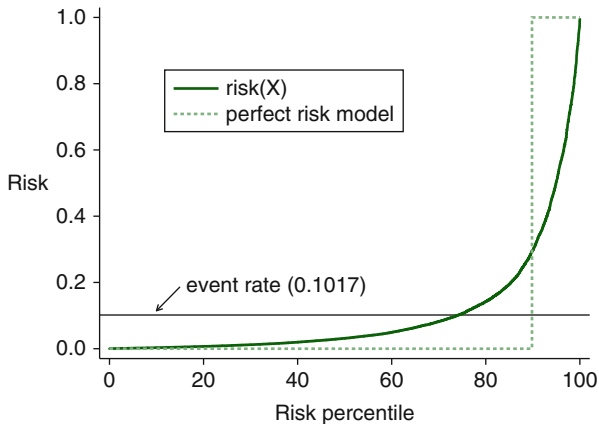


Fig. 6 The predictiveness curve that shows the quantiles of risk(X) in the population. The predictiveness curve for the ideal model where all cases have risk = 1 and all controls have risk = 0, is shown as a *dashed step* function

calculated from the overall population distribution of risk. Using $f(\text{risk}(X) = r)$ to denote the pdf for the true risk(X) at r , this follows because

$$\begin{aligned}
 f(\text{risk}(X) = r|D = 1) &= \frac{P(D = 1|\text{risk}(X) = r)}{P(D = 1)} f(\text{risk}(X) = r) \\
 &= \frac{r f(\text{risk}(X) = r)}{P(D = 1)} \\
 f(\text{risk}(X) = r|D = 0) &= \frac{(1 - r) f(\text{risk}(X) = r)}{P(D = 0)}
 \end{aligned}$$

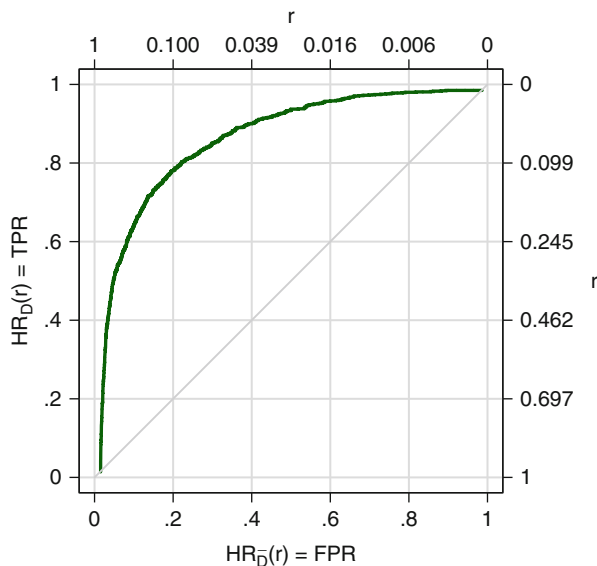
The overall cdf of risk(X) in the population as a whole is shown by the predictiveness curve that displays quantiles of risk in the population (Fig. 6). However we find the case- and control-specific cdfs shown in Fig. 5 to be a more informative display.

The *receiver operating characteristic* (ROC) curve that plots $\text{HR}_D(r)$ versus $\text{HR}_{\bar{D}}(r)$ (Fig. 7) is another popular visual display to assess performance. When the case and control distributions are well separated, the ROC curve

$$\text{ROC}(f) = \text{HR}_D(\text{HR}_{\bar{D}}^{-1}(f))$$

lies close to the upper left hand corner of the $[0, 1] \times [0, 1]$ quadrant; in contrast a useless risk model has an ROC curve that follows the diagonal 45° line. Huang and Pepe [11] show that the ROC curve and prevalence, $\rho = P[D = 1]$, together can be used to calculate the predictiveness curve, again assuming weak calibration of the risk model. And, since the predictiveness curve can be used to calculate the case and control distributions of risk, it follows that $(\text{ROC}(f), f \in (0, 1) ; \rho)$ contain all the

Fig. 7 ROC plot with risk thresholds r shown on top and right axes



information available in the case and control distributions of risk(X). However, the risk thresholds r corresponding to the points on the ROC curve, $(HR_{\bar{D}}(r), HR_D(r))$, are not visible from the ROC curve detracting from its interpretation. When plotting a single ROC curve it is possible to add the risk thresholds to the axes of the ROC plot as shown in Fig. 7. However, with two or more ROC curves this is not possible. Moreover since ROC curves do not align models according to the same risk thresholds and do not display risk thresholds they are less useful for evaluating prediction models than they are for evaluating diagnostic tests whose numeric scales are often irrelevant in data displays.

Risk Thresholds

How should one choose the risk threshold for designating a patient as at sufficiently high risk to warrant treatment? Intuitively the costs and benefits associated with treatment dictate the choice. If the treatment is very costly in terms of toxicities, monetary expense or inconvenience, a high threshold may be warranted. If the treatment is very likely to be effective at preventing a bad outcome, this might lower the threshold for treatment. In the extreme, an ineffective treatment or a prohibitively costly treatment, such as mastectomy for breast cancer prevention, dictates use of a risk threshold close to 1 corresponding to few people being treated. At the other extreme, a highly effective inexpensive treatment with few toxicities, such as statins for cardiovascular disease prevention, dictates that many people should be treated, i.e., use of a low risk threshold.

An explicit relationship is given in the next result between the risk threshold for treatment, r_H , and the net costs and benefits of treatment. Write the *net benefit of treatment* to a subject who would have an event in the absence of treatment as B . For example, if treatment reduces the risk of an event by 50%, the benefit might be $0.5 \times \{\text{value of an event}\}$ and the net benefit is the benefit less the costs associated with treatment for a subject who would otherwise have an event. Note that these costs must be put on the same scale as the benefit, i.e. $\{\text{value of an event}\}$ in our example. A subject who would not have an event in the absence of treatment suffers only costs and no benefit from being treated. We use C to denote the corresponding cost for such a subject.

Result 1. The risk threshold for treatment that should be employed to ensure subjects with risk values $\text{risk}(X)$ benefit on average is

$$r_H = C / (C + B).$$

The threshold does not depend on the model or on the predictors in the model, assuming the model is weakly calibrated.

Proof. The expected net benefit for subjects with $\text{risk}(X) = r$ is

$$B \cdot P(D = 1 | \text{risk}(X) = r) - C \cdot P(D = 0 | \text{risk}(X) = r) = B \cdot r - C \cdot (1 - r)$$

which is positive if

$$\frac{r}{1 - r} > \frac{C}{B}$$

In other words, to ensure a positive average benefit for subjects with risk values r they should opt for treatment if $r / (1 - r) > C / B$ and should not opt for treatment if $r / (1 - r) < C / B$. That is the treatment risk threshold is $C / (C + B)$. \square

The result agrees with the intuition that high costs and/or low benefits should be correspond to high values of the risk threshold for opting for treatment.

Example. Suppose treatment reduces the risk of breast cancer within 5 years by 50% but it can cause other bad outcomes such as other cancers, cardiovascular events and hip fractures that are considered equally bad. Assume other adverse outcomes, A , occur with a frequency of x in the absence of treatment but with a frequency of z in the presence of treatment regardless of whether $D = 1$ or 0. Let $D(0), D(1), A(0), A(1)$ denote the potential outcomes with and without treatment. A woman who would not develop breast cancer absent treatment suffers the increased risk of other bad events; her net cost of treatment is

$$\begin{aligned} &P(D = 1 \text{ or } A = 1 | T = 1, D(0) = 0) - P(D = 1 \text{ or } A = 1 | T = 0, D(0) = 0) \\ &= P(A = 1 | T = 1, D(0) = 0) - P(A = 1 | T = 0, D(0) = 0) \\ &= P(A = 1 | T = 1) - P(A = 1 | T = 0) = z - x, \end{aligned}$$

where the second line uses the assumption that treatment cannot cause disease, i.e. $D(0) = 0$ implies $D(1) = 0$. The net benefit of treating a subject who would get breast cancer absent treatment is the reduction in her event probability,

$$\begin{aligned} & P(D = 1 \text{ or } A = 1 | T = 0, D(0) = 1) - P(D = 1 \text{ or } A = 1 | T = 1, D(0) = 1) \\ &= 1 - [P(D = 1 | T = 1, D(0) = 1) + P(D = 0 \text{ and } A = 1 | T = 1, D(0) = 1)] \\ &= 1 - [0.5 + (1 - 0.5) \cdot P(A = 1 | T = 1)] = 0.5 + 0.5z \end{aligned}$$

The treatment risk threshold is therefore $\frac{z-x}{z-x+0.5(1+z)}$.

In practice it is often difficult to specify costs and benefits associated with treatment. It can be especially difficult to specify them on a common scale when they are qualitatively different entities. On the other hand a treatment threshold for risk is often easier to specify. For example, the ATP guidelines recommend that subjects with risks above 20% consider longterm treatment with cholesterol lowering therapy to reduce risk of cardiovascular events. Individuals make decisions such as whether or not to have genetic testing of their fetus based on their risk of having a child with genetic abnormalities. Their chosen risk threshold is often derived intuitively from their knowledge of the qualitative costs and benefits of amniocentesis. Similarly we make decisions about procuring insurance using our tolerance for risk. Result 1 tells us the explicit relationship between the risk threshold and the perceived cost-benefit ratio. For example, by choosing a risk threshold equal to 20% for cholesterol lowering therapy, we are implicitly stating that the net benefit of therapy for a would-be case is 4 times the net cost of therapy for a would-be control because $\frac{r_H}{1-r_H} = .2/(1-.2) = 1/4$.

Summary Statistics when a Risk Threshold Is Available

In this section we consider settings where a risk threshold, r_H , exists that defines high risk status with possible recommendation for treatments or, perhaps, for entry into a clinical trial. The context is then essentially reduced to a binary classification rule, high risk or not high risk, and measures commonly used to summarize performance of binary classifiers are appropriate. We already defined the proportions of cases and controls classified as high risk as

$$\begin{aligned} \text{HR}_D(r_H) &= P(\text{risk}(X) > r_H | D = 1) \\ \text{HR}_{\bar{D}}(r_H) &= P(\text{risk}(X) > r_H | D = 0). \end{aligned}$$

A perfect model classifies all cases and no controls as high risk, $\text{HR}_D(r_H) = 1$ and $\text{HR}_{\bar{D}}(r_H) = 0$. A good model classifies a large proportion of cases as high risk and a low proportion of controls as high risk.

The $\text{HR}_D(r_H)$ is a good attribute of a prediction model while $\text{HR}_{\bar{D}}(r_H)$ is a negative attribute. The expected population net benefit of using the model with risk threshold r_H , $\text{NB}(r_H)$, combines the two into an overall population measure that balances the positive and negative attributes:

$$\begin{aligned}\text{NB}(r_H) &= P(\text{risk}(X) > r_H) \{B \cdot P(D = 1 | \text{risk}(X) > r_H) - C \cdot P(D = 0 | \text{risk}(X) > r_H)\} \\ &= B \cdot P(\text{risk}(X) > r_H | D = 1) P(D = 1) - C \cdot P(\text{risk}(X) > r_H | D = 0) P(D = 0) \\ &= B \cdot \text{HR}_D(r_H) \rho - C \cdot \text{HR}_{\bar{D}}(r_H) (1 - \rho).\end{aligned}$$

Observe that $\text{NB}(r_H)$ is an expectation over the entire population and assumes treatment is offered if $\text{risk}(X) > r_H$. In contrast the expected net benefit in the proof of Result 1 concerns only subjects with $\text{risk}(X) = r$ and considers their net benefit if they receive treatment.

Recall that Result 1 tells us that use of the risk threshold r_H implies $C/B = r_H/(1 - r_H)$. Substituting into the above gives us an expression for expected net benefit that depends only on model performance parameters ($\text{HR}_D(r_H), \text{HR}_{\bar{D}}(r_H)$) and the constants (ρ, r_H).

$$\text{NB}(r_H) = \left\{ \rho \text{HR}_D(r_H) - \frac{r_H}{1 - r_H} (1 - \rho) \text{HR}_{\bar{D}}(r_H) \right\} B$$

Vickers and Elkin [29] propose measuring net benefit in units that assign B a value 1. In those units $\text{NB}(r_H) = \rho \text{HR}_D(r_H) - \frac{r_H}{1 - r_H} (1 - \rho) \text{HR}_{\bar{D}}(r_H)$. The inverse of $\text{NB}(r_H)$ can be interpreted as the number of X measurements required to yield the benefit corresponding to detecting one true positive and no false positives. Baker and Kramer [2] call this the number needed to test (NNT).

A standardized version of $\text{NB}(r_H)$ is found by dividing $\text{NB}(r_H)$ by the maximum possible value that can be achieved, namely ρB , corresponding to a perfect model with $\text{HR}_D(r_H) = 1$ and $\text{HR}_{\bar{D}}(r_H) = 0$. Baker et al. used the term ‘relative utility’ but we prefer the more descriptive term ‘standardized net benefit’ and use notation that corresponds:

$$\begin{aligned}s\text{NB}(r_H) &= \text{NB}(r_H) / \max(\text{NB}(r_H)) = \text{NB}(r_H) / \rho B \\ &= \text{HR}_D(r_H) - \frac{r_H}{(1 - r_H)} \frac{(1 - \rho)}{\rho} \text{HR}_{\bar{D}}(r_H).\end{aligned}$$

An advantage of standardizing net benefit is that it no longer depends on the measurement unit B , an entity that is sometimes difficult to digest. $s\text{NB}(r_H)$ is a unit-less numerical summary in the range $(0,1)$.

Another interpretation for $s\text{NB}(r_H)$ is that it discounts the true positive rate $\text{HR}_D(r_H)$ using the scaled false positive rate $\text{HR}_{\bar{D}}(t)$ to yield a discounted true positive rate. The FPR is scaled so that the units are on the same scale as the TPR. $s\text{NB}(r_H)$ can therefore be interpreted as the true positive rate of a prediction model that has no false positives but equal benefit.

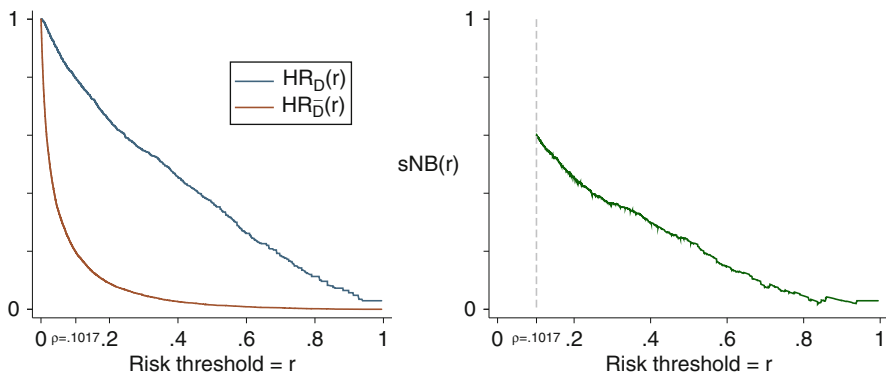


Fig. 8 Proportions of cases and controls above risk threshold r and corresponding standardized net benefit

In our opinion a reasonably complete reporting of a prediction model’s performance is given by $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and $sNB(r_H)$. One also needs to keep in mind the prevalence, ρ , and the risk threshold, r_H , in order to interpret $sNB(r_H)$ and its components ($HR_D(r_H)$, $HR_{\bar{D}}(r_H)$).

Example. Figure 8 shows $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and their weighted average $sNB(r_H)$ for various choices of risk threshold, r_H . For example, at $r_H = 0.2$ we have that 65.2% of cases and 8.9% of controls are classified as high risk. This corresponds to a benefit that is 45.6% of the maximum possible benefit, where the maximum possible benefit would be achieved by classifying all 10.17% cases and no controls as high risk. Alternatively we can consider that the observed true positive rate of 65.2% is discounted to 45.6% by the 8.9% of controls that are designated as high risk.

The plot in Fig. 8 allows one to view the performance achieved with different choices of risk threshold. Observe that the net benefit curve is plotted only for $r_H > \rho$. This is because the assumed default action is to *not* treat subjects. In the absence of predictors, all subjects are assigned risk values ρ . Therefore a risk value of ρ must correspond to the ‘no treatment’ rule. To be consistent and rational we still assign no treatment if $risk(X) < \rho$ when predictors are available. Therefore risk thresholds for treatment below ρ are not relevant. If one were to instead assign treatment as the default decision and use the model for decisions to forego treatment, then the expected net benefit and its standardized version would be calculated differently: $sNB(r_L) = (1 - HR_{\bar{D}}(r_L)) - \frac{\rho}{(1-\rho)} \frac{(1-r_L)}{r_L} (1 - HR_D(r_L))$, where r_L is the low risk threshold below which subjects would *not* receive treatment. Moreover this version of $sNB(r_L)$ would only be calculated for $r_L < \rho$. See Baker [1] for details.

In many circumstances a fixed risk threshold for assigning treatment does not exist. It can be useful to consider a variety of thresholds. Consider that a prediction model is often developed to determine eligibility criteria for a clinical trial of a new treatment. For example, new treatments for acute kidney injury are under development and prediction models are sought to identify high risk subjects for

Table 2 Proportions of subjects in each of 3 risk categories. Risk refers to 10-year probability of a cardiovascular event

Risk category	Cases	Controls
Low (<5%)	.112	.657
Medium (5–20%)	.236	.253
High (>20%)	.652	.089

upcoming clinical trials. Potentially toxic or expensive treatments will require higher risk thresholds than less toxic, inexpensive treatments. Having displays that show performance across a range of risk thresholds will allow researchers to entertain use of risk prediction models for designing trials of different types of treatment.

Another important reason to display performance as a function of risk threshold is that it allows individuals with different tolerances to assess the value of ascertaining predictor information for them. If the distribution of risk thresholds among individuals in the population were known, those could be overlaid on the plots. One could summarize the information by integrating over the distribution of risk thresholds: $HR_{\bar{D}} = E(HR_{\bar{D}}(r_H))$ = the overall proportions of controls that do not receive treatment; $HR_D = E(HR_D(r_H))$ = the overall proportion of cases who receive treatment; and $E(NB(r_H))$ = expected net benefit.

Although we emphasize $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and $sNB(r_H)$ as the key measures of prediction performance for settings where a risk threshold reduces the model to a binary decision rule, other measures of performance for binary classifiers could also be reported. Classic measures include: the misclassification rate, $(1 - HR_D(r_H))\rho + HR_{\bar{D}}(r_H)(1 - \rho)$; Youden’s index, $HR_D(r_H) - HR_{\bar{D}}(r_H)$; and the Brier score, $E(D - risk(X))^2$. These measures, like $sNB(r_H)$, are functions of $HR_D(r_H)$ and $HR_{\bar{D}}(r_H)$ but seem to lack its compelling interpretation and practical relevance. Therefore we do not endorse them for evaluating risk prediction models.

Multiple Risk Categories

For prevention of cardiovascular events two possible treatment strategies are recommended. Long-term treatment with cholesterol lowering drugs is recommended for high risk subjects while an inexpensive, non-toxic intervention, namely healthy lifestyle changes, is recommended for subjects at moderately elevated risk. Three risk categories are therefore of interest in this clinical setting: low risk (risk $\leq 5\%$), moderate risk (5–20%) and high risk ($\geq 20\%$). The parameters HR_D and $HR_{\bar{D}}$ are easily generalized to the setting of multiple risk categories. One reports the proportions of cases in each category and the proportions of controls in each category. These fractions can be read off of the distribution function displays in Fig. 8. Values are shown in Table 2.

The standardized net benefit function can also be generalized to accommodate more than two categories of risk, but it requires specifying some relative costs and benefits explicitly in addition to the cost-benefit ratios that are implied by the risk threshold values that define the risk categories.

As an example, suppose there are 3 categories of risk with treatment recommendations being none for the low risk category ($\text{risk} \leq r_{\text{low}}$), intermediate treatment for the medium risk category ($r_{\text{low}} \leq \text{risk} < r_{\text{high}}$) and intense treatment for the high risk category ($\text{risk} > r_{\text{high}}$). Suppose that in the absence of a predictive model all subjects are recommended intermediate treatment. We use the following notation for costs and benefits: B_{high} = net benefit of intense treatment to a subject who would be a case; C_{high} = net cost of intense treatment to a would-be control; B_{low} = net benefit of no treatment to a would-be control and C_{low} = net cost of no treatment to a would-be case. All of these quantities are relative to the intermediate treatment and as always, cases (controls) are those who would (would not) get the bad outcome in the absence of treatment. The expected net benefit in the population associated with use of the risk model is:

$$\rho\{-C_{\text{low}}\text{LR}_D + B_{\text{high}}\text{HR}_D\} + (1 - \rho)\{B_{\text{low}}\text{LR}_{\bar{D}} - C_{\text{high}}\text{HR}_{\bar{D}}\}$$

where LR and HR are probabilities of being in the low and high risk categories and the subscripts D and \bar{D} indicate cases and controls as usual. Let r_L and r_H denote the risk thresholds that separate low from medium risk and medium from high risk, respectively. The arguments of Result 1 implies that $(r_L/1 - r_L) = B_{\text{low}}/C_{\text{low}}$ and $(r_H/1 - r_H) = C_{\text{high}}/B_{\text{high}}$. Let $\lambda = B_{\text{high}}/C_{\text{low}}$ be the ratio of the net benefit of intense treatment to the net cost of no treatment for a subject that would be a case. The population net benefit of using the model with these risk categories can then be written as

$$B_{\text{high}} \left\{ \rho \left\{ -\frac{1}{\lambda} \text{LR}_D + \text{HR}_D \right\} + (1 - \rho) \left\{ \frac{r_L}{(1 - r_L)\lambda} \text{LR}_{\bar{D}} - \frac{r_H}{1 - r_H} \text{HR}_{\bar{D}} \right\} \right\}.$$

This can be standardized by the maximum possible benefit that is achieved with a perfect prediction model, $B_{\text{high}}\{\rho + (1 - \rho)\frac{r_L}{(1 - r_L)\lambda}\}$, yielding the standardized net benefit function

$$s\text{NB}(r_M, r_H) = \frac{\rho\{\text{HR}_D - \text{LR}_D/\lambda\} + (1 - \rho)\left\{\frac{r_L}{(1 - r_L)\lambda} \text{LR}_{\bar{D}} - \frac{r_H}{(1 - r_H)} \text{HR}_{\bar{D}}\right\}}{\rho + (1 - \rho)\frac{r_L}{(1 - r_L)\lambda}}$$

This expression is a function of the risk thresholds, prevalence, and case and control risk distributions, and it requires specifying another parameter, namely λ . If we assume the net benefit of cholesterol lowering treatment relative to healthy lifestyle is 10 times as large as the net cost of no treatment relative to healthy lifestyle intervention to a would-be case, then the standardized net benefit associated with the fitted model $\text{risk}(X)$ is

$$\begin{aligned} & \frac{0.1017\{0.652 - 0.112/10\} - 0.8983\left\{\frac{0.05}{0.95}\frac{0.657}{10} - \frac{0.20}{0.80} \times 0.089\right\}}{0.1017 + 0.8983 \times \frac{0.05}{0.95 \times 10}} \\ & = 77.1\% \end{aligned}$$

This example demonstrates that calculation of net benefit associated with a risk model becomes quite complicated with three treatment categories compared with the calculation when only two treatment categories exist.

Implicit Use of Risk Thresholds

In some circumstances a risk threshold for treatment that maximizes expected benefit cannot be adopted. Policy makers may require that alternative criteria are met. For example, Pfeiffer and Gail [24] consider using a risk threshold that results in a proportion v of the population recommended for treatment. Allocation of financial resources might determine such a policy. The risk threshold is written as

$$r_H(v) : v = P(\text{risk}(X) > r_H(v)).$$

Having fixed v , they propose the proportion of cases that meet the treatment threshold as a measure of model performance:

$$\text{PCF}(v) = P(\text{risk}(X) > r_H(v) | D = 1),$$

with larger values indicating better performance. Observe that in our previous notation we can write

$$\text{PCF}(v) = \text{HR}_D(r_H(v)).$$

Another policy based criterion might require that a fixed proportion of the cases are recommended for treatment. In this case the treatment threshold is

$$r_H(w) : w = P(\text{risk}(X) > r_H(w) | D = 1)$$

and the prediction model performance measure proposed by Pfeiffer and Gail is the corresponding proportion of the population needed to follow, i.e., testing positive,

$$\text{PNF}(w) = P(\text{risk}(X) > r_H(w)).$$

Smaller values of $\text{PNF}(w)$ are more desirable.

These measures are closely related to the ROC curve that plots the true positive rate $\text{ROC}(f) = P(\text{risk}(X) > r_H | D = 1)$ versus the false positive rate $f = P(\text{risk}(X) > r_H | D = 0)$ for all possible thresholds $r_H \in (0, 1)$. In fact a little algebra shows that

$$\text{PNF}(w) = \rho w + (1 - \rho) \text{ROC}^{-1}(w)$$

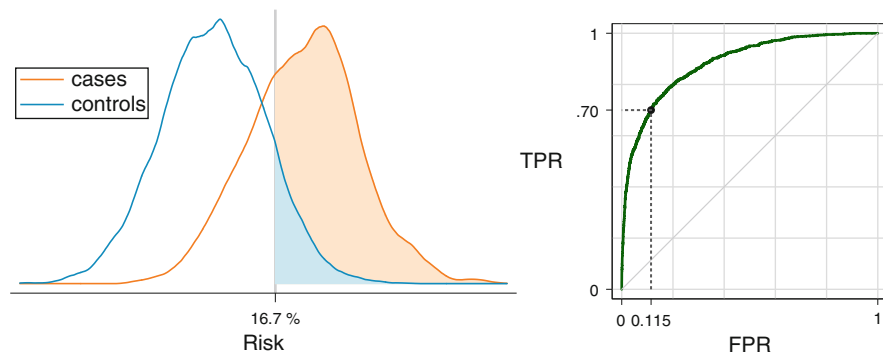


Fig. 9 Density of risk among cases and controls and the ROC curve for risk. For the risk threshold that yields 70% of cases at high risk, 11.5% of controls are also designated high risk

and

$$PCF(v) = ROC(f(v))$$

where $f(v)$ is found by solving $v = \rho ROC(f) + (1 - \rho)f$.

One can also directly use ROC curve points to characterize performance and it follows from arguments above that this approach is essentially equivalent to Pfeiffer and Gail’s approach. In ROC analysis one fixes the proportion of cases deemed at high risk, derives the corresponding threshold $r_H(w)$ defined above, and evaluates the corresponding proportion of controls classified as high risk,

$$ROC^{-1}(w) = P(\text{risk}(X) > r_H(w) | D = 0).$$

Alternatively, one can fix the proportion of controls classified as high risk at f , derive the corresponding threshold

$$r_H(f) : f = P(\text{risk}(X) > r_H(f) | D = 0)$$

and use as the performance measure the corresponding proportion of cases classified as high risk

$$ROC(f) = P(\text{risk}(X) > r_H(f) | D = 1).$$

Example. Using our dataset, suppose we require that $w = 70\%$ of cases go forward for treatment. We calculate that the corresponding risk threshold will be $r_H(w) = 0.167$ and that 11.5% of controls will exceed this threshold with use of the model (see Fig. 9). Since the prevalence is 10.17%, the overall proportion of the population that will undergo treatment is 17.5%. Using our notation:

$$w = 70\%, \quad r_H(w) = 0.167, \quad ROC^{-1}(w) = .115, \quad PNF(w) = .175$$

Measures Independent of Risk Thresholds

When a risk threshold for decision making is not forthcoming, a descriptive summary of the risk distributions in cases and controls may be of interest. In particular, one can describe the separation between the case and control distributions of risk. For example, we could report: the average risk in cases, 0.391 in our example; the average risk in controls, 0.069 in our example, and the difference that we write as MRD, the *mean risk difference*:

$$\text{MRD} = E(\text{risk}(X)|D = 1) - E(\text{risk}(X)|D = 0)$$

which is 0.322 in our example.

The MRD statistic is also called Yates' slope. It is closely related to the integrated discrimination improvement (IDI) statistic proposed by Pencina et al. [17] for comparing nested risk prediction models. Specifically, if we consider the baseline model as the null model without predictors, so all subjects have estimated risk equal to ρ , then the IDI for comparing the model, $\text{risk}(X)$, with the null model is the MRD. It has also been shown [21] that the MRD can be interpreted as the proportion of explained variation or coefficient of determination, $R^2 = \text{var}\{E(D|X)\}/\text{var}(D)$.

Another way to summarize the distance between the case and control risk distributions is with the *above average risk difference*, AARD:

$$\text{AARD} \equiv P(\text{risk}(X) > \rho|D = 1) - P(\text{risk}(X) > \rho|D = 0).$$

Noting that the average risk in the population is ρ , this measure compares the proportion of cases with risks exceeding ρ , $\text{HR}_D(\rho) = 0.797$ in our example, with the corresponding proportion of controls, $\text{HR}_{\bar{D}}(\rho) = 0.198$, in our example, and calculates the difference, $\text{AARD} = 0.797 - 0.198 = 0.599$.

The AARD has several additional noteworthy interpretations. First, Youden's index for a dichotomous diagnostic test is defined as the true positive rate minus the false positive rate. We see that AARD is Youden's index for the rule that classifies subjects as positive when $\text{risk}(X) > \rho$. Second, we see that $\text{AARD} = s\text{NB}(\rho)$, the standardized net benefit defined earlier, for the decision rule that uses ρ as the high risk threshold. Third, the AARD is closely related to the net reclassification index (NRI) that will be defined in section "The Net Reclassification Index". The NRI is currently a very popular measure for comparing nested models. It can be shown that for comparing the model with predictors X , $\text{risk}(X)$, to the null model that assigns all subjects a risk of ρ , $\text{AARD} = \text{NRI}/2$ and $\text{AARD} = c\text{NRI}(\rho)/2$ where NRI is known as the continuous NRI and $c\text{NRI}(\rho)$ is the categorical NRI with two risk categories defined by the risk threshold ρ .

Interestingly, it can be shown that the difference, $\text{HR}_D(r) - \text{HR}_{\bar{D}}(r)$, is maximized at $r = \rho$ (see proof of Theorem A.1 [23]). Therefore, the AARD can also be interpreted as a Kolmogorov-Smirnov distance between the case and control risk distributions. Finally, Huang and Pepe [11] and Gu and Pepe [8] showed that the

standardized total gain statistic proposed by Bura and Gastwirth [3] as a measure of predictive capacity of a model $\text{risk}(X)$, $sTG = \int |\text{risk}(X) - \rho| dF(X) / 2\rho(1 - \rho)$ is equal to the AARD.

Another nonparametric measure of distance between the case and control risk distributions is the area under the ROC curve (AUC):

$$\text{AUC} = P(\text{risk}(X_i) \geq \text{risk}(X_j) | D_i = 1, D_j = 0)$$

where X_i and X_j are predictors for randomly drawn independent subjects from the case and control distributions, respectively. This is also known as the Mann-Whitney U-statistic and is calculated as $\text{AUC} = 0.884$ for our data. The AUC has a long history of use in evaluating diagnostic tests and other classifiers including risk models. It is still the most popular metric in use. However, its use in evaluating risk models has been criticized [4, 20]. One of the criticisms leveled against the AUC is that the measure has no practical relevance. Certainly this is true. If subjects were presented in case-control pairs to the physician for deciphering which one is the case, the measure $P(\text{risk}(X_i) > \text{risk}(X_j) | D_i = 1, D_j = 0)$ would be of interest. But this is not the usual clinical task. Another criticism of the AUC is that the measure may be dominated by differences in risk distributions that are clinically irrelevant. In particular in a low prevalence or incidence setting, the AUC is dominated by the low end of the risk range where most of the population's risks lie. Yet small differences in the distributions over that range are of no clinical relevance. Consequently the AUC may not be sensitive to differences in risk distributions over more clinically relevant ranges.

These two criticisms, practical irrelevance and insensitivity to clinically important differences in distributions, however, apply not only to the AUC, but also apply to other measures of distance between case and control risk distributions such as the MRD and AARD. We see no particular advantage to MRD or AARD or the related reclassification measures (IDI and NRI) that will be described later. We caution against using any of these measures as a sole focus for evaluating and comparing models. Rather they may be more suitable to assessing if a model is at one extreme or the other in terms of prediction performance. As such, their use may be justified in algorithms to sift through many models in order to select some that have predictive performance worthy of more thorough evaluation, i.e., discovery research.

Recommendations

For evaluating a single risk prediction model we have the following recommendations:

- (i) Assess the model for its calibration in the population of interest and if necessary recalibrate the model.
- (ii) Plot the case and control risk distributions, possibly providing a summary index of distance between the distributions as a descriptive adjunct.

- (iii) In collaboration with clinical and health policy colleagues, elicit a risk threshold or several thresholds that could be used for making treatment decisions. Evaluate the model in terms of corresponding case and control classification rates and in terms of net benefit of using the model.

Comparing Two Risk Models

In this section we consider the comparison of two risk models for their prediction performance. In a nutshell we recommend that (i) each model be evaluated for its calibration; (ii) plots of risk distributions and net benefit be prepared for each model; (iii) having chosen a clinically relevant risk threshold for treatment (or several), compare the corresponding case and control categorized risk distributions for the two models and (iv) compare the corresponding net benefits associated with the two models. An illustrative example is provided in section “Example”. This approach applies when the two models are nested (where one model has predictors X and the other has additional predictors Y) and when the models are not nested. Several methods have been proposed in recent years for the specific problem of comparing nested models, notably risk reclassification methods. We describe those risk reclassification methods in section “Risk Reclassification to Compare Two Models”. Finally, we provide a result concerning the equivalence of null hypotheses about improvement in prediction performance gained by adding Y to a set of baseline predictors X and the classic null hypothesis about Y as a risk factor after controlling for X .

Example

We compare the logistic models, $\text{risk}(X)$ and $\text{risk}(X, Y)$, fit to our illustrative data, as described in section “Illustrative Data”. Predictiveness curves for the fitted models superimposed with observed event rates in each decile of fitted risk are shown in Fig. 10. Both models are well calibrated. The risk distributions in cases and controls are shown in the left panel of Fig. 11 using 1 minus cdfs. That is, for each risk threshold r we show the proportions of cases and controls above that threshold. We see that at all thresholds more cases and fewer controls have risks above the threshold when using the model $\text{risk}(X, Y)$ than with model $\text{risk}(X)$. Consequently the case and control risk distributions are more separated by the model $\text{risk}(X, Y)$ than by the model $\text{risk}(X)$. The measures of separation, AUC, MRD, and AARD, are presented in Table 3, and confirm that greater separation is achieved with the model including Y as a predictor.

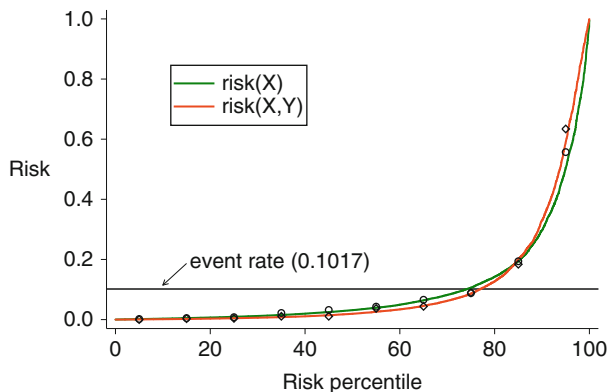


Fig. 10 Event rates for subjects in each decile of estimated risk align well with the risk values for subjects in each decile

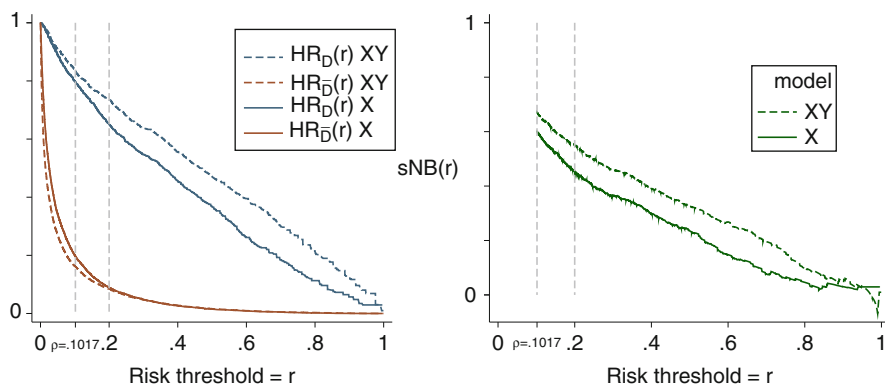


Fig. 11 Plots showing high risk classification for cases (D) and controls (\bar{D}) under the baseline model risk(X) and the expanded model risk(X, Y). A comparison of standardized net benefit is also shown

Table 3 Summary measures of performance for baseline and expanded models

	risk(X)	risk(X, Y)	Difference
AUC	0.884	0.920	0.036
MRD	0.322	0.416	0.094 ^a
AARD	0.599	0.673	0.074
$HR_D(0.20)$	0.652	0.735	0.084
$HR_{\bar{D}}(0.20)$	0.089	0.084	-0.005
$sNB(0.20)$	0.455	0.550	0.095
$PNF(0.70)$	0.174	0.134	-0.040

^aDifference in MRDs is called the IDI [17]

The right hand side of Fig. 11 shows that, regardless of which risk threshold is employed for recommending treatment, the standardized net benefit is larger when Y is included in the risk model. This is not surprising since the change in the standardized net benefit is a weighted sum of the increase in the value $HR_D(r)$ plus the decrease in the value of $HR_{\bar{D}}(r)$, both of which are positive, i.e.,

$$sNB^{(X,Y)}(r) - sNB^X(r) = \{HR_D^{(X,Y)}(r) - HR_D^X(r)\} + \frac{(1-\rho)}{\rho} \frac{r}{(1-r)} \times \{HR_{\bar{D}}^X(r) - HR_{\bar{D}}^{(X,Y)}(r)\},$$

where X and X,Y superscripts denote measures of performance for risk(X) and risk(X,Y) models, respectively.

Suppose that subjects with risks above 20% are recommended for treatment because the net benefit of treatment to a (would-be) case is considered 4 times the net cost of treatment to a (would be) control. The model that includes Y recommends 8.4% more cases and .5% fewer controls for treatment. Relative to a perfect prediction model, the model with X only achieves 45.5% of maximum benefit while that including Y as well as X achieves 55.0% of maximum benefit. That is, the standardized net benefit or discounted true positive rate is improved by 9.5%.

We also consider performance when certain criteria are set by policy makers. Suppose that a treatment risk threshold will be employed that will guarantee $w = 70\%$ of cases are treated. Using the model with X only will require treating $PNF(w) = 17.4\%$ of the population (and correspondingly $ROC^{-1}(w) = 0.114$ of controls) since the largest risk threshold exceeded by 70% of cases is $r_H^X(w) = 0.167$. On the other hand, a higher risk threshold $r_H^{(X,Y)}(w) = 0.231$, can be employed with the model risk(X,Y) as the case risk distribution is higher. Consequently only $PNF(w) = 13.4\%$ of the population (and only $ROC^{-1}(w) = 0.07$ of controls) will be treated if risk(X,Y) is used for assigning treatment.

In this example, better performance is achieved by including Y in the risk model regardless of how performance is measured.

Risk Reclassification Within Subpopulations Defined by risk(X)

In addition to determining whether or not use of the model risk(X,Y) is better than use of risk(X) in the population as a whole, one might ask if the additional information provided by knowledge of Y is useful in subsets of the population. Specifically, one might consider subsets of the population determined to be at low (or high or intermediate) risk according to risk(X) and evaluate use of the expanded model risk(X,Y) in that subpopulation. This is one motivation for constructing the risk reclassification table illustrated in Table 4.

Table 4 Event and nonevent risk reclassification tables

Events				
$r(X)$	risk(X, Y)			Total
	<5%	5–20%	≥20%	
≤5%	72	38	4	114 (11.2%)
5–20%	21	105	114	240 (23.6%)
≥20%	0	33	630	663 (65.2%)
Total	93 (9.1%)	176 (17.3%)	748 (73.5%)	1017
Nonevents				
$r(X)$	risk(X, Y)			Total
	<5%	5–20%	≥20%	
≤5%	5486	399	21	5906 (65.8%)
5–20%	1015	990	272	2277 (25.3%)
≥20%	40	296	464	800 (8.9%)
Total	6541 (72.8%)	1685 (18.8%)	757 (8.4%)	8983

Table 5 Performance of $r(X, Y)$ within strata defined by $r(X)$

Population	$\rho(X)$ Event rate (%)	Cases $HR_D(0.20)$	Controls $HR_{\bar{D}}(0.20)$	% of max benefit $sNB(0.20)$
Low risk $r(X)$	1.89	0.035	0.004	−1.7
Med risk $r(X)$	9.54	0.475	0.119	19.3
High risk $r(X)$	45.32	0.950	0.580	25.4

Table 4 shows risk reclassification tables for cases in the ‘Events’ panel of the table and for controls in the ‘Nonevents’ panel of the table. In cardiovascular disease the tables are often constructed using 3 categories corresponding to the 3 treatment recommendations. Here we focus on the most important reclassifications to or from the high risk category that corresponds to cholesterol lowering treatment, and ignore reclassification between the medium and low risk categories that are less consequential. Table 5 summarizes the performance of the expanded model $risk(X, Y)$ within each of the subpopulations determined to be at low (<5%), medium(5–20%) and high(>20%) risk according to the baseline model, $risk(X)$.

In the low risk population, where the event rate $\rho = 1.89\%$, only 3.5% of cases are reclassified by Y to the high risk category and almost no controls (0.4%) are reclassified. The maximum possible benefit is to reclassify all cases (18.9 per 1000 subjects) and no controls. Consequently, the standardized net benefit of using the model is negligible. (The negative value, −1.7%, must be due to sampling variability as the net benefit cannot be negative if the $risk(X)$ and $risk(X, Y)$ models are correct, which is true for our simulated data.) It appears that use of Y in the low risk population is not beneficial.

In the medium risk population, 47.5% of the cases are favorably reclassified to the high risk group while only 11.9% of the controls are. The maximum possible benefit is that achieved if all 95.4/1000 cases were moved to the high risk category

without moving any of the 904.6/1000 controls. With use of Y it appears that $sNB(0.2) = 19.3\%$ of this maximum possible benefit is achieved. In other words the benefit reached by measuring Y in the medium risk population is the same as that achieved by a model that moved about 193/1000 cases to treatment with statins without moving any controls into that high risk category.

In the high risk group, benefit can be obtained by moving controls down to a lower risk category but at the possible cost of moving cases down. The reclassification tables show however that with use of Y only 5% of the cases are moved down while 42% of controls move down. The maximum benefit in a population of 1000 subjects would be that achieved by moving none of the 453 cases but all of the 547 controls down. With use of Y , we are able to move 230 controls off treatment at the expense of moving 23 cases off treatment. Is this a net benefit? Arguments similar to those in section “Summary Statistics When a Risk Threshold is Available” can be used to show the standardized net benefit of a rule that denies treatment when the risk $< r_H$ in a population with prevalence ρ is

$$sNB(r_H) = \{1 - HR_{\bar{D}}\} - \frac{\rho}{1 - \rho} \frac{1 - r_H}{r_H} \{1 - HR_D(r_H)\}.$$

We calculate that $sNB(r_H) = 25.4\%$ of maximum benefit is achieved with use of Y in the population deemed at high risk according to risk(X). This is equivalent to moving $.254 \times 547 = 140$ controls down without moving any cases down in a set of 1000 subjects. This benefit seems substantial.

We find risk reclassification tables useful for evaluating an expanded model risk(X, Y) within subpopulations defined by risk levels calculated according to the baseline model risk(X), as illustrated above. One might also choose to plot risk distributions and calculate other statistics for evaluating a risk model within each subpopulation using techniques described in section “Measuring Prediction Performance of a Single Model”. However, there are other analyses of risk reclassification tables that have been proposed for purposes beyond evaluation use of risk(X, Y) within subpopulations defined by risk(X). In particular, analyses intended to compare the two models in the entire population have been proposed. In the next section we describe the two main approaches and point out problems encountered when using these approaches to compare risk models.

Risk Reclassification to Compare Two Models

The Cook and Ridker Analysis Method

Cook and Ridker [5] combine the event and nonevent reclassification tables in a single table with the elements shown in Table 6.

Table 6 Risk reclassification tables showing numbers of subjects and event rates (%) in each cell

$r(X)$	risk(X, Y)			Total
	$\leq 5\%$	5–20%	$> 20\%$	
$\leq 5\%$	5558	437	25	6020
	1.30	8.71	16.00	1.89
5–20 %	1036	1095	386	2517
	2.03	9.59	29.53	9.54
$> 20\%$	40	329	1094	1463
	0.00	10.03	57.59	45.32
Total	6634	1861	1505	10,000
	1.40	9.46	49.70	10.17

They calculate the following entities that are explained below:

- (i) The overall reclassification rate = RC
- (ii) The percent correctly reclassified = RC-correct
- (iii) The baseline model reclassification calibration statistic: RCC^X and its p-value
- (iv) The expanded model reclassification calibration statistic: $RCC^{(X,Y)}$ and its p-value

The RC is the proportion of subjects in the off-diagonal cells of the table, 22.5% in our example. This is a descriptive statistic that is not useful for comparing risk models. A small RC value indicates that treatment recommendations will be changed for few subjects by measuring Y in addition to X , and a large value indicates that many subjects will have different treatment recommendations when Y is added to X .

The RC-correct is defined to be the proportion of subjects in off-diagonal cells where the observed event rate is within the risk(X, Y) category label and not within the risk(X) category label. In our data RC-correct = 100%. If the RC-correct value is large, this is taken as evidence that prediction performance with the model that includes Y is better than prediction performance with the baseline model risk(X). However, it has been shown that by definition, when models are well calibrated in the standard sense, RC-correct $\approx 100\%$ in large samples. This follows because for observations in an off-diagonal cell where risk(X) $\in A$ and risk(X, Y) $\in B$ the expected event rate

$$\begin{aligned}
 &P(D = 1 | \text{risk}(X) \in A, \text{risk}(X, Y) \in B) \\
 &= E(D | \text{risk}(X) \in A, \text{risk}(X, Y) \in B) \\
 &= E(E(D | X, Y) | \text{risk}(X) \in A, \text{risk}(X, Y) \in B) \\
 &= E(\text{risk}(X, Y) | \text{risk}(X) \in A, \text{risk}(X, Y) \in B).
 \end{aligned}$$

This average risk is in the interval B because for all subjects in the off-diagonal cell, risk(X, Y) $\in B$. Moreover, the average risk is not in the interval A because, being an off-diagonal cell, the interval A lies outside of interval B . It follows that for each

off-diagonal cell, in large samples the event rate is in the interval defined by the expanded model. That is, we expect that RC-correct = 100%. Any deviation from 100% that occurs must be due to sampling variability. Therefore, RC-correct cannot be used to compare the performances of the two risk models, and we see no purpose in calculating a statistic that is $\approx 100\%$ by definition.

The reclassification calibration statistics are:

$$RCC^X = \sum_{k=1}^K \frac{(\hat{p}_k - ave(\text{risk}(X))_k)^2}{ave(\text{risk}(X))_k(1 - ave(\text{risk}(X))_k)/n_k}$$

and

$$RCC^{(X,Y)} = \sum_{k=1}^K \frac{(\hat{p}_k - ave(\text{risk}(X,Y))_k)^2}{ave(\text{risk}(X,Y))_k(1 - ave(\text{risk}(X))_k)/n_k}$$

where the summation is over the K interior cells of the table with sample sizes $n_k \geq 20$, $ave(\text{risk}(\))_k$ is the average estimated risk for subjects in cell k and \hat{p}_k is the observed event rate in that cell. Cook and Ridker compare the reclassification calibration statistics to chi-squared distributions with $K - 2$ degrees of freedom for calculating p-values.

The arguments above show that in large samples the observed event rates in off-diagonal cells converge to $ave(\text{risk}(X,Y))_k$ but not to $ave(\text{risk}(X))_k$. Therefore the implicit null hypothesis for the statistic $RCC^{(X,Y)}$ is satisfied. That is, the expanded model will not be rejected at a rate above the nominal significance level in large samples. On the other hand the implicit null hypothesis for the baseline model will be rejected in large samples assuming that Y is a risk factor that moves even a small proportion of subjects to off-diagonal cells. In other words, if there are subjects in off-diagonal cells, there is no point in performing the $RCC(X,Y)$ statistical test since the result is predetermined in large samples. If there are no subjects in off-diagonal cells the setting is degenerate and there is obviously no point in performing the test either. We implemented the RCC tests on our illustrative data and in agreement with our arguments above, the baseline model was rejected, $p < .001$, while the expanded model was not, $p = 0.29$.

In conclusion, we regard the risk-reclassification table proposed by Cook and Ridker [5] as a useful descriptive device. However, the analysis strategy based on the table is not informative for comparing risk models.

The Net Reclassification Index

Pencina et al. [17, 18] introduced the Net Reclassification Index (NRI) as a measure to compare the prediction performance of nested risk models. The statistic is calculated as the sum of two components, NRI_D , calculated among events or cases, and $NRI_{\bar{D}}$, calculated among nonevents or controls. When risk categories exist,

Table 7 Example where $cNRI > 0$ but there is no performance improvement

		$r(X, Y)$			
		Low	Med	High	
$r(X)$	Events				
	Low	10	10	0	20
	Med	5	20	10	35
	High	5	5	35	45
		20	35	45	100
		$r(X, Y)$			
		Low	Med	High	
Non-Events	Low	500	100	0	600
	Med	100	200	0	300
	High	0	0	100	100
			600	300	100

the data are summarized in reclassification tables of the form shown in Table 4 and the corresponding NRI statistics are

$$cNRI_D = P[risk_c(X, Y) > risk_c(X) | D = 1] - P[risk_c(X, Y) < risk_c(X) | D = 1]$$

$$cNRI_{\bar{D}} = P[risk_c(X, Y) < risk_c(X) | D = 0] - P[risk_c(X, Y) > risk_c(X) | D = 0]$$

$$cNRI = cNRI_D + cNRI_{\bar{D}}$$

where $risk_c(X, Y)$ is the risk category in which the subject’s value for $risk(X, Y)$ falls, $risk_c(X)$ is that in which his value of $risk(X)$ falls, and $cNRI$ stands for categorical NRI. In words, $cNRI_D$ is the proportion of cases above the diagonal of the event reclassification table minus the proportion below the diagonal. The counterpart, $cNRI_{\bar{D}}$ is the proportion of controls below the diagonal minus the proportion of controls above the diagonal. Their sum, $cNRI$, takes values between 0 and 2. For our data with 3 risk-categories, $cNRI_D = 0.100$, $cNRI_{\bar{D}} = 0.073$ and $cNRI = 0.174$.

The $cNRI$ statistic provides a descriptive summary of the reclassification tables. However, it is not well-suited to the purpose of comparing the prediction performance of the models $risk(X)$ and $risk(X, Y)$ because, in general, it does not represent a comparison of the performance of $risk(X, Y)$ with that of $risk(X)$. To see this, consider that the performance of the model $risk(X)$ must be derived from the case and control distributions of $risk(X)$. These distributions are contained in the *vertical margins* of the event and nonevent reclassification tables, respectively. Similarly the performance of the model $risk(X, Y)$ must be derived from the *horizontal margins* of the reclassification tables. However, the $cNRI$ statistic is a function that depends on the interior cells of the tables, not just their margins. This is illustrated in Table 7 that shows an example where the margins of the reclassification tables are equal but $cNRI > 0$.

Table 8 Notation for entries in the risk reclassification tables with two risk categories

<i>Events</i>	$r(X, Y)$	
	Low	High
Low	a	b
High	c	d

<i>Non-Events</i>	$r(X, Y)$	
	Low	High
Low	e	f
High	g	h

That is, in Table 7, prediction performance for the model risk(X, Y) is the same as that for the model risk(X) since the vertical and horizontal margins are equal. However, the cNRI statistic = $(20 - 15)/100 = 0.05 > 0$, indicating that performance has improved.

Although the cNRI statistic was originally proposed for use with 3 or more risk categories, it is interesting to consider it in the simpler and more common setting where only two risk categories exist that are separated at a treatment risk threshold r_H . Using the notation in Table 8, it is easy to see that with two categories

$$\begin{aligned}
 \text{cNRI}_D &= b - c = b + d - (c + d) = \text{HR}_D^{(X, Y)}(r_H) - \text{HR}_D^X(r_H) \\
 \text{cNRI}_{\bar{D}} &= f - g = f + h - (g + h) = \text{HR}_{\bar{D}}^X(r_H) - \text{HR}_{\bar{D}}^{(X, Y)}(r_H) \\
 \text{cNRI} &= \{ \text{HR}_D^{(X, Y)}(r_H) - \text{HR}_D^X(r_H) \} - \{ \text{HR}_{\bar{D}}^{(X, Y)}(r_H) - \text{HR}_{\bar{D}}^X(r_H) \}.
 \end{aligned}$$

That is, cNRI_D is the increase in the proportion of cases classified as high risk and $\text{cNRI}_{\bar{D}}$ is the decrease in the proportion of controls classified as high risk. The simple summation that is cNRI however, does not weight the relative contributions appropriately unless $r_H = \rho$. To see this, recall that the change in the standardized net benefit does weight the contributions appropriately and is written as

$$s\text{NB}^{(X, Y)}(r_H) - s\text{NB}^X(r_H) = \{ \text{HR}_D^{(X, Y)}(r_H) - \text{HR}_D^X(r_H) \} - \frac{1 - \rho}{\rho} \frac{r_H}{(1 - r_H)} \{ \text{HR}_{\bar{D}}^{(X, Y)}(r_H) - \text{HR}_{\bar{D}}^X(r_H) \}.$$

Only when $r_H = \rho$ does cNRI correspond to the appropriately weighted combination, the change in $s\text{NB}(r_H)$. Interestingly a weighted version of the two-category NRI statistic has recently been proposed to correspond with the form of change in $s\text{NB}(r_H)$, by weighting $\text{cNRI}_{\bar{D}}$ by $\frac{(1 - \rho)}{\rho} \frac{r_H}{(1 - r_H)}$ [18]. However, weighted versions of the cNRI statistic have not been proposed to correspond with change in standardized net benefit when more than two categories are involved.

A continuous version of the NRI statistic has been proposed for use when no clinically relevant risk categories exist:

$$\begin{aligned}
 \text{NRI}_D &= P(\text{risk}(X, Y) > \text{risk}(X) | D = 1) - P(\text{risk}(X, Y) < \text{risk}(X) | D = 1) \\
 &= 2P(\text{risk}(X, Y) > \text{risk}(X) | D = 1) - 1
 \end{aligned}$$

$$\begin{aligned}
NRI_{\bar{D}} &= P(\text{risk}(X, Y) < \text{risk}(X) | D = 0) - P(\text{risk}(X, Y) > \text{risk}(X) | D = 0) \\
&= 1 - 2 P(\text{risk}(X, Y) > \text{risk}(X) | D = 0) \\
NRI &= 2\{P(\text{risk}(X, Y) > \text{risk}(X) | D = 1) - P(\text{risk}(X, Y) > \text{risk}(X) | D = 0)\}.
\end{aligned}$$

The statistic NRI is also denoted by $NRI(> 0)$. Again, since it is not a function of the marginal case and control risk distributions, NRI does not seem well-suited to quantifying the improvement in prediction performance of $\text{risk}(X, Y)$ versus $\text{risk}(X)$. It is not composed as a difference between a measure of the performance of $\text{risk}(X, Y)$ and a measure of the performance of $\text{risk}(X)$. However it is an interesting easily understood descriptive statistic about the joint distributions of $(\text{risk}(X), \text{risk}(X, Y))$ based on the comparison of $\text{risk}(X, Y)$ and $\text{risk}(X)$ within individuals. In our data we calculate that for 69.4% of cases their calculated risks increased with addition of Y and for 29.5% of controls their risks increased with addition of Y . Consequently $NRI_D = 0.388$, $NRI_{\bar{D}} = 0.411$ and $NRI = 0.799$.

Hypothesis Testing for Nested Models

When evaluating if a model that includes predictor Y in addition to X improves performance over the baseline model that includes X only, it is common practice to do several hypothesis tests. One will typically test the hypothesis $H_0^1 : \text{risk}(X, Y) = \text{risk}(X)$, using, for example, likelihood techniques based on regression models. If H_0^1 is rejected, one may test if the prediction performance of $\text{risk}(X, Y)$ is equal to that of $\text{risk}(X)$ using one or more statistics, such as the difference in the AUCs or the IDI statistic, which is the difference in MRDs, amongst others. The following result indicates that the null hypotheses concerning many measures of performance improvement are identical to $H_0^1 : \text{risk}(X, Y) = \text{risk}(X)$.

Result 2. The following conditions are equivalent

$$\begin{aligned}
H_0^1 &: \text{risk}(X, Y) = \text{risk}(X) \\
H_0^2 &: \text{ROC}^{(X, Y)}(f) = \text{ROC}^X(f) \quad \forall f \\
H_0^3 &: \text{AUC}^{(X, Y)} = \text{AUC}^X \\
H_0^4 &: \text{MRD}^{(X, Y)} = \text{MRD}^X \\
H_0^5 &: \text{AARD}^{(X, Y)} = \text{AARD}^X \\
H_0^6 &: NRI(> 0) = 0
\end{aligned}$$

□

For a proof of Result 2 and additional related results see Pepe [23].

The practical implication of this result is that if H_0^1 is rejected, one can conclude that the other hypotheses listed in Result 2 are also rejected. Testing any one of hypotheses $H_0^2 - H_0^6$ is equivalent to testing H_0^1 . We recommend using standard well-studied methods from regression modeling to test the hypothesis formulated as H_0^1 . Corresponding statistical techniques are well-developed and they are often efficient. In contrast techniques based on estimates of the performance measures in $H_0^2 - H_0^6$ are likely to be less efficient and have in some cases been shown to have bizarre distributions under the null hypothesis of no change in performance [14]. This is an active area of research.

Concluding Remarks

We now recap some of the main points made in this chapter. First, a necessary condition for a useful risk prediction model is that it be well-calibrated. Whereas the scale of a marker used for classifying individuals according to disease status is not in and of itself of interest, risks calculated using a prediction model are used to advise patients and to make medical decisions. The scale of the risk model predictions is therefore a fundamental aspect of the model's utility. Good calibration is essential.

The distributions of risk predicted by the model, for cases and for controls, are the building blocks for evaluating model performance. Summary measures are functions of these distributions. A variety of summary measures have been described here which rely on specification of a high risk threshold (or multiple risk thresholds) for classifying subjects. Our preferred measures are the proportions of cases and controls classified as high risk (or classified into each risk category) and the standardized net benefit of using the model. Performance measures that do not rely on risk thresholds can be useful for screening many models in order to select a subset for further evaluation.

The choice of the risk threshold(s) should be based on an assessment of costs and benefits associated with a high risk (or each risk category) designation. These costs and benefits are also used in calculating the scaled net benefit of the model.

When comparing models, our recommendation is to compare measures of marginal performance. This is in contrast to basing comparisons on statistics that summarize the cross-classification of the two models. Assessing cross-classification is useful for descriptive analysis but cannot be used as the basis for forming conclusions regarding the relative performance of the two models.

When testing the incremental value of a new predictor added to a risk model, standard likelihood methods should be used. Tests based on contrasts of performance measures between the baseline and expanded risk models are testing the same null hypothesis. These tests have, in some cases, been shown to poorly control the type-I error rate and to be less powerful than likelihood-based inference [14, 23, 31].

This chapter has focused on conceptual approaches to evaluation assuming a very large sample size. In practice, where sample sizes are finite, all measures of model performance should be accompanied by confidence intervals to characterize the level of uncertainty. This practice is much more informative than reporting p-values based on hypothesis tests; moreover in some instances as mentioned above, tests based on model performance measures have poor properties. Bootstrapping is a simple and flexible technique that can be used to construct confidence intervals. The bootstrapping should reflect the actual data analysis that was performed; if the risk model was fit using the data (versus using a separate dataset), the model should be re-fit and performance estimated in each bootstrap sample. Bootstrapping is flexible in that unique attributes of the original study design can be accommodated, such as repeated measures or case-control sampling.

When a risk model is fit and evaluated using the same data, the apparent performance will tend to be over-optimistic. This is particularly true if an intensive model-selection procedure was employed. Standard approaches to dealing with this problem include separating the data into “training” and “test” portions, and more efficient methods such as cross-validation [9, 10]. The disadvantage of the latter approach is the requirement for a prespecified and automated model selection procedure. If either approach is used, it should be reflected in the bootstrapping described above. Specifically, in each bootstrap sample the complete model selection procedure should be performed.

In many contexts there are covariates that need to be taken into account when predicting risk and evaluating risk model performance. We distinguish between covariates (Z) that predict risk of bad outcome, e.g. age, and covariates that modify the performance of a risk calculator ($\text{risk}(X)$), e.g., the laboratory in which a biomarker X is assayed. Of course some covariates, such as disease comorbidities, may have both types of effects. For covariates that predict risk only, the approach is to simply include these as predictors in the risk model, i.e. to model $P(D = 1|X, Z) = \text{risk}(X, Z)$. In this way, the covariates Z are treated as additional predictors and the methods described in this chapter apply directly. On the other hand, covariates that modify the distribution of $\text{risk}(X)$ will have an impact on the performance of the model. Evaluating how the performance of $\text{risk}(X)$ varies with Z will generally require modeling X as a function of Z and we refer the reader to Huang [12] for details. Covariates that both predict risk and modify performance can be accommodated using the same methods where $\text{risk}(X, Z)$ is the risk calculator and the joint distribution of (X, Z) is estimated as a function of Z .

Our discussion of the choice of risk threshold(s) assumed implicitly that the cost and benefit of being treated are constant across individuals, and in particular are independent of the predictor X . Predictors (X) that predict the benefit or cost of treatment have greater potential net benefit [13, 25, 30]. However evaluating whether this is so requires data from a randomized trial where the predictor X is measured at baseline, in order to assess the cost and benefit of treatment as a function of X .

References

1. Baker, S.: Putting risk prediction in perspective: relative utility curves. *J. Natl. Cancer Inst.* **101**(22), 1538–1542 (2009)
2. Baker, S., Kramer, B.: Evaluating a new marker for risk prediction: decision analysis to the rescue. *Discov. Med.* **14**(76), 181–188 (2012)
3. Bura, E., Gastwirth, J.: The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biom. J.* **43**(1), 5–21 (2001)
4. Cook, N.: Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**(7), 928–935 (2007)
5. Cook, N., Ridker, P.: Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann. Intern. Med.* **150**(11), 795–802 (2009)
6. Gail, M., Costantino, J.: Validating and improving models for projecting the absolute risk of breast cancer. *J. Natl. Cancer Inst.* **93**(5), 334–335 (2001)
7. Gail, M., Pfeiffer, R.: On criteria for evaluating models of absolute risk. *Biostatistics* **6**(2), 227–239 (2005)
8. Gu, W., Pepe, M.: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **5**(1), Article 27 (2009). doi:10.2202/1557-4679.1188
9. Harrell, F.: *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York (2001)
10. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer, New York (2001)
11. Huang, Y., Pepe, M.: A parametric ROC model-based approach for evaluating the predictive-ness of continuous markers in case-control studies. *Biometrics* **65**(4), 1133–1144 (2009)
12. Huang, Y., Pepe, M.S.: Semiparametric methods for evaluating the covariate-specific predictive-ness of continuous markers in matched case-control studies. *J. R. Stat. Soc., Ser. C (Appl. Stat.)* **59**(3), 437–456 (2010)
13. Janes, H., Pepe, M.S., Bossuyt, P.M., Barlow, W.E.: Measuring the performance of markers for guiding treatment decisions. *Ann. Intern. Med.* **154**, 253–259 (2011)
14. Kerr, K.F., McClelland, R.L., Brown, E.R., Lumley, T.: Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am. J. Epidemiol.* **174**(3), 364–374 (2011)
15. Lemeshow, S., Hosmer Jr, D.: A review of goodness of fit statistics for use in the development of logistic regression models. *Am. J. Epidemiol.* **115**(1), 92–106 (1982)
16. Parikh, C.R., Devarajan, P., Zappitelli, M., Sint, K., Thiessen-Philbrook, H., Li, S., Kim, R.W., Koyner, J.L., Coca, S.G., Edelstein, C.L., Shlipak, M.G., Garg, A.X., Krawczeski, C.D., TRIBE-AKI Consortium: Postoperative biomarkers predict acute kidney injury and poor outcomes after pediatric cardiac surgery. *J. Am. Soc. Nephrol.* **22**(9), 1737–1747 (2011)
17. Pencina, M., D’Agostino, R., D’Agostino, R., Vasan, R.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**(2), 157–172 (2008)
18. Pencina, M., D’Agostino Sr, R., Steyerberg, E.: Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **30**(1), 11–21 (2011)
19. Pepe, M.: Problems with risk reclassification methods for evaluating prediction models. *Am. J. Epidemiol.* **173**(11), 1327 (2011)
20. Pepe, M., Janes, H.: Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J. Natl. Cancer Inst.* **100**(14), 978–979 (2008)
21. Pepe, M., Feng, Z., Gu, J.: Comments on ‘Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond’ by MJ Pencina et al. *Stat. Med.* **27**(2), 173–181 (2008). doi:10.1002/sim.2929
22. Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I., Zheng, Y.: Integrating the predictiveness of a marker with its performance as a classifier. *Am. J. Epidemiol.* **167**(3), 362 (2008)

23. Pepe, M., Kerr, K., Longton, G., Wang, Z.: Testing for improvement in prediction model performance. *Stat. Med.* **32**(9), 1467–1482 (2013)
24. Pfeiffer, R., Gail, M.: Two criteria for evaluating risk prediction models. *Biometrics* **67**(3), 1057–1065 (2011)
25. Sargent, D.J., Conley, B.A., Allegra, C., Collette, L.: Clinical trial designs for predictive marker validation in cancer treatment trials. *J. Clin. Oncol.* **23**(9), 2020–2027 (2005)
26. Seymour, C.W., Kahn, J.M., Cooke, C.R., Watkins, T.R., Heckbert, S.R., Rea, T.D.: Prediction of critical illness during out-of-hospital emergency care. *JAMA* **304**(7), 747–754 (2010)
27. Steyerberg, E.: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, New York (2009)
28. Steyerberg, E., Borsboom, G., van Houwelingen, H., Eijkemans, M., Habbema, J.: Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat. Med.* **23**(16), 2567–2586 (2004)
29. Vickers, A., Elkin, E.: Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**(6), 565 (2006)
30. Vickers, A.J., Kattan, M.W., Daniel, S.: Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* **5**, 8–14 (2007)
31. Vickers, A.J., Cronin, A.M., Begg, C.B.: One statistical test is sufficient for assessing new predictive markers. *BMC Med. Res. Methodol.* **11**, 13 (2011)
32. Wilson, P., D’Agostino, R., Levy, D., Belanger, A., Silbershatz, H., Kannel, W.: Prediction of coronary heart disease using risk factor categories. *Circulation* **97**(18), 1837–1847 (1998)

Estimating Improvement in Prediction with Matched Case-Control Designs

Aasthaa Bansal and Margaret Sullivan Pepe

Abstract When an existing risk prediction model is not sufficiently predictive, additional variables are sought for inclusion in the model. This paper addresses study designs to evaluate the improvement in prediction performance that is gained by adding a new predictor to a risk prediction model. We consider studies that measure the new predictor in a case-control subset of the study cohort, a practice that is common in biomarker research. We ask if matching controls to cases in regards to baseline predictors improves efficiency. A variety of measures of prediction performance are studied. We find through simulation studies that matching improves the efficiency with which most measures are estimated, but can reduce efficiency for some. Efficiency gains are less when more controls per case are included in the study. A method that models the distribution of the new predictor in controls appears to improve estimation efficiency considerably.

Introduction

Medical decisions are often based on an individual's calculated risk of having or developing a condition. For example, decisions to prescribe long-term cholesterol lowering statin therapy are often made with use of the Framingham risk of a cardiovascular event [1, 12, 21, 34] that uses as input information the individual's

This paper appeared in volume 19 (2013) of Lifetime Data Analysis.

A. Bansal (✉)

University of Washington, 15T UW Tower, 4333 Brooklyn Ave NE,
Box 359461, Seattle, WA 98195, USA

e-mail: abansal@u.washington.edu

M.S. Pepe

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North,
M2-B500, Seattle, WA 98109, USA

e-mail: mspepe@u.washington.edu

sex, age, blood pressure, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, smoking behavior and diabetes status. The Breast Cancer Risk Assessment Tool (BCRAT) is used to calculate 10 year risk of breast cancer for individuals, using information on age, personal medical history (number of previous breast biopsies and the presence of atypical hyperplasia in any previous breast biopsy specimen), reproductive history (age at the start of menstruation and age at the first live birth of a child) and family history of breast cancer. If a woman's risk exceeds an age-specific threshold, she may be recommended for hormone therapy that reduces the risk at least in some women. Risk prediction models can also be used to determine if a person's risk is low enough to forgo certain unpleasant or costly medical interventions [10, 11].

Our ability to predict risk with currently available clinical predictors is often very poor. For example the BCRAT model has a very modest capacity to discriminate women who develop breast cancer within 10 years from those who do not. The area under the age-specific receiver operating characteristic curve is approximately 0.56 [24]. Therefore new predictors are sought for their capacity to improve upon its prediction performance. Recent advances in and wider availability of molecular and imaging biotechnologies offer the potential for new powerful predictors. Recent studies have examined the use of data on genetic polymorphisms and breast density to improve the performance of BCRAT.

This paper concerns study designs to estimate the improvement in prediction performance that is gained by adding a new predictor Y to a set of baseline predictors X , to predict the risk of an outcome D ($D = 1$ for a bad outcome and $D = 0$ for a good outcome). When resources are limited and Y is difficult to ascertain, it may not be feasible to measure it on all subjects in a study cohort. Consider, for example, if the new predictor is a biomarker measured on biological samples obtained and stored while women were healthy at enrollment in the Women's Health Initiative. The preciousness of such biological samples dictates that they be used with maximum efficiency. Typically therefore a case-control study design is employed wherein Y is measured on a random subset of cases (denoted by $D = 1$) and a selected subset of controls ($D = 0$).

Our specific interest concerns whether or not the controls on whom Y is measured should be selected to frequency match the cases with regard to the baseline predictors X . Matching is in fact routinely done in practice in order to avoid observing associations between Y and D that are solely due to associations of X with both Y and D . However, the effect of this practice on estimation of performance improvement is not fully understood. We have raised concerns about matching with regards to bias, emphasizing that naïve analyses typically employed are misleading, as they underestimate performance [30]. The effect of matching on the estimation of incremental value with regards to efficiency has not been examined. Nevertheless, the practice is entrenched in the field of biomarker research. Here, we propose a two-stage estimator that accounts for matching to produce unbiased estimates. Using this estimator, we look to address the question of whether matching can improve the efficiency of estimating the increment in performance. This is an important question given that matching also necessitates a somewhat more complicated

analysis algorithm than is required for an unmatched study. We ask whether there is a large enough (or any) efficiency gain that justifies the common practice of matching and a more complicated analysis.

Matching is known to improve efficiency for estimating the odds ratio for Y in a risk model that includes X [4]. However, the odds ratio, $\frac{P(D=1|X,Y=y+1)/P(D=0|X,Y=y+1)}{P(D=1|X,Y=y)/P(D=0|X,Y=y)}$, does not characterize prediction performance or improvement in prediction performance gained by including Y in the risk model over and above use of X alone. The distribution of (X, Y) in the population is an additional component that enters into the calculation of prediction performance. Janes and Pepe [19] showed that matching on X is also optimal for estimating the covariate adjusted ROC curve, which is a measure of prediction performance. However, Janes and Pepe [18] show that the covariate adjusted ROC curve that characterizes the ROC performance of Y within populations where X is fixed, does not quantify the improvement in the ROC curve gained by including Y in the risk model. It is currently unknown if matching leads to gains in efficiency for estimating performance improvement.

There are many metrics available for gauging improvement in prediction performance, and there is much confusion in the field about which metrics are most worthy for reporting. In section “Measures of Improvement in Prediction Performance”, we review the most popular measures, providing some novel insights about their interpretations and inter-relationships. We provide rationale for the measures we selected to study here. In section “Estimation from Matched and Unmatched Designs”, we describe how these measures can be estimated from matched and unmatched studies. Simulation studies that were performed to evaluate the properties of the estimators and the efficiencies of matched designs are described in section “Simulation Studies” using a simulated dataset and a real dataset concerning the prediction of renal artery stenosis. In section “Bootstrap Method for Inference”, we propose a bootstrap approach for inference and demonstrate its validity through simulation studies. In section “Illustration with Renal Artery Stenosis Study”, we illustrate our methodology in the context of renal artery stenosis. We close with some recommendations and suggestions for further research.

Measures of Improvement in Prediction Performance

We first consider the most popular measures used to quantify improvement in prediction performance. Table 1 presents definitions for these measures. In this section, we review the measures in more detail.

Notation

Recall our use of D for the outcome variable, $D = 1$ denoting a case with a bad outcome and $D = 0$ denoting a control with a good outcome. We use X for predictors in the baseline risk function, $\text{risk}(X) = P(D = 1|X)$, Y for the novel

Table 1 Definitions of performance measures. The subscript X or (X, Y) denotes if the measure is calculated with the baseline or expanded risk models

Name	Definition and notation	Performance improvement measure
High risk cases (r)	$HR^D(r) = P(\text{risk} > r D = 1)$	$\Delta HR^D(r) = HR^D_{(X,Y)}(r) - HR^D_X(r)$
High risk controls (r)	$HR^D(r) = P(\text{risk} > r D = 0)$	$\Delta HR^D(r) = HR^D_{(X,Y)}(r) - HR^D_X(r)$
Standardized benefit (r)	$B(r) = HR^D(r) - \frac{(1-p)r}{p} HR^D(r)$	$\Delta B(r) = B_{(X,Y)}(r) - B_X(r)$
Cases above control defined threshold (p^D)	$ROC(p^D) = P(\text{risk} > r(p^D) D = 1)$	$\Delta ROC(p^D) = ROC_{(X,Y)}(p^D) - ROC_X(p^D)$
Controls above case defined threshold (p^D)	$ROC^{-1}(p^D) = P(\text{risk} > r(p^D) D = 0)$	$\Delta ROC^{-1}(p^D) = ROC^{-1}_{(X,Y)}(p^D) - ROC^{-1}_X(p^D)$
Area under the ROC curve	$AUC = P(\text{risk}_i > \text{risk}_j D_i = 1, D_j = 0)$	$\Delta AUC = AUC_{(X,Y)} - AUC_X$
Mean risk difference	$MRD = E(\text{risk} D = 1) - E(\text{risk} D = 0)$	$\Delta MRD = MRD_{(X,Y)} - MRD_X = \text{IDI}$
Above average risk difference	$AARD = \{P(\text{risk} > p D = 1) - P(\text{risk} > p D = 0)\}$	$\Delta AARD = AARD_{(X,Y)} - AARD_X$
Net reclassification improvement	$NRI(> 0)$	$NRI = 2\{P(\text{risk}(X, Y) > \text{risk}(X) D = 1) - P(\text{risk}(X, Y) > \text{risk}(X) D = 0)\}$

predictors to be added and we write $\text{risk}(X, Y) = P(D = 1 | X, Y)$. All measures of prediction performance involve the distributions of $\text{risk}(X)$ and $\text{risk}(X, Y)$ in cases and controls. We write these distributions as:

$$\begin{aligned} F_X^D(r) &= P(\text{risk}(X) \leq r | D = 1) \\ F_X^{\bar{D}}(r) &= P(\text{risk}(X) \leq r | D = 0) \\ F_{X,Y}^D(r) &= P(\text{risk}(X, Y) \leq r | D = 1) \\ F_{X,Y}^{\bar{D}}(r) &= P(\text{risk}(X, Y) \leq r | D = 0) \end{aligned}$$

The joint distributions of $(\text{risk}(X), \text{risk}(X, Y))$ in cases and controls will be denoted by $F^D(r, r')$ and $F^{\bar{D}}(r, r')$ respectively.

Proportions at High Risk and Net Benefit

In some settings a threshold exists for high risk classification and patients designated as ‘high risk’ receive an intervention. For example, patients whose 10-year risk of a cardiovascular event exceeds 20% are recommended for cholesterol lowering therapy [8]. A risk model performs well, in the sense of treating people who would have an event in the absence of therapy, i.e. the cases, if a large proportion of those subjects are placed in the high risk category by the model, i.e. if $\text{HR}^D(r) \equiv P[\text{risk} > r | D = 1]$ is large. Conversely, one must consider to what extent subjects that would not have an event in the absence of intervention, i.e. the controls, are inappropriately given intervention. A good model will place few of the controls in the high risk category, i.e. $\text{HR}^{\bar{D}}(r) \equiv P[\text{risk} > r | D = 0]$ is small. The changes in $\text{HR}^D(r)$ and $\text{HR}^{\bar{D}}(r)$ that are gained by adding Y to the risk model are therefore key entities for quantifying improvement in model performance for decision making when a therapeutic threshold for risk exists:

$$\begin{aligned} \Delta \text{HR}^D(r) &\equiv P[\text{risk}(X, Y) > r | D = 1] - P[\text{risk}(X) > r | D = 1] \\ \Delta \text{HR}^{\bar{D}}(r) &\equiv P[\text{risk}(X) > r | D = 0] - P[\text{risk}(X, Y) > r | D = 0]. \end{aligned}$$

These measures are also called changes in the true and false positive rates. Note that our goal is to increase $\text{HR}^D(r)$ and reduce $\text{HR}^{\bar{D}}(r)$ by adding Y to the baseline risk model. Therefore positive values of ΔHR^D and $\Delta \text{HR}^{\bar{D}}$ are desirable.

There is a net expected benefit (B) associated with designating a case as high risk and a net expected cost (C) associated with designating a control as high risk. It has been noted that a rational choice of risk threshold is $r = C/(C + B)$ [25, 35] and that the expected population net benefit associated with use of a risk model and threshold r to assign treatment is $\text{NB}(r) = \{\rho \text{HR}^D(r) - (1 - \rho) \frac{r}{(1-r)} \text{HR}^{\bar{D}}(r)\} B$ where ρ is the population prevalence, $P[D = 1]$. Baker [5] suggests standardizing

$NB(r)$ by the maximum possible benefit, ρB , achieved when all cases and no controls are designated as high risk. This standardized measure $B(r) \equiv HR^D(r) - \frac{(1-\rho)}{\rho} \frac{r}{(1-r)} HR^{\bar{D}}(r)$, the proportion of maximum benefit, can also be viewed as the true positive rate $HR^D(r)$ discounted (appropriately) for the false positive rate $HR^{\bar{D}}(r)$. The change in $B(r)$ that is achieved by adding Y to the risk model is an appropriate summary of its components $\Delta HR^D(r)$ and $\Delta HR^{\bar{D}}(r)$:

$$\Delta B(r) = \Delta HR^D(r) + \frac{1-\rho}{\rho} \frac{r}{1-r} \Delta HR^{\bar{D}}(r).$$

In some settings all subjects receive treatment by default and use of a prediction model is to identify low risk subjects that can forego treatment. Parameters analogous to $\Delta HR^D(r)$, $\Delta HR^{\bar{D}}(r)$ and $\Delta B(r)$ can be defined but we do not focus on those here.

Performance Measures Related to Fixed Points on the ROC Curve

When risk thresholds or costs and benefits are not available, other approaches to summarizing prediction performance have been proposed. Points on the ROC curve or on its inverse are commonly used in practice because of their use in evaluating diagnostic tests and classifiers. We define

$$\Delta ROC(p^{\bar{D}}) = ROC_{(X,Y)}(p^{\bar{D}}) - ROC_X(p^{\bar{D}})$$

where $ROC(p^{\bar{D}})$ is the proportion of cases with risks above the threshold $r(p^{\bar{D}})$ that allows the fraction $p^{\bar{D}}$ of controls to be classified as high risk. Analogously,

$$\Delta ROC^{-1}(p^D) = ROC_X^{-1}(p^D) - ROC_{(X,Y)}^{-1}(p^D)$$

where $ROC^{-1}(p^D)$ is the proportion of controls with risks above the threshold $r(p^D)$ that is exceeded by the fraction p^D of cases.

Interestingly, the ROC points are closely related to measures proposed by Pfeiffer and Gail [32] for quantifying prediction performance. They argue for choosing a high risk threshold $r(p^D)$ so that a specified proportion of cases (p^D) are designated as high risk and define the proportion needed to follow, $PNF(p^D) = P[\text{risk} > r(p^D)]$, as a performance metric. In words, $PNF(p^D)$ is the proportion of the population designated as high risk in order that p^D of the cases are classified as high risk. A little algebra shows that $PNF(p^D) = \rho p^D + (1-\rho)ROC^{-1}(p^D)$. The reduction in the proportion of the population needed to follow in order to identify p^D of the cases (ΔPNF) that is gained by adding Y to the model is

$$\Delta PNF(p^D) = (1-\rho)\Delta ROC^{-1}(p^D).$$

We choose to study $\Delta\text{ROC}^{-1}(p^D)$ here as it does not depend on the prevalence. Pfeiffer and Gail [32] also define a performance metric that is the proportion of cases followed, $\text{PCF}(p)$, when a fixed proportion p of the population is designated as highest risk. This measure relates directly to the ROC:

$$\text{PCF}(p) = \text{ROC}(p^{\bar{D}})$$

where $p^{\bar{D}}$ is the point on the x-axis of the ROC plot such that $p = \rho\text{ROC}(p^{\bar{D}}) + (1 - \rho)p^{\bar{D}}$. We study $\Delta\text{ROC}(p^{\bar{D}})$ rather than $\Delta\text{PCF}(p)$ here because of its widespread use and its independence from the prevalence.

Global Performance Measures that Do Not Specify a Risk Threshold

The above measures require explicit or implicit choices for risk thresholds. Measures that average over all risk thresholds in some sense are popular in part because they avoid the need to choose a risk threshold. The change in the area under the ROC curve by adding Y to the model, denoted ΔAUC , is the most commonly used measure in practice. The AUC is often written as

$$\text{AUC} = P(\text{risk}_i > \text{risk}_j | D_i = 1, D_j = 0)$$

and

$$\Delta\text{AUC} = \text{AUC}_{(X,Y)} - \text{AUC}_X.$$

A more recently proposed measure, called the integrated discrimination improvement (IDI) index, is the change in the difference in mean risks between cases and controls:

$$\text{IDI} = \Delta\text{MRD} = \text{MRD}_{(X,Y)} - \text{MRD}_X$$

where

$$\text{MRD} = E(\text{risk} | D = 1) - E(\text{risk} | D = 0).$$

Both the AUC and the MRD are measures of distance between the case and control distributions of modeled risks. Another measure of distance between distributions is the above average risk difference:

$$\text{AARD} = P(\text{risk} > \rho | D = 1) - P(\text{risk} > \rho | D = 0),$$

the name deriving from the fact that $E(\text{risk}) = \rho$ regardless of the risk model. We study the AARD because it is related to several other measures of prediction

performance. We note in particular that $\text{AARD} = \text{B}(\rho)$. Youden's index is a measure of diagnostic performance for binary tests and we write $\text{YI}(r) = \text{HR}^D(r) - \text{HR}^{\bar{D}}(r)$. We note that $\text{AARD} = \text{YI}(\rho)$. Moreover, theory from Gu and Pepe [13] implies that $\text{YI}(\rho) = \max(\text{ROC}(\rho) - \rho) = \max(\text{YI}(r))$. Therefore, $\text{AARD} = \max(\text{YI}(r))$. This is also known as the Kolmogorov-Smirnov measure of distance between the case and control risk distributions. Finally, Gu and Pepe [13] also showed that this statistic is equal to the standardized total gain statistic [6], a measure derived from the population distribution of risk. The measure of improvement in prediction performance that we consider is the difference in measures calculated with $\text{risk}(X, Y)$ compared with when calculated with $\text{risk}(X)$:

$$\Delta \text{AARD} = \text{AARD}_{(X, Y)} - \text{AARD}_X.$$

Risk Reclassification Performance Measures

Reclassification measures of performance compare $\text{risk}(X, Y)$ with $\text{risk}(X)$ within individuals and summarize across subjects. The most popular measure is the net reclassification improvement (NRI) index [26]. We focus on the continuous NRI [27], written $\text{NRI}(> 0)$:

$$\begin{aligned} \text{NRI}(> 0) &\equiv P(\text{risk}(X, Y) > \text{risk}(X) | D = 1) - P(\text{risk}(X, Y) < \text{risk}(X) | D = 1) \\ &\quad + P(\text{risk}(X, Y) < \text{risk}(X) | D = 0) - P(\text{risk}(X, Y) > \text{risk}(X) | D = 0) \\ &= 2\{P(\text{risk}(X, Y) > \text{risk}(X) | D = 1) - P(\text{risk}(X, Y) > \text{risk}(X) | D = 0)\} \end{aligned}$$

It is interesting to consider the $\text{NRI}(> 0)$ statistic when the baseline model contains no covariates, i.e. when all subjects are assigned $\text{risk} = \rho$. In this setting it is related to measures mentioned previously:

$$\text{NRI} = 2\{\text{HR}^D(\rho) - \text{HR}^{\bar{D}}(\rho)\} = 2\text{AARD}(\rho) = 2\text{YI}(\rho) = 2\text{B}(\rho).$$

Originally the NRI was proposed for categories of risk and was defined as the net proportion of cases that moved to a higher risk category plus the net proportion of controls that moved to a lower risk category. When there are two categories, above or below the risk threshold r , the $\text{NRI} = \Delta \text{HR}^D(r) + \Delta \text{HR}^{\bar{D}}(r) = \Delta \text{YI}(r)$. Similar to $\Delta \text{B}(r)$, it is a weighted summary of improvements in true and false positive rates but unfortunately it uses inappropriate weights.

Another risk reclassification measure is the integrated discrimination improvement (IDI), also defined as:

$$\text{IDI} = E\{\text{risk}(X, Y) - \text{risk}(X) | D = 1\} + E\{\text{risk}(X) - \text{risk}(X, Y) | D = 0\}.$$

Interestingly, because of the linearity, this measure of individual changes in risk due to adding Y to the model can also be interpreted as a difference of two population performance measures. That is, as noted earlier

$$\Delta\text{MRD} = \text{MRD}_{(X,Y)} - \text{MRD}_X = \text{IDI}.$$

Estimation from Matched and Unmatched Designs

We now consider how the measures defined above can be estimated from a cohort study within which a case-control study of a new predictor is nested.

Data

We assume that data on the outcome and baseline covariates are available on a simple random sample of N independent identically distributed observations: $(D_k, X_k), k = 1 \dots, N$. We select a simple random sample of n_D cases from the cohort to ascertain $Y: Y_i, i = 1, \dots, n_D$. The controls on whom Y is ascertained $\{Y_j, j = 1, \dots, n_D\}$ may be obtained as a simple random sample in an unmatched design. Alternatively, in a matched design, a categorical variable W is defined as a function of $X, W = W(X)$, and the number of controls within each level of W is chosen to equal a constant K times the number of cases with that value for W .

As shown in Table 1, all performance improvement measures are defined as functions of the risk distributions (notation in section “Notation”). We estimate $\text{risk}(X)$ and $\text{risk}(X, Y)$ first, then estimate their distributions in cases and controls and substitute the estimated distributions into expressions for the performance improvement measures.

Estimating Risk Functions

For the baseline model, we fit a regression model to the cohort data $\{(D_k, X_k), k = 1, \dots, N\}$ and calculate predicted risks, $\widehat{\text{risk}}(X)$, for each individual in the cohort. For the expanded model, $\text{risk}(X, Y)$, we consider two approaches.

Case-control with adjustment We fit a model to data from the case-control subset, yielding fitted values $\widehat{\text{risk}}^{cc}(X, Y)$, and then adjust the intercept to the prevalence in the cohort

$$\text{logit } \widehat{\text{risk}}^{adj}(X, Y) = \text{logit } \widehat{\text{risk}}^{cc}(X, Y) - \text{logit} \left(\frac{n_D}{n} \right) + \text{logit} \left(\frac{N_D}{N} \right),$$

where $n = n_D + n_{\bar{D}}$ and N_D is the number of cases in the cohort. This is a well-known and standard approach to estimation of absolute risk for epidemiologic case-control studies [2]. It draws upon the results of Prentice and Pyke [33], which suggested that a prospective logistic model can be fit to retrospective data from a case-control study with a slight modification that adds an offset term to the logistic model. The approach maximizes the pseudo- (or conditional-) likelihood that an observation in the case-control sample is a case or a control [3, 9].

However this approach does not account for matching. Pencina et al. [27] presented a similar approach that used intercept adjustment to estimate $\text{NRI}(> 0)$ in the context of simple case-control studies.

Two-Stage Two-stage methods acknowledge that selection of subjects for whom Y is measured, i.e. the second stage of sampling, may depend on their values of (D, X) found in the first stage. In particular, they account for matching. We generalize the intercept adjustment idea presented above to account for matching on X . This requires using the cohort to adjust the odds ratio associated with X . The odds ratio associated with Y is correctly estimated using standard logistic regression applied to the case-control dataset. We use the corresponding fitted values but adjust them using fitted values from the baseline model fit to the cohort and to the case-control datasets. Specifically, if we let $\widehat{\text{risk}}^{\text{cohort}}(X)$ and $\widehat{\text{risk}}^{\text{cc}}(X)$ denote the fitted values for the baseline models, then the two-stage estimator of the absolute risk is:

$$\widehat{\text{logit risk}}^{2\text{-stage}}(X, Y) = \widehat{\text{logit risk}}^{\text{cc}}(X, Y) - \widehat{\text{logit risk}}^{\text{cc}}(X) + \widehat{\text{logit risk}}^{\text{cohort}}(X)$$

Using ‘cohort’ and ‘cc’ to denote sampling in the cohort or in the case-control subset, rationale for $\widehat{\text{risk}}^{2\text{-stage}}(X, Y)$ derives from the facts that

$$\text{logit } P(D = 1 | X, Y, \text{cohort}) = \text{logit } P(D = 1 | X, \text{cohort}) + \log \text{DLR}_X(Y)$$

and

$$\text{logit } P(D = 1 | X, Y, \text{cc}) = \text{logit } P(D = 1 | X, \text{cc}) + \log \text{DLR}_X(Y)$$

where the covariate-specific diagnostic likelihood ratio

$$\text{DLR}_X(Y) = P(Y | X, D = 1) / P(Y | X, D = 0)$$

is the same in the (matched or unmatched) case-control and cohort populations. The equations are a simple application of Bayes’ theorem [14]. Substituting the expression for $\log \text{DLR}_X(Y)$ derived from the case-control equation into that for the cohort equation gives the expression above for $\widehat{\text{logit risk}}^{2\text{-stage}}(X, Y)$.

Estimating Distributions of Risk

To estimate the risk distributions, we draw upon previously proposed methods for the estimation of risk distributions in simple case-control studies [14, 16, 17]. Here, we propose methodology for estimation with matched nested case-control data, which has not been previously considered. We estimate the baseline risk distributions, F_X^D and $F_X^{\bar{D}}$, using the empirical distributions of $\widehat{\text{risk}}(X)$ in the cohort data. Since the cases in the case-control set are drawn as a simple random sample from the cases in the cohort, we use the empirical distribution of $\widehat{\text{risk}}(X, Y)$ in the cases as the estimator of $F_{X,Y}^D$. For estimation of the distribution of $\widehat{\text{risk}}(X, Y)$ in the controls, we propose nonparametric and semiparametric approaches.

Nonparametric Estimation In unmatched case-controls studies we can also use the empirical distribution of $\widehat{\text{risk}}(X, Y)$ among the controls to estimate $F_{X,Y}^{\bar{D}}$. However in matched designs the controls are not a simple random sample and the distribution of $\widehat{\text{risk}}(X, Y)$ must be reweighted to reflect the distribution in the population. Specifically, letting $c = 1, \dots, C$ represent the distinct levels of the matching variable we can write

$$\begin{aligned}
 F_{X,Y}^D(r) &= P\{\text{risk}(X, Y) \leq r | D = 0\} \\
 &= \sum_{c=1}^C P\{\text{risk}(X, Y) \leq r | D = 0, W = c\} P(W = c | D = 0). \tag{1}
 \end{aligned}$$

A nonparametric estimator substitutes the observed proportions in the cohort for $P(W = c | D = 0)$ and the observed empirical stratum specific distributions of $\widehat{\text{risk}}(X, Y)$ for $P\{\text{risk}(X, Y) | D = 0, W = c\}$. We also consider a semiparametric estimator that substitutes semiparametric stratum specific estimates for $P\{\text{risk}(X, Y) \leq r | D = 0, W = c\}$.

Semiparametric Estimation Observe that

$$P\{\text{risk}(X, Y) \leq r | D = 0, W = c\} = E\{P(\text{risk}(X, Y) \leq r | D = 0, X) | D = 0, W = c\}. \tag{2}$$

A semiparametric location-scale model for the distribution of Y conditional on $(D = 0, X)$ is written

$$Y = \mu^{\bar{D}}(X) + \sigma^{\bar{D}}(X)\varepsilon$$

where the distribution of ε is unspecified, $\varepsilon \sim F_0$, and $\mu^{\bar{D}}(X)$, and $\sigma^{\bar{D}}(X)$ are parametric functions of X [15]. After fitting the regression functions $\hat{\mu}^{\bar{D}}(X)$ and $\hat{\sigma}^{\bar{D}}(X)$, the empirical distribution of the residuals $\hat{\varepsilon}_j = (Y_j - \hat{\mu}^{\bar{D}}(X_j)) / \hat{\sigma}^{\bar{D}}(X_j)$,

$j = 1, \dots, n_D$, yields an estimator \hat{F}_0 . The semiparametric estimate of the distribution of Y is then

$$\begin{aligned} \hat{P}(Y \leq y | D = 0, X) &= \hat{P} \left\{ \frac{Y - \hat{\mu}^{\bar{D}}(X)}{\hat{\sigma}^{\bar{D}}(X)} \leq \frac{y - \hat{\mu}^{\bar{D}}(X)}{\hat{\sigma}^{\bar{D}}(X)} \middle| D = 0, X \right\} \\ &= \hat{P} \left\{ \hat{\varepsilon} \leq \frac{y - \hat{\mu}^{\bar{D}}(X)}{\hat{\sigma}^{\bar{D}}(X)} \middle| D = 0, X \right\} \\ &= \hat{F}_0 \left\{ \frac{y - \hat{\mu}^{\bar{D}}(X)}{\hat{\sigma}^{\bar{D}}(X)} \right\}, \end{aligned} \quad (3)$$

which in turn yields $\hat{P}\{\text{risk}(X, Y) \leq r | D = 0, X\}$. For example, if we use a logistic model for $\text{risk}(X, Y)$ and write $\widehat{\text{logit risk}}(X, Y) = \hat{\theta}_0 + \hat{\theta}_1 X + \hat{\theta}_2 Y$ where $\hat{\theta}_2 > 0$, then

$$\begin{aligned} \hat{P}\{\text{risk}(X, Y) \leq r | D = 0, X\} &= \hat{P}\{\widehat{\text{logit risk}}(X, Y) \leq \text{logit}(r) | D = 0, X\} \\ &= \hat{P}\{\hat{\theta}_0 + \hat{\theta}_1 X + \hat{\theta}_2 Y \leq \text{logit}(r) | D = 0, X\} \\ &= \hat{P} \left\{ Y \leq \frac{\text{logit}(r) - \hat{\theta}_0 - \hat{\theta}_1 X}{\hat{\theta}_2} \middle| D = 0, X \right\} \\ &= \hat{F}_0 \left\{ \frac{\frac{\text{logit}(r) - \hat{\theta}_0 - \hat{\theta}_1 X}{\hat{\theta}_2} - \hat{\mu}^{\bar{D}}(X)}{\hat{\sigma}^{\bar{D}}(X)} \right\}, \end{aligned}$$

by substituting into (3). In turn, we estimate (2) as

$$\hat{P}\{\text{risk}(X, Y) \leq r | D = 0, W = c\} = \frac{\sum_{j=1}^N \hat{P}\{\text{risk}(X_j, Y) \leq r | D_j = 0, X_j\} I\{W(X_j) = c, D_j = 0\}}{N_D^c}$$

where N_D^c is the number of controls in the cohort with matching covariate value $W = c$. This estimator is then substituted into (1) to get $\hat{F}_{X,Y}^{\bar{D}}(r)$. As noted above, a nonparametric estimator substitutes the observed proportions in the cohort for $P(W = c | D = 0)$, so that $\hat{P}(W = c | D = 0) = \frac{N_D^c}{N_D}$. The semiparametric estimator then simplifies to

$$\hat{F}_{X,Y}^{\bar{D}}(r) = \hat{P}\{\text{risk}(X, Y) \leq r | D = 0\} = \frac{\sum_{j=1}^N \hat{P}\{\text{risk}(X_j, Y) \leq r | D_j = 0, X_j\} I\{D_j = 0\}}{N_D}$$

for both matched and unmatched studies.

Both nonparametric and semiparametric estimators of $F_{X,Y}^{\bar{D}}$ are accompanied by a nonparametric estimator of $F_{X,Y}^D$.

Estimates of Performance Improvement Measures

In Table 1, we presented the definitions of all performance improvement measures being studied here. Observe that estimates of $\Delta HR^D(r)$, $\Delta HR^{\bar{D}}(r)$, $\Delta B(r)$ and $\Delta AARD(r)$ follow directly from the estimators described above for the cumulative distributions of $risk(X)$ and $risk(X, Y)$ in cases and in controls. Note that since $\Delta HR^D(r)$ relies only on $F_{X,Y}^D$, what we refer to as nonparametric and semiparametric estimates of $\Delta HR^D(r)$ are in fact the same empirical estimate.

The pointwise ROC measures are also calculated directly, after noting that $ROC(p^{\bar{D}}) = 1 - F^D(r(p^{\bar{D}}))$ where $r(p^{\bar{D}})$ is such that $1 - F^{\bar{D}}(r(p^{\bar{D}})) = p^{\bar{D}}$ and $ROC^{-1}(p^D) = 1 - F^{\bar{D}}(r(p^D))$ where $r(p^D)$ is such that $1 - F^D(r(p^D)) = p^D$.

For ΔAUC , we use the usual simple empirical estimator with cohort data for the baseline value AUC_X , while we use

$$\widehat{AUC}_{(X,Y)} = \frac{1}{n_D} \sum_{i=1}^{n_D} \widehat{F}_{X,Y}^{\bar{D}} \{ \widehat{risk}(X_i, Y_i) \},$$

where the summation is over cases, for the enhanced model. Note that this is equal to the usual empirical estimator in an unmatched study but that it also yields an estimate of $P\{risk(X_j, Y_j) \leq risk(X_i, Y_i) | D_i = 1, D_j = 0\}$ in the matched design setting.

The baseline MRD is calculated empirically from the cohort values of $\widehat{risk}(X)$ while the enhanced model MRD is calculated as

$$MRD_{(X,Y)} = \frac{1}{n_D} \sum_{i=1}^{n_D} \widehat{risk}(X_i, Y_i) - \sum_{c=1}^C \widehat{E} \{ \widehat{risk}(X, Y) | D = 0, W = c \} P(W = c | D = 0).$$

Here $\widehat{E} \{ \widehat{risk}(X, Y) | D = 0, W = c \}$ are the stratum specific sample averages of $\widehat{risk}(X, Y)$ for controls in the case-control study for the nonparametric estimator. For the semiparametric estimator $\widehat{E} \{ \widehat{risk}(X, Y) | D = 0, W = c \}$ is calculated as the average of

$$\int \widehat{risk}(X_i, y) d\widehat{F}_0 \left\{ \frac{y - \mu^{\bar{D}}(X_i)}{\hat{\sigma}^{\bar{D}}(X_i)} \right\} = \frac{1}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \widehat{risk} \left\{ X_i, \frac{Y_j - \hat{\mu}^{\bar{D}}(X_j)}{\hat{\sigma}^{\bar{D}}(X_j)} \hat{\sigma}^{\bar{D}}(X_i) + \hat{\mu}^{\bar{D}}(X_i) \right\}$$

over the controls in the cohort stratum with $W = c$.

The $NRI(> 0)$ statistic uses the observed proportion of cases with $\widehat{risk}(X, Y) > \widehat{risk}(X)$ in the case-control study for the event NRI component, which requires estimation of $P\{risk(X, Y) > risk(X) | D = 1\}$. The non-event NRI component requires $P\{risk(X, Y) < risk(X) | D = 0\}$, which is estimated as a weighted average of the stratum specific observed proportions for the nonparametric estimator and as $\frac{1}{n_D} \sum_{i=1}^{n_D} \widehat{P} \{ \widehat{risk}(X_i, Y) < \widehat{risk}(X_i) | D_i = 0, X_i \}$ for the semiparametric estimator.

Further details of the performance measure estimators obtained in each scenario are presented in the Appendix, Tables 8–10.

Summary of Estimation Approaches

In Table 1, we showed that all performance improvement measures are functions of the risk distributions. Therefore, regardless of which measure is used, estimation of performance improvement is a two-fold task that requires estimating: (1) the risk functions $\text{risk}(X)$ and $\text{risk}(X, Y)$, and (2) the distributions of the risk functions in cases and in controls. We then substitute the estimated distributions into expressions for the performance improvement measures.

We estimated both risk functions parametrically using simple logistic models with linear terms. Other more flexible forms may be used in practice. In section “Estimating Risk Functions”, we presented two different modeling approaches for estimating $\text{risk}(X, Y)$ under the logistic regression framework. The first method (M_{adj}) is a commonly used approach which utilizes only the data in the case-control subset and is valid only for an unmatched design. The second method ($M_{2-stage}$) is a two-stage estimator which utilizes additional data from the cohort and is valid for both matched and unmatched designs. By comparing these two approaches to modeling the risk function, we aim to demonstrate that matching invalidates commonly used naïve analysis. Additionally, we investigate whether utilizing the parent cohort data for X improves the efficiency of risk function estimation.

In section “Estimating Distributions of Risk”, we turned our attention to the estimation of the risk distributions in cases and in controls. We estimated the distributions of $\text{risk}(X)$ using the empirical distributions estimated from the cohort. We also estimated the distribution of $\text{risk}(X, Y)$ in cases empirically. For the estimation of the risk distribution in controls, we proposed nonparametric and semiparametric approaches for matched and unmatched case-control designs. The nonparametric approach has the advantage of making no modeling assumptions for the distribution of Y given X in controls. On the other hand, the semiparametric approach does make modeling assumptions and borrows information across strata of controls, and is therefore expected to be more efficient. One would therefore use the nonparametric approach in situations where there was uncertainty about how to model the distribution of Y given X in controls. The semiparametric approach would be preferable in situations with sparse controls. Using these two approaches for estimating the risk distribution, we aim to compare the efficiency of semiparametric estimation to that of nonparametric estimation.

Finally, using the above methods, we aim to answer the question of whether matching in the nested case-control subset improves efficiency in the estimation of performance improvement measures.

Simulation Studies

We investigated the performances of the estimators and the merits of matched study designs using two small simulation studies – in the first study, we generated the data from a bivariate binormal model and in the second study, we used a real dataset.

Simulation Study 1: Bivariate Binormal Data

Data Generation

We generated bivariate binormal cohort data of size $N = 5,000$ for cases ($D = 1$) and controls ($D = 0$) with population prevalence $\rho = P(D = 1) = 0.10$, so that the cohort contained $N_D = 500$ cases and $N_{\bar{D}} = 4,500$ controls:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN} \left(\begin{pmatrix} \mu_X(D) \\ \mu_Y(D) \end{pmatrix}, \begin{pmatrix} 1 & \text{corr}(X, Y|D) \\ \text{corr}(X, Y|D) & 1 \end{pmatrix} \right)$$

where $\mu_X(0) = \mu_Y(0) = 0$ and $\mu_X(1) = \mu_Y(1) = 0.742$. The corresponding AUC values associated with X and Y alone are $\text{AUC}_X = \text{AUC}_Y = \Phi(0.742/\sqrt{2}) = 0.7$. Data for $N = 5,000$ subjects were generated, so that $\{(D_i, X_i), i = 1, \dots, N\}$ constitutes the study cohort data. A random sample of $n_D = 250$ cases were selected from the cohort and their Y values added to the dataset. For the unmatched design, Y values for a random sample of $n_{\bar{D}} = 500$ controls were also added to the dataset. For the matched design, we generated the matching variable W using quartiles of X in the control population and selected two controls randomly for each case in each of the four W strata.

Results

Using the notation M for a generic performance improvement measure, Table 2 shows mean values for estimates derived from 5,000 simulations. Estimates calculated using the adjusted case-control modeling approach for risk(X, Y) are denoted by M_{adj} , while estimates calculated using the two-stage modeling approach are denoted by $M_{2-stage}$. Bias estimates are calculated by subtracting the mean values from the true value for each measure. We see that the M_{adj} estimators are valid in unmatched designs, in the sense that mean values are close to the true values. However, M_{adj} estimators are biased in matched designs because they do not account for matching. Note that the direction and size of the bias is such that performance appears to decrease rather than increase with addition of Y to the model. In contrast the $M_{2-stage}$ estimators provide estimates that are centered around the true values in matched and unmatched designs.

Table 2 Mean estimates of improvement in prediction performance for measures defined in Table 1. Results are from 5,000 simulations of nested case-control studies ($n_D = 250, n_B = 500$) with a cohort of 5,000 subjects. Data were generated from the bivariate binormal model described in the text with $corr(X, Y|D) = 0.5$. Estimates calculated with $\widehat{M}_{adj}^{(a)}$ are denoted by M_{adj} and those calculated with $\widehat{M}_{2-stage}^{(a)}$ (X, Y) are denoted by $M_{2-stage}$. (a) Nonparametric and (b) semiparametric estimates are presented

Measure	True value	Unmatched design				Matched design			
		M_{adj}		$M_{2-stage}$		M_{adj}		$M_{2-stage}$	
		Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias
$\Delta HR^D(0.20)$	0.067	0.069	0.002	0.069	0.002	-0.169	-0.236	0.069	0.002
$\Delta HR^B(0.20)$	-0.013	-0.013	0.000	-0.013	0.000	0.056	0.069	-0.013	0.000
$\Delta B(0.20)$	0.038	0.040	0.002	0.039	0.001	-0.043	-0.081	0.039	0.001
$\Delta ROC^{-1}(0.80)$	0.046	0.048	0.002	0.048	0.002	-0.030	-0.076	0.046	0.000
$\Delta ROC(0.10)$	0.041	0.046	0.005	0.045	0.004	-0.029	-0.070	0.041	0.000
ΔAUC	0.028	0.028	0.000	0.028	0.000	-0.020	-0.048	0.028	0.000
$\Delta MRD = IDI$	0.020	0.021	0.001	0.020	0.000	-0.027	-0.047	0.020	0.000
$\Delta AARD$	0.042	0.043	0.001	0.043	0.001	-0.030	-0.072	0.043	0.001
$NRI(> 0)$	0.337	0.339	0.002	0.337	0.000	-0.270	-0.607	0.336	-0.001

(a) Nonparametric estimates

(b) Semiparametric estimates

Measure	True value	Unmatched design				Matched design			
		M_{adj}		$M_{2-stage}$		M_{adj}		$M_{2-stage}$	
		Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias
$\Delta HR^D(0.20)$	0.067	0.069	0.002	0.069	0.002	-0.169	-0.236	0.069	0.002
$\Delta HR^D(0.20)$	-0.013	-0.013	0.000	-0.013	0.000	0.056	0.069	-0.013	0.000
$\Delta B(0.20)$	0.038	0.039	0.001	0.040	0.002	-0.043	-0.081	0.040	0.002
$\Delta ROC^{-1}(0.80)$	0.046	0.046	0.000	0.046	0.000	-0.030	-0.076	0.046	0.000
$\Delta ROC(0.10)$	0.041	0.042	0.001	0.042	0.001	-0.030	-0.071	0.042	0.001
ΔAUC	0.028	0.028	0.000	0.028	0.000	-0.020	-0.048	0.028	0.000
$\Delta MRD = IDI$	0.020	0.021	0.001	0.021	0.001	-0.027	-0.047	0.020	0.000
$\Delta AARD$	0.042	0.043	0.001	0.043	0.001	-0.030	-0.072	0.043	0.001
$NRI(> 0)$	0.337	0.336	-0.001	0.337	0.000	-0.270	-0.607	0.338	0.001

Table 3 Efficiency of $M_{2-stage}$ in matched and unmatched designs relative to the nonparametric M_{adj} estimator from the unmatched design. Shown are the ratios of the standard deviations of estimates found in simulation studies divided by reference standard deviations (M_{adj} -NP; unmatched), so smaller values show more efficiency. NP and SP represent nonparametric and semiparametric estimation, respectively, of the distribution of risk(X, Y) in controls

Measure	Unmatched design		Matched design	
	$M_{2-stage}$ -NP (%)	$M_{2-stage}$ -SP (%)	$M_{2-stage}$ -NP (%)	$M_{2-stage}$ -SP (%)
$\Delta HR^D(0.20)$	75.3	75.3	74.3	74.3
$\Delta HR^{\bar{D}}(0.20)$	109.1	53.4	74.0	47.5
$\Delta B(0.20)$	99.4	77.8	82.8	75.1
$\Delta ROC^{-1}(0.80)$	99.8	87.0	95.7	88.5
$\Delta ROC(0.10)$	98.9	77.7	83.1	75.3
ΔAUC	100.0	84.1	86.1	84.0
$\Delta MRD = IDI$	71.1	69.3	65.3	64.7
$\Delta AARD$	99.5	83.4	91.2	83.7
$NRI(> 0)$	61.6	61.3	62.5	59.3

The relative efficiencies of estimators are considered in Table 3 using ratios of standard deviations, with the standard deviation of the nonparametric M_{adj} estimator in the unmatched studies as the reference.

In the unmatched design, we found that the nonparametric $M_{2-stage}$ estimator is more efficient than M_{adj} for estimating $\Delta HR^D(0.20)$, ΔMRD and $NRI(> 0)$. Interestingly, $M_{2-stage}$ performs slightly worse than M_{adj} for $\Delta HR^{\bar{D}}(0.20)$, but has similar performance to M_{adj} for all other performance measures.

To evaluate the impact of matching on efficiency we only consider $M_{2-stage}$ because M_{adj} estimators are biased. Comparing $M_{2-stage}$ in matched versus unmatched designs, we see that matching improves precision with which performance improvement is estimated for most measures. For example, with nonparametric estimation of the ROC related measures, the standard deviations in matched studies are 80–90 % the size of those in unmatched studies.

Interestingly, the improvement observed from matching can often be achieved in unmatched data by using the semiparametric estimator. In fact, for many of the measures, the efficiency is improved more by modeling $P(Y|X, D = 0)$ in an unmatched study than by matching controls to cases in the design and using the nonparametric estimator. For example, the standard deviation of the nonparametric estimate of $\Delta HR^{\bar{D}}(0.20)$ in matched studies is 74.0% of the reference, while the semiparametric estimate in unmatched studies has a standard deviation that is 53.4% of the reference. Some intuition for this result is provided by the fact that semiparametric estimation borrows information across strata of controls. While matching enriches strata with larger numbers of cases, it also makes those strata with fewer cases more sparse with respect to the number of controls. Therefore, both matched and unmatched data are prone to sparseness of controls in certain strata and nonparametric estimation suffers in such scenarios. The semiparametric approach, however, is less affected as it borrows information across strata.

Simulation Study 2: Renal Artery Stenosis Data

Study Description

The kidneys play several major regulatory roles in the human body, including regulation of blood pressure. The renal arteries aid in the proper functioning of the kidneys by supplying them with blood. Narrowing of the renal arteries is a condition termed *renal artery stenosis* (RAS); it inhibits blood flow to the kidneys and can lead to treatment-resistant hypertension.

The gold standard diagnostic test for RAS is an invasive and expensive procedure called renal angiography. In order to avoid unnecessarily performing angiography on individuals with a low likelihood of having disease, a clinical decision rule was developed to predict RAS based on patient characteristics and thus identify high-risk patients as candidates for the procedure [23].

We illustrate the proposed methodology using data from a RAS study [20]. For 426 patients, information is available on disease diagnosis from angiography, as well as age (10-year units), BMI, gender, recent onset of hypertension, presence of atherosclerotic vascular disease and serum creatinine (SCr) concentration. We model baseline risk using the first five characteristics and look to estimate the incremental value gained from adding SCr concentration to the model. Age and BMI were mean-centered. SCr concentration was log-transformed and standardized to have mean 0 and standard deviation 1. The study cohort includes 98 cases and 328 controls.

Methods

We simulated nested case-control studies using this dataset. Specifically, we resampled 426 observations with replacement from the cohort, selected all the cases and twice the number of controls, and disregarded SCr concentration data for patients who were not in the selected case-control subset. In one set of analyses the controls were selected unmatched as a simple random sample from all controls. In a second set of analyses the controls were selected to match the cases in regards to estimated baseline risk category. In particular, we created a three-level risk category variable, W , defined as: low if $\widehat{\text{risk}}(X) < 0.10$, medium if $0.10 < \widehat{\text{risk}}(X) < 0.20$ and high if $\widehat{\text{risk}}(X) > 0.20$. We selected two controls per case at random without replacement within each baseline risk category for the matched controls datasets. We also evaluated settings with 1:1 case-control ratios.

Results from Renal Artery Stenosis Dataset

Tables 4 and 5 summarize results of 1,000 nested case-control studies based on the renal artery stenosis dataset. We see that the M_{adj} estimators are only valid in

Table 4 Nonparametric estimates of improvement in prediction performance from the complete renal artery stenosis dataset and from simulated nested case-control datasets derived from it using a 1:2 case-control ratio. Shown are mean (a) nonparametric and (b) semiparametric estimates. Estimates calculated with $\widehat{\text{risk}}^{adj}(X, Y)$ are denoted by M_{adj} and those calculated with $\widehat{\text{risk}}^{2-stage}(X, Y)$ are denoted by $M_{2-stage}$. True values are obtained using the original renal artery stenosis dataset of all 426 subjects

Measure	True value	(a) Nonparametric estimates							
		Unmatched design			Matched design				
		M_{adj}		$M_{2-stage}$	M_{adj}		$M_{2-stage}$		
	Estimate	Bias	Estimate	Bias	Estimate	Bias			
$\Delta HR^D(0.40)$	0.051	0.054	0.003	0.055	0.004	-0.170	-0.221	0.065	0.014
$\Delta HR^D(0.40)$	-0.003	0.013	0.016	0.010	0.013	0.077	0.080	0.005	0.008
$\Delta B(0.40)$	0.045	0.084	0.039	0.079	0.034	0.002	-0.043	0.077	0.032
$\Delta ROC^{-1}(0.80)$	0.027	0.045	0.018	0.045	0.018	0.014	-0.013	0.050	0.023
$\Delta ROC(0.10)$	0.081	0.084	0.003	0.082	0.001	0.051	-0.030	0.081	0.000
ΔAUC	0.027	0.028	0.001	0.027	0.000	0.014	-0.013	0.028	0.001
$\Delta MRD = IDI$	0.069	0.068	-0.001	0.068	-0.001	-0.039	-0.108	0.075	0.006
$\Delta AARD$	-0.032	0.034	0.066	0.032	0.064	-0.008	0.024	0.036	0.068
$NRI(> 0)$	0.501	0.438	-0.063	0.467	-0.034	-0.290	-0.791	0.465	-0.036

Table 5 Efficiency of estimates of improvement in prediction performance in studies simulated from the renal artery stenosis dataset. Shown are standard deviations (SD) and the ratios of the standard deviations divided by reference standard deviations (M_{adj} -NP; unmatched), so smaller values show more efficiency. NP and SP represent nonparametric and semiparametric estimation of the distribution of risk (X, Y) in controls, respectively

Measure	Unmatched design			Matched design		
	M_{adj} -NP		ratio (%)	$M_{2-stage}$ -NP		ratio (%)
	SD			SD		
Case-control ratio=1:1						
$\Delta HR^D(0.40)$	0.060		85.0	0.051	85.0	0.053
$\Delta HR^D(0.40)$	0.025		120.0	0.026	104.0	0.025
$\Delta B(0.40)$	0.084		104.8	0.084	100.0	0.078
$\Delta ROC^{-1}(0.80)$	0.069		95.7	0.056	81.2	0.065
$\Delta ROC(0.10)$	0.078		92.3	0.072	92.3	0.071
ΔAUC	0.018		105.6	0.022	122.2	0.017
$\Delta MRD = IDI$	0.040		77.5	0.031	77.5	0.029
$\Delta AARD$	0.058		101.7	0.047	81.0	0.057
$NRI(> 0)$	0.237		75.1	0.170	71.7	0.203
Case-control ratio=1:2						
$\Delta HR^D(0.40)$	0.052		94.2	0.049	94.2	0.053
$\Delta HR^D(0.40)$	0.018		111.1	0.015	83.3	0.020
$\Delta B(0.40)$	0.067		103.0	0.059	88.1	0.069
$\Delta ROC^{-1}(0.80)$	0.059		98.3	0.055	93.2	0.061
$\Delta ROC(0.10)$	0.064		98.4	0.057	89.1	0.064
ΔAUC	0.015		93.3	0.013	86.7	0.015
$\Delta MRD = IDI$	0.030		86.7	0.026	86.7	0.026
$\Delta AARD$	0.049		100.0	0.044	89.8	0.052
$NRI(> 0)$	0.197		81.2	0.156	79.2	0.179

unmatched case-control studies. Interestingly, the bias in M_{adj} in matched studies is such that prediction performance appears to disimprove considerably with addition of Y when the IDI, $NRI(> 0)$ or ΔHR^D performance measures are employed. This is very similar to results in Table 2 for the simulated bivariate normal distributions. Also as in Table 2, we see that $M_{2-stage}$ is valid in matched and unmatched designs.

Comparing the efficiency of $M_{2-stage}$ to M_{adj} in unmatched designs where both are valid, we see trends in the top panel of Table 5 that are similar to those observed in Table 3. For a case-control ratio of 1:1, $M_{2-stage}$ -NP is more efficient than M_{adj} -NP, but only for ΔHR^D , ΔMRD and $NRI(> 0)$. For a larger number of controls (case-control ratio = 1:2), $M_{2-stage}$ loses some of its efficiency advantage. As before, $M_{2-stage}$ has worse performance than M_{adj} for the estimation of ΔHR^D , although again, this effect is lessened with the larger case-control ratio of 1:2.

Turning to the main question concerning efficiency due to matching, we again see some trends in the top panel of Table 5 that are similar to observations made for the bivariate binormal simulations in Table 3. Comparing $M_{2-stage}$ -NP in matched versus unmatched designs, matching appears to improve the efficiency with which ΔHR^D is estimated. However, ΔHR^D is not affected by matching and estimation of $NRI(> 0)$ may be worse in matched studies. With larger numbers of controls, we see in the bottom panel of Table 5 that there is no gain from matching with regards to efficiency of $M_{2-stage}$ -NP.

Semiparametric estimation improves efficiency much more than matching does in these simulations. Again, this is consistent with the earlier simulation results.

Bootstrap Method for Inference

Performance improvement estimates obtained from nested case-control data incorporate variability from both the cohort and the nested case-control subset. However, simple bootstrap resampling from observed data cannot be implemented in this setting, as data on Y are observed only for subjects selected in the original case-control subset. Below we discuss our proposed strategy for bootstrapping with nested case-control data.

Proposed Approach

We propose a parametric bootstrap method that combines resampling observations in the cohort and resampling residuals in the case-control subset [7]. To begin, we have the original study cohort for which X and disease status are available and a nested case-control subsample on which Y is measured. We first bootstrap a cohort

(say, cohort^{*}) from the original cohort and proceed to generate the matching variable W^* based on quartiles of X^* in the bootstrapped cohort^{*}. A matched or unmatched case-control subsample^{*} is then constructed in the same fashion as before. However, note that in this bootstrapped case-control subsample^{*}, the only subjects that have Y data are those who were selected to be in the original case-control subsample. We generate Y^* values for all subjects in the bootstrapped case-control subsample^{*} using a parametric bootstrap method combined with residual resampling.

Specifically, we use the original case-control subsample to model $Y|X, D = 0$ semiparametrically as in section “Estimating Distributions of Risk”,

$$Y^{\bar{D}} = \mu(X^{\bar{D}}) + \sigma\varepsilon.$$

Fitting this model on the original case-control subsample gives us estimated values $\hat{\mu}$, $\hat{\sigma}$ and residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{n_{\bar{D}}}$. Then, for each control^{*} in the bootstrapped case-control subsample^{*}, we use that subject’s covariate values, X^* , and sample with replacement a residual from among $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{n_{\bar{D}}}$ to generate a Y^* value using $\hat{\mu}$ and $\hat{\sigma}$:

$$Y_i^* = \hat{\mu}(X_i^*) + \hat{\sigma}\hat{\varepsilon}_i^*, i = 1, \dots, n_{\bar{D}}^*.$$

We fit a separate model for $Y|X, D = 1$ in the original case-control subsample and take a similar approach to generate $Y_1^*, \dots, Y_{n_D^*}^*$ for cases in the bootstrapped case-control subsample^{*}.

Simulation Study

We assessed the performance of the proposed bootstrap method with a simulation study using bivariate binormal data generated as in section “Data Generation”. We carried out 1,000 simulations, each time generating a new study cohort of size $N = 5,000$ and from this study cohort, selecting a nested case-control subsample of size 250 cases and 500 controls. We used both the matched and unmatched designs. Within each simulation, we carried out 200 bootstrap repetitions using the procedure described above. For each performance measure estimate obtained in that simulation, we estimated its standard error as the standard deviation across the 200 bootstrap repetitions and used it to calculate normality-based 95 % confidence intervals. Coverage was averaged over all 1,000 simulations.

Results are presented in Table 6. Not surprisingly, M_{adj} estimators, which are biased in matched designs, also generate confidence intervals with poor coverage. For all other settings, coverage of the 95 % bootstrap confidence intervals is good.

Table 6 Coverage of normality-based 95% bootstrap confidence intervals. Results are from 1,000 simulations of nested case-control studies ($n_D = 250$, $n_B = 500$) with a cohort of 5,000 subjects. 200 bootstrap repetitions were carried out in each simulation. Data were generated from the bivariate binormal model described in the text with $corr(X, Y|D) = 0.5$. Estimates calculated with \widehat{M}_{adj} (X, Y) are denoted by M_{adj} and those calculated with $\widehat{M}_{2-stage}$ (X, Y) are denoted by $M_{2-stage}$. Nonparametric and semiparametric estimates are presented

Measure	Nonparametric estimation				Semiparametric estimation			
	Unmatched design		Matched design		Unmatched design		Matched design	
	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)
$\Delta HR^D(0.20)$	95.1	95.2	1.3	95.2	95.1	95.2	1.3	95.2
$\Delta HR^B(0.20)$	95.3	94.8	1.1	96.6	94.7	94.5	95.1	95.3
$\Delta B(0.20)$	95.9	94.5	27.8	95.0	96.1	95.6	0.6	96.0
$\Delta ROC^{-1}(0.80)$	94.4	94.7	66.7	94.9	93.9	93.9	78.3	95.0
$\Delta ROC(0.10)$	94.7	94.6	55.2	95.0	94.5	93.9	0.1	96.4
ΔAUC	94.5	94.5	33.2	94.9	95.2	95.1	30.2	95.4
$\Delta MRD = IDI$	95.4	94.1	0.9	95.6	95.7	94.5	0.9	95.4
$\Delta AARD$	93.9	94.4	55.6	94.8	94.7	95.2	39.5	95.7
$NRI(> 0)$	95.3	94.5	0.1	96.0	95.3	94.5	7.0	95.7

Illustration with Renal Artery Stenosis Study

We illustrate our methodology on the renal artery stenosis dataset by simulating a single nested case-control dataset using the unmatched design and a single dataset using the matched design with a 1:2 case-control ratio. We include bootstrap standard errors and normality-based 95% confidence intervals (CIs), obtained from 500 bootstrap repetitions following the approach described in section “Bootstrap Method for Inference”. Instead of repeating numerous simulations as in section “Simulation Study 2: Renal Artery Stenosis Data”, we have a single study cohort and a single two-phase dataset here that we bootstrap from.

Results are presented in Table 7. We see that the two-phase estimates are quite different from the full-data estimates. We used only a single two-phase sample here to mimic a real-life two-phase dataset. Repeating the sampling 100 times and averaging estimates across repetitions showed that the estimates are unbiased (data not shown). The observed inconsistency is a result of sampling variability. As before, we see that a standard adjusted analysis (M_{adj}) underestimates performance improvement in a matched design. $M_{2-stage}$ produces valid estimates. Conclusions regarding the incremental value of SCr concentration are similar using any of the valid estimation methods in this setting. We use estimates from $M_{2-stage}$ with semiparametric estimation and a matched design to draw conclusions in the following paragraph.

The incremental value of SCr concentration appears to be significant using ΔMRD and NRI as the measures of interest. Values of 0 for both measures would indicate no improvement from SCr concentration. $\widehat{\Delta MRD}$ is 0.069 {95 % CI (0.013, 0.124)}, indicating that the change in the difference in mean risks between cases and controls is approximately 0.069. \widehat{NRI} is 0.547 {95 % CI (0.259, 0.836)}; given that NRI has a range of $(-2, 2)$, this seems like a moderate level of improvement in risk reclassification. Small improvements that are not statistically significant are seen using all other measures.

Discussion

Matching controls to cases on baseline risk factors is a common practice in epidemiologic studies of risk. It has also become common practice in biomarker research [29]. It allows one to evaluate from simple two-way analyses of Y and D if there is any association between Y and D and to be assured that the association is not explained by the matching factors. Matching also allows for efficient estimation of the relative risk associated with Y controlling for baseline predictors X in a risk

Table 7 Results from a matched and an unmatched two-phase study simulated from the renal artery stenosis dataset. 95 % bootstrap confidence intervals were obtained from 500 bootstrap repetitions, using $M_{2-stage}$ and M_{adj} for estimation of risk(X, Y) and (a) nonparametric and (b) semiparametric estimation of the distribution of risk(X, Y) in controls

(a) Nonparametric estimates							
Measure	Full data estimate	$M_{2-stage}$			M_{adj}		
		Estimate	Std err	95 % CI	Estimate	Std err	95 % CI
Unmatched study design							
$\Delta HR^D(0.20)$	-0.010	-0.020	0.058	(-0.135, 0.094)	-0.020	0.057	(-0.131, 0.091)
$\Delta HR^{\bar{D}}(0.20)$	0.043	0.082	0.063	(-0.042, 0.206)	0.077	0.062	(-0.045, 0.199)
$\Delta B(0.20)$	0.026	0.048	0.082	(-0.113, 0.210)	0.044	0.081	(-0.114, 0.203)
$\Delta ROC^{-1}(0.80)$	0.027	0.067	0.098	(-0.124, 0.258)	0.057	0.097	(-0.134, 0.248)
$\Delta ROC(0.10)$	0.081	0.071	0.089	(-0.103, 0.245)	0.081	0.089	(-0.094, 0.256)
ΔAUC	0.027	0.037	0.034	(-0.029, 0.103)	0.039	0.034	(-0.027, 0.105)
$\Delta MRD = IDI$	0.069	0.081	0.026	(0.031, 0.132)	0.087	0.030	(0.028, 0.146)
$\Delta AARD$	-0.032	0.006	0.048	(-0.089, 0.101)	0.037	0.047	(-0.055, 0.129)
$NRI(> 0)$	0.501	0.531	0.155	(0.226, 0.835)	0.510	0.195	(0.129, 0.892)
Matched study design							
$\Delta HR^D(0.20)$	-0.010	-0.020	0.034	(-0.087, 0.046)	-0.112	0.049	(-0.209, -0.015)
$\Delta HR^{\bar{D}}(0.20)$	0.043	0.043	0.038	(-0.031, 0.117)	0.108	0.040	(0.030, 0.185)
$\Delta B(0.20)$	0.026	0.016	0.048	(-0.078, 0.109)	-0.022	0.059	(-0.137, 0.093)
$\Delta ROC^{-1}(0.80)$	0.027	0.028	0.063	(-0.095, 0.150)	0.018	0.073	(-0.125, 0.161)
$\Delta ROC(0.10)$	0.081	0.051	0.067	(-0.081, 0.183)	0.020	0.079	(-0.134, 0.174)
ΔAUC	0.027	0.030	0.015	(0.001, 0.060)	0.021	0.019	(-0.016, 0.059)
$\Delta MRD = IDI$	0.069	0.069	0.027	(0.015, 0.122)	-0.041	0.036	(-0.112, 0.029)
$\Delta AARD$	-0.032	-0.024	0.052	(-0.126, 0.078)	-0.046	0.063	(-0.169, 0.077)
$NRI(> 0)$	0.501	0.596	0.181	(0.241, 0.951)	-0.359	0.226	(-0.801, 0.084)
(b) Semiparametric estimates							
Measure	Full data estimate	$M_{2-stage}$			M_{adj}		
		Estimate	Std err	95 % CI	Estimate	Std err	95 % CI
Unmatched study design							
$\Delta HR^D(0.20)$	-0.010	-0.020	0.058	(-0.135, 0.094)	-0.020	0.057	(-0.131, 0.091)
$\Delta HR^{\bar{D}}(0.20)$	0.043	0.059	0.063	(-0.066, 0.183)	0.051	0.064	(-0.074, 0.176)
$\Delta B(0.20)$	0.026	0.029	0.080	(-0.129, 0.186)	0.022	0.080	(-0.134, 0.178)
$\Delta ROC^{-1}(0.80)$	0.027	0.039	0.101	(-0.159, 0.236)	0.015	0.101	(-0.183, 0.212)
$\Delta ROC(0.10)$	0.081	0.102	0.087	(-0.068, 0.272)	0.112	0.087	(-0.059, 0.283)
ΔAUC	0.027	0.034	0.034	(-0.032, 0.100)	0.033	0.034	(-0.033, 0.099)
$\Delta MRD = IDI$	0.069	0.077	0.028	(0.023, 0.131)	0.080	0.029	(0.022, 0.138)
$\Delta AARD$	-0.032	0.000	0.043	(-0.084, 0.084)	0.023	0.041	(-0.058, 0.103)
$NRI(> 0)$	0.501	0.532	0.150	(0.237, 0.826)	0.463	0.187	(0.095, 0.830)
Matched study design							
$\Delta HR^D(0.20)$	-0.010	-0.020	0.034	(-0.087, 0.046)	-0.112	0.049	(-0.209, -0.015)
$\Delta HR^{\bar{D}}(0.20)$	0.043	0.035	0.027	(-0.017, 0.087)	0.080	0.034	(0.013, 0.148)
$\Delta B(0.20)$	0.026	0.009	0.042	(-0.074, 0.091)	-0.045	0.055	(-0.152, 0.062)
$\Delta ROC^{-1}(0.80)$	0.027	0.013	0.059	(-0.102, 0.128)	-0.046	0.069	(-0.181, 0.089)

(continued)

Table 7 (continued)

(b) Semiparametric estimates							
Measure	Full data estimate	$M_{2-stage}$			M_{adj}		
		Estimate	Std err	95 % CI	Estimate	Std err	95 % CI
$\Delta\text{ROC}(0.10)$	0.081	0.081	0.062	(-0.041, 0.203)	0.010	0.070	(-0.127, 0.147)
ΔAUC	0.027	0.027	0.015	(-0.002, 0.057)	0.009	0.019	(-0.027, 0.046)
$\Delta\text{MRD} = \text{IDI}$	0.069	0.069	0.028	(0.013, 0.124)	-0.044	0.038	(-0.118, 0.030)
ΔAARD	-0.032	-0.014	0.046	(-0.104, 0.076)	-0.044	0.058	(-0.159, 0.071)
$\text{NRI}(> 0)$	0.501	0.547	0.147	(0.259, 0.836)	-0.232	0.210	(-0.643, 0.179)

model for $\text{risk}(X, Y)$. However, the impact of matching on estimates of prediction performance measures has not been explored previously.

We demonstrated the intuitive result that matching invalidates standard estimates of performance improvement. Our estimators that simply adjust for population prevalence but not for matching, M_{adj} , substantially underestimated the performance of the risk model $\text{risk}(X, Y)$ and therefore underestimated the increment in performance gained by adding Y to the set of baseline predictors X . Intuitively, this underestimation can be attributed to the fact that matching causes the distribution of X to be more similar to cases in study controls than in population controls and therefore the distribution of $\text{risk}(X, Y)$ is also more similar to cases in study controls than in population controls.

We derived two-stage estimators that are valid in matched or unmatched nested case-control studies. We were unable to derive analytic expressions for the variances of these estimates. Therefore we investigated efficiency in two simple simulation studies. Our results suggest that the impact of two-stage estimation and of matching varies with the performance measure in question. In our simulations two-stage estimation in unmatched studies had little impact on efficiencies of ROC measures but was advantageous for estimating the reclassification measures $\text{NRI}(> 0)$ and $\text{IDI} = \Delta\text{MRD}$. On the other hand, matching improved efficiency of estimates of ROC related measures but did little to improve estimation of reclassification measures.

Our preferred measures of performance increment are neither ROC measures nor risk reclassification measures. We argue for use of the changes in high risk proportions of cases, $\Delta\text{HR}_D(r)$, high risk proportion of controls, $\Delta\text{HR}_{\bar{D}}(r)$, and the linear combination $\Delta\text{B}(r)$. These measures are favored due to their practical value for quantifying effects on improved medical decisions [28].

In our simulations we found that two-stage estimation improved efficiency of ΔHR_D but that matching had little to no further impact. Note that matching only affects the two-stage estimator for ΔHR_D through the influence of controls on the estimator of $\text{risk}(X, Y)$. That is, given estimates of $\text{risk}(X, Y)$, the empirical estimator of ΔHR_D is employed in both matched and unmatched designs as the cases

are a simple random sample from the cohort. We conclude that the improvement in estimating $\text{risk}(X, Y)$ that is gained with matched data does not carry over to substantially impact on estimation of the distribution of $\text{risk}(X, Y)$ in cases. On the other hand, matching improved estimation of $\Delta\text{HR}_{\bar{D}}(r)$, at least with smaller control to case ratio.

We implemented a semiparametric method that modeled the distribution of Y given X among controls. This had a profound positive influence on efficiency with which most measures were estimated, especially in unmatched designs. If one is comfortable with making necessary assumptions to model Y given X in controls, it seems that little additional efficiency is gained by using a matched design.

We recognize that the simulation scenarios we studied are limited and our conclusions may not apply to other scenarios. There are a number of factors to consider with respect to study design and estimation and changing one of these factors could affect results. In fact, we see this happen in our two simulation studies. For example, in our second simulation study, changing the case-control ratio from 1:1 to 1:2 alone lessens the advantage of matching on results. Moreover, the effect of matching is different on different performance measures. More work is needed to derive analytic results that could generalize our observations. In the meantime our practical suggestion is to use simulation studies based on the application of interest in order to guide decisions about matching and other aspects of study design. Simulation studies may be based on hypothesized joint distributions for biomarkers, as in our first simulation study (section “Simulation Study 1: Bivariate Binormal Data”). If pilot data are available one could base simulation studies on that, as we did with the renal artery stenosis data (section “Simulation Study 2: Renal Artery Stenosis Data”). Simulation studies can be used to guide the design of another larger study, by simulating both matched and unmatched nested case-control studies by varying factors related to study design and estimation approach and investigating which approaches would maximize efficiency for the performance improvement measures of interest.

Another consideration in the decision to match is that inference is complicated by matching. Asymptotic distribution theory is not available for two-stage estimators of performance measures. The difficulty in deriving analytic expressions comes from the fact that there are multiple sources of variability that must be accounted for, given the complicated analytic approach and study design. Simple bootstrap resampling cannot be implemented in this setting because the nested case-control design implies that Y is only available for the study controls. We proposed a parametric bootstrap approach that generates Y for all cohort subjects using semiparametric models for Y given X fit to the original data. We showed that this method was valid with good coverage in simulation studies. We recommend this approach with the caveat that near the null, estimates tend to be skewed and in turn, inference tends to be problematic near the null for all measures of performance improvement. We and others have noted severe problems with bootstrap methods

and inference in general for estimates of performance improvement even in cohort studies and especially with weakly predictive markers [22, 31, 36]. In practice, we recommend doing simulations similar to those suggested above to determine if valid inference is possible with the given data and study design or if the performance improvement is too close to the null. Continued effort is needed to develop methods for inference about performance improvement measures in cohort studies and then to extend them to nested case-control designs.

Acknowledgements Support for this research was provided by RO1-GM-54438 and PO1-CA-053996. The authors thank Mr. Jing Fan for his contribution to the simulation studies.

Appendix

Table 8 Estimators of performance measures: Nonparametric estimators using the baseline risk model and cohort data

Name	Estimator
$HR_X^D(r)$	$\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r, D_i = 1\}$
$HR_X^{\bar{D}}(r)$	$\frac{1}{N_{\bar{D}}} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r, D_i = 0\}$
$B_X(r)$	$\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r, D_i = 1\} - \frac{N_{\bar{D}}}{N_D} \frac{r}{(1-r)} \frac{1}{N_{\bar{D}}} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r, D_i = 0\}$
$ROC_X(p^{\bar{D}})$	$\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r(p^{\bar{D}}), D_i = 1\},$ where $r(p^{\bar{D}})$ s.t. $\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r(p^{\bar{D}}), D_i = 0\} = p^{\bar{D}}$
$ROC_X^{-1}(p^D)$	$\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r(p^D), D_i = 0\},$ where $r(p^D)$ s.t. $\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > r(p^D), D_i = 1\} = p^D$
AUC_X	$\frac{1}{N_D} \sum_{i=1}^N \frac{1}{N_D} \sum_{j=1}^N I\{\widehat{\text{risk}}(X_j) \leq \widehat{\text{risk}}(X_i), D_i = 1, D_j = 0\}$
MRD_X	$\frac{1}{N_D} \sum_{i=1}^N \widehat{\text{risk}}(X_i) I(D_i = 1) - \frac{1}{N_D} \sum_{i=1}^N \widehat{\text{risk}}(X_i) I(D_i = 0)$
\overline{AARD}_X	$\frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > \frac{N_D}{N}, D_i = 1\} - \frac{1}{N_D} \sum_{i=1}^N I\{\widehat{\text{risk}}(X_i) > \frac{N_D}{N}, D_i = 0\}$
$NRI(> 0)$	N/A

Table 9 Estimators of performance measures: Nonparametric estimators using the enhanced risk model and case-control subset data

Name	Unmatched design	Matched design
$HR_{(x,y)}^D(r)$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=1\}$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=1\}$
$HR_{(x,y)}^D(r)$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=0\}$	$\sum_{c=1}^{N_D} \frac{N_{D,c}}{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=0, W_i=c\}$
$B_{(x,y)}(r)$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=1\}$ $-\frac{N_D}{n_D} \frac{r}{1-r} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=0\}$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=1\}$ $-\frac{N_D}{n_D} \frac{r}{1-r} \sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r, D_i=0, W_i=c\}$
$ROC_{(x,y)}(p^D)$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=1\}$, where $r(p^D)$ s.t. $\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=0\} = p^D$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=1\}$, where $r(p^D)$ s.t. $\sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=0, W_i=c\} = p^D$
$ROC_{(x,y)}^{-1}(p^D)$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=0\}$, where $r(p^D)$ s.t. $\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=1\} = p^D$	$\sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=0, W_i=c\}$, where $r(p^D)$ s.t. $\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > r(p^D), D_i=1\} = p^D$
$AUC_{(x,y)}$	$\frac{1}{n_D} \sum_{i=1}^{n_D} \frac{1}{n_D} \sum_{j=1}^{n_D} I\{\widehat{\text{risk}}(X_j, Y_j) \leq \widehat{\text{risk}}(X_i, Y_i), D_i=1, D_j=0\}$	$\frac{1}{n_D} \sum_{i=1}^n \left[\sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{j=1}^n I\{\widehat{\text{risk}}(X_j, Y_j) \leq \widehat{\text{risk}}(X_i, Y_i), D_i=1, D_j=0, W_j=c\} \right]$
$MRD_{(x,y)}$	$\frac{1}{n_D} \sum_{i=1}^{n_D} \widehat{\text{risk}}(X_i, Y_i) I(D_i=1) - \frac{1}{n_D} \sum_{i=1}^{n_D} \widehat{\text{risk}}(X_i, Y_i) I(D_i=0)$	$\frac{1}{n_D} \sum_{i=1}^n \widehat{\text{risk}}(X_i, Y_i) I(D_i=1) - \sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n \widehat{\text{risk}}(X_i, Y_i) I(D_i=0, W_i=c)$
$AAARD_{(x,y)}$	$\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > \frac{N_D}{N}, D_i=1\}$ $-\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > \frac{N_D}{N}, D_i=0\}$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > \frac{N_D}{N}, D_i=1\}$ $-\sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > \frac{N_D}{N}, D_i=0, W_i=c\}$
$NRI(< 0)$	$2 \left[\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > \widehat{\text{risk}}(X_i), D_i=1\} \right]$ $-\frac{1}{n_D} \sum_{i=1}^{n_D} I\{\widehat{\text{risk}}(X_i, Y_i) > \widehat{\text{risk}}(X_i), D_i=0\}$	$2 \left[\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > \widehat{\text{risk}}(X_i), D_i=1\} \right]$ $-\sum_{c=1}^{N_D} \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i, Y_i) > \widehat{\text{risk}}(X_i), D_i=0, W_i=c\}$

Table 10 Estimators of performance measures; Semiparametric estimators using the enhanced risk model and both cohort and case-control subset data. We let superscripts ‘cohort’ and ‘cc’ denote data from the cohort and the case-control subset, respectively

Name	Estimator
$HR_{(X,Y)}^D(r)$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r, D_i^{cc} = 1\}$
$HR_{(X,Y)}^D(r)$	$1 - \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\text{logit}(r) - \hat{\theta}_0 - \hat{\theta}_1 X_j^{cohort}}{\hat{\sigma}^D(X_j^{cohort})} - \beta^D(X_j^{cohort}) \right\} I\{D_j^{cohort} = 0\}$
$B_{(X,Y)}(r)$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r, D_i^{cc} = 1\} - \frac{N_D}{N_D} \frac{r}{1-r} \left[1 - \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\text{logit}(r) - \hat{\theta}_0 - \hat{\theta}_1 X_j^{cohort}}{\hat{\sigma}^D(X_j^{cohort})} - \beta^D(X_j^{cohort}) \right\} I\{D_j^{cohort} = 0\} \right]$
$ROC_{(X,Y)}(p^D)$	$\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r(p^D), D_i^{cc} = 1\}$, where $r(p^D)$ s.t. $\frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\text{logit}(r(p^D)) - \hat{\theta}_0 - \hat{\theta}_1 X_j^{cohort}}{\hat{\sigma}^D(X_j^{cohort})} - \beta^D(X_j^{cohort}) \right\} I\{D_j^{cohort} = 0\} = 1 - p^D$
$ROC_{(X,Y)}^{-1}(p^D)$	$1 - \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\text{logit}(r(p^D)) - \hat{\theta}_0 - \hat{\theta}_1 X_j^{cohort}}{\hat{\sigma}^D(X_j^{cohort})} - \beta^D(X_j^{cohort}) \right\} I\{D_j^{cohort} = 0\}$, where $r(p^D)$ s.t. $\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r(p^D), D_i^{cc} = 1\} = p^D$
$AUC_{(X,Y)}$	$\frac{1}{n_D} \sum_{i=1}^n \left[I(D_i^{cc} = 1) \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\text{logit}(\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) - \hat{\theta}_0 - \hat{\theta}_1 X_j^{cohort}}{\hat{\sigma}^D(X_j^{cohort})} - \beta^D(X_j^{cohort})) \right\} I\{D_j^{cohort} = 0\} \right]$

$$\text{MRD}_{(X,Y)} = \frac{1}{n_D} \sum_{i=1}^n \widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) I(D_i^{cc} = 1) - \frac{1}{N_D} \sum_{i=1}^N \frac{1}{n_D} \sum_{j=1}^n \widehat{\text{risk}} \left\{ X_i^{\text{cohort}}, \frac{Y_j^{cc} - \hat{\mu}^D(X_j^{cc})}{\hat{\sigma}^D(X_j^{cc})} \hat{\sigma}^D(X_i^{\text{cohort}}) + \hat{\mu}^D(X_i^{\text{cohort}}) \right\} I(D_i^{\text{cohort}} = 0, D_j^{cc} = 0)$$

$$\text{AARD}_{(X,Y)} = \frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > \frac{N_D}{N}, D_i^{cc} = 1\} - \left[1 - \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\frac{\text{logit}(\frac{N_D}{N}) - \theta_0 - \theta_1 X_j^{\text{cohort}}}{\theta_2} - \hat{\mu}^D(X_j^{\text{cohort}})}{\hat{\sigma}^D(X_j^{\text{cohort}})} \right\} I\{D_j^{\text{cohort}} = 0\} \right]$$

$$\text{NRI}(> 0) = 2 \left(\frac{1}{n_D} \sum_{i=1}^n I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > \widehat{\text{risk}}(X_i^{\text{cohort}}), D_i^{cc} = 1\} - \left[1 - \frac{1}{N_D} \sum_{j=1}^N \hat{F}_0 \left\{ \frac{\frac{\text{logit}(\widehat{\text{risk}}(X_j^{\text{cohort}}) - \theta_0 - \theta_1 X_j^{\text{cohort}})}{\theta_2} - \hat{\mu}^D(X_j^{\text{cohort}})}{\hat{\sigma}^D(X_j^{\text{cohort}})} \right\} I\{D_j^{\text{cohort}} = 0\} \right] \right)$$

References

1. Anderson, M., Wilson, P.W., Odell, P.M., Kannel, W.B.: An updated coronary risk profile: a statement for health professionals. *Circulation* **83**, 356–362 (1991)
2. Breslow, N.E.: Statistics in epidemiology: the case-control study. *J. Am. Stat. Assoc.* **91**(433), 14–27 (1996)
3. Breslow, N.E., Cain, K.C.: Logistic regression for two-stage case-control data. *Biometrika* **75**(1), 11–20 (1988)
4. Breslow, N.E., Day, N.E.: Statistical methods in cancer research, vol. 1 - The analysis of case-control studies. International Agency for Research on Cancer, Lyon (1980)
5. Baker, S.G.: Putting risk prediction in perspective: relative utility curves. *J. Natl. Cancer Inst.* **101**, 1538–1542 (2009)
6. Bura, E., Gastwirth, J.L.: The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biom. J.* **43**, 5–21 (2001)
7. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. Chapman & Hall/CRC, New York (1993)
8. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults.: Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *J. Am. Med. Assoc.* **285**(19), 2486–2497 (2001)
9. Fears, T.R., Brown, C.C.: Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **42**, 955–960 (1986)
10. Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Shairer, C., Mulvihill, J.J.: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**(24), 1879–1886 (1989)
11. Gail, M.H., Costantino, J.P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K., Vogel, V.: Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *J. Natl. Cancer Inst.* **91**(21), 1829–1846 (1999)
12. Gordon, T., Kannel, W.B.: Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. *Am. Heart J.* **103**, 1031–1039 (1982)
13. Gu, W., Pepe, M.: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **5**, article 27 (2009)
14. Gu, W., Pepe, M.S.: Estimating the capacity for improvement in risk prediction with a marker. *Biostatistics* **10**(1), 172–186 (2009)
15. Heagerty, P.J., Pepe, M.S.: Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in children. *Appl. Stat.* **48**(4), 533–551 (1999)
16. Huang, Y., Pepe, M.S.: Semiparametric methods for evaluating risk prediction markers in case-control studies. *Biometrika* **96**(4), 991–997 (2009)
17. Huang, Y., Pepe, M.S., Feng, Z.: Evaluating the predictiveness of a continuous marker. *Biometrics* **63**(4), 1181–1188 (2007)
18. Janes, H., Pepe, M.S.: Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics* **64**, 1–9 (2008)
19. Janes, H., Pepe, M.S.: Adjusting for covariate effects on classification accuracy using the covariate adjusted ROC curve. *Biometrika* **96**, 371–382 (2009)
20. Janssens, A.C.J.W., Deng, Y., Borsboom, G.J.J.M., Eijkemans, M.J.C., Habemma, J.D.F., Steyerberg, E.W.: A new logistic regression approach for the evaluation of diagnostic test results. *Ann. Intern. Med.* **25**(2), 168–177 (2005)
21. Kannel, W.B., McGee, D., Gordon, T.: A general cardiovascular risk profile: the Framingham study. *Am. J. Cardiol.* **38**, 46–51 (1976)
22. Kerr, K.F., McClelland, R.L., Brown, E.R., Lumley, T.: Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am. J. Epidemiol.* **174**(3), 364–374 (2011)

23. Krijnen, P., van Jaarsveld, B.C., Steyerberg, E.W., Man in't Veld, A.J., Schalekamp, M.A.D.H., Habbema, J.D.F.: A clinical prediction rule for renal artery stenosis. *Stat. Med.* **129**(9), 705–711 (1998)
24. Mealiffe, M.E., Stokowski, R.P., Rhees, B.K., Prentice, R.L., Pettinger, M., Hinds, D.A.: Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natl. Cancer Inst.* **102**(21), 1618–1627 (2010)
25. Pauker, S.G., Kassierer, J.P.: The threshold approach to clinical decision making. *N. Engl. J. Med.* **302**, 1109–1117 (1980)
26. Pencina, M.J., D'Agostino, R.B. Sr., D'Agostino, R.B. Jr., Vasan, R.S.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172 (2008)
27. Pencina, M.J., D'Agostino, R.B. Sr., Steyerberg, E.W.: Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **30**, 11–21 (2011)
28. Pepe, M., Janes, H.: *Methods for evaluating prediction performance of biomarkers and tests.* University of Washington Working Paper 384. The Berkley Electronic Press, Berkley (2012)
29. Pepe, M.S., Feng, Z., Janes, H., Bossuyt, P., Potter, J.: Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J. Natl. Cancer Inst.* **100**(20), 1432–1438 (2008)
30. Pepe, M.S., Fan, J., Seymour, C.W., Li, C., Huang, Y., Feng, Z.: Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clin. Chem.* **58**(8), 1242–1251 (2012)
31. Pepe, M.S., Kerr, K.F., Longton, G., Wang, Z.: Testing for improvement in prediction model performance. *Stat. Med.* **32**(9), 1467–1482 (2013)
32. Pfeiffer, R.M., Gail, M.H.: Two criteria for evaluating risk prediction models. *Biometrics* **67**, 1057–1065 (2011)
33. Prentice, R.L., Pyke, R.: Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411 (1979)
34. Truett, J., Cornfield, J., Kannel, W.: A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chronic Dis.* **20**, 511–524 (1967)
35. Vickers, A.J., Elkin, E.B.: Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making.* **26**, 565–574 (2006)
36. Vickers, A.J., Cronin, A.M., Begg, C.M.: One statistical test is sufficient for assessing new predictive markers. *BMC Med. Res. Methodol.* **11**(1), 13 (2011)

ROC Analysis for Multiple Markers with Tree-Based Classification

Mei-Cheng Wang and Shanshan Li

Abstract Multiple biomarkers are frequently observed or collected for detecting or understanding a disease. The research interest of this paper is to extend tools of ROC analysis from univariate marker setting to multivariate marker setting for evaluating predictive accuracy of biomarkers using a tree-based classification rule. Using an arbitrarily combined and-or classifier, an ROC function together with a weighted ROC function (WROC) and their conjugate counterparts are introduced for examining the performance of multivariate markers. Specific features of the ROC and WROC functions and other related statistics are discussed in comparison with those familiar properties for univariate marker. Nonparametric methods are developed for estimating the ROC and WROC functions, and area under curve (AUC) and concordance probability. With emphasis on population average performance of markers, the proposed procedures and inferential results are useful for evaluating marker predictability based on multivariate marker measurements with different choices of markers, and for evaluating different and-or combinations in classifiers.

Introduction

The Receiver Operating Characteristic (ROC) analysis has been widely used as tools for assessing the discriminant performance for biomarkers. Based on a univariate or combined-to-univariate marker, the ROC curve is known as a plot of the true positive rate versus the false positive rate for each possible cut point, for summarizing

The paper appeared in volume 19 (2013) of *Lifetime Data Analysis*.

M.-C. Wang (✉) • S. Li

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,
615 N. Wolfe Street, Baltimore, MD 21205, USA
e-mail: mcwang@jhsph.edu; shli@jhsph.edu

sensitivity and specificity of a binary classifier system when marker measurements are continuous. In nonparametric, semiparametric or parametric models, the ROC curve and its associated measures such as area under curve (AUC) or partial area under curve (pAUC) have been used as useful indices for evaluating the predictive accuracy of markers or diagnostic tests [17]. In statistical literature, different measures have been developed to summarize and compare the predictive accuracy of biomarkers ([2, 8] among others).

This paper considers situations when multiple markers (M_1, M_2, \dots, M_k) are available for classification of disease state. The research interest is to establish criterion and tools for assessing predictive accuracy based on multivariate markers or multivariate test measurements, (M_1, M_2, \dots, M_k) , from observed data or a training data set. The proposed work includes at least two types of applications: (i) to quantify the result of dual or multiple readings from a single diagnostic test, or readings from multiple tests; (ii) to evaluate the predictability of combined multiple markers for a disease, where each marker characterizes a specific biological function for the disease. For the first type of applications, (i), multiple reading is employed for either reducing uncertainty of test classification or comparison of multiple diagnostic modalities [9, 15]. Applications of the second type, (ii), are important when multivariate markers are used as prognostic measurements for predicting or understanding the disease.

To analyze multiple marker data, several approaches have been developed to handle the correlation structure of marker measurements for different research goals. The most common approach is perhaps to combine multiple markers into a single composite score using logistic regression model, and evaluate the predictability of markers by the one-dimensional composite score [14]. For high-dimensional markers, or when markers come from different biological sources, it may not be analytically appropriate to combine the markers into a composite score and, in such situations, the tree-based regression model could serve as a good alternative for identifying a classification rule. The tree-based classification method is sometimes referred to as recursive partitioning, which is frequently used in data mining, machine learning and clinical practice as a predictive model [3, 22]. For example, Baker [1] and Etzioni et al. [6] considered discretized markers by keeping the marker values in multi-dimensional settings and proposed new definitions for ROC curves.

When markers are continuous, Jin and Lu [13] considered bivariate markers and proposed to use the area under the upper boundary of ROC region to evaluate diagnostic utilities. Jin and Lu's work can be viewed as an extension of Baker's approach [1] from discrete markers to continuous markers. Wang and Li [21] defined an ROC function for bivariate continuous markers via generalized inverse set of the quantile function FP , where the ROC function possesses a conditional expectation expression. In this paper, we generalize Wang and Li's results from bivariate marker to multivariate marker setting, and develop methods and inference for ROC analysis.

Assume a k -dimensional marker vector (M_1, M_2, \dots, M_k) is available and the disease state is determined by a sequence of arbitrarily combined and-or classifier with positivity specified in either direction of marker values; for example, $I((M_1 \geq$

m_1 or $M_2 < m_2$) and ($M_3 < m_3$ or $M_4 \geq m_4$)). This extension links to potential applications related to classification tree with binary decision diagrams. The research interest is to establish criterion and tools for assessing predictive accuracy based on multivariate markers, (M_1, M_2, \dots, M_k) . Specifically, the ROC function is extended from univariate case to multivariate case, and a weighted ROC (WROC) function is introduced for examining the performance of predictive accuracy with arbitrarily combined and-or classifiers.

Let $(M_{l1}, M_{l2}, \dots, M_{lk})$, $l = 0, 1$, be the marker vector for a non-diseased or diseased subject. Let the arbitrarily combined and-or classifier be expressed as $I\{(M_{l1}, M_{l2}, \dots, M_{lk}) \in D(m_1, m_2, \dots, m_k)\}$ with $D(m_1, m_2, \dots, m_k) \subseteq R^k$ defined as the region for marker-based positivity. To simplify notation and formulation, hereafter we shall use bold face \mathbf{m} to represent the vector (m_1, m_2, \dots, m_k) , and let \mathbf{m}_l and \mathbf{M}_l , $l = 0, 1$, represent the vectors $(m_{l1}, m_{l2}, \dots, m_{lk})$ and $(M_{l1}, M_{l2}, \dots, M_{lk})$. Define the false and true positive rates respectively as

$$FP(\mathbf{m}) = P\{\mathbf{M}_0 \in D(\mathbf{m})\}, \quad TP(\mathbf{m}) = P\{\mathbf{M}_1 \in D(\mathbf{m})\}$$

The research interest is to extend rules and tools from univariate marker to multivariate marker setting for assessment of predictive accuracy of markers.

Using the US Alzheimer's Disease Neuroimaging Initiative (ADNI) data set as an example, the biomarkers of interest include measurements from different biological systems related to neuroimaging, genetics, CSF (Cerebrospinal fluid) and cognition. As the k markers are identified from different biological sources, it may not be appropriate to combine them using, say, a linear combination of the measurements. The and-or classifier also signifies the importance of interaction between markers. For example, using an Alzheimer's Disease study that the authors are currently involved (the BIOCARD study at Johns Hopkins School of Medicine), decreases in CSF Amyloid beta-42 and/or increases in total tau or phosphorylated-tau (p-tau) are hypothesized as strong predictors for AD or AD-related symptoms. It would be interesting to keep the k markers in multivariate setting and explore their respective roles and interaction nonparametrically.

The paper is organized as follows. Section "Univariate Marker Case" briefly reviews some of the fundamental definitions and properties for univariate ROC analysis, where emphasis is placed on those which will be extended to multivariate setting. In sections "Multivariate Markers: ROC, WROC and AUC" and "Other Types of ROC and WROC Functions", a set of ROC and ROC-related functions are introduced with discussion focused on contrasting features between univariate and multivariate cases. Section "Nonparametric Estimation" considers nonparametric estimators for ROC-related functions, AUC and concordance probabilities. Simulation and a real data analysis are presented in section "Simulation and Data Example" to illustrate the applicability of the proposed procedures. Section "Discussion" concludes the paper with a brief discussion.

Univariate Marker Case

In the section we consider the univariate marker case, $k = 1$. Suppose the disease outcome D takes binary values 0 or 1, and M is a continuous marker variable. Let M_0 and M_1 respectively represent the marker variable from non-diseased ($D = 0$) and diseased ($D = 1$) group. Define $TP(m) = P(M_1 > m) = P(M > m | D = 1)$ as the true positive rate (sensitivity), and $FP(m) = P(M_0 > m) = P(M > m | D = 0)$ the false positive rate ($1 - \text{specificity}$). Assume M_0 and M_1 are independent. Define $F_0(m) = 1 - FP(m)$ and $F_1(m) = 1 - TP(m)$ respectively as the cumulative distribution function of M_0 and M_1 .

There are multiple ways to define the ROC function for a univariate marker. A mathematically simple definition $ROC(q) = TP[FP^{-1}(q)]$, $q \in [0, 1]$, evaluates the magnitude of true positive rate at controlled false positive rate through inverse functional mapping between FP and TP . The comparison of two ROC functions from different markers should thus be interpreted as the comparison of TP values with the same FP rate. The partial area under ROC curve for false positive rate less than p , $0 \leq p \leq 1$, is defined as $AUC(p) = \int I(0 \leq q \leq p) ROC(q) dq$. The area under ROC curve is defined as the total area with the FP rate ranging from 0 to 1, that is, $AUC(1)$. Define the partial concordance probability as $CON(p) = P(M_1 > M_0, FP(M_0) \leq p)$. For univariate marker model, the quantile variable $Q_0 = FP(M_0)$ is Uniform $[0, 1]$ distributed and thus $CON(p)$ can be calculated using probability measure on (M_1, Q_0) and is simplified to

$$\begin{aligned} CON(p) &= P(M_1 > FP^{-1}(Q_0), Q_0 \leq p) = \int_0^p \int I(m_1 > FP^{-1}(q)) dF_1(m_1) dq \\ &= \int_0^p ROC(q) dq = AUC(p) \end{aligned}$$

Thus, an alternative way to define $ROC(p)$ is to obtain it as the derivative of the partial concordance probability with respect to p , namely $ROC(p) = CON'(p)$. By definition, $CON(p)$ can also be expressed as

$$CON(p) = \int \int I(m_1 > m_0) I(FP(m_0) \leq p) dF_1(m_1) dF_0(m_0) \tag{1}$$

The equivalence between $CON(p)$ and $AUC(p)$ has led to development of non-parametric approaches for estimating $AUC(p)$ using the formula in (1). Dodd and Pepe [4] showed that the partial area under curve possesses a concordance probability expression: Let $p_0^* = FP(TP^{-1}(p_0))$ and assume $p_0^* < p_1$, then

$$\int I(p_0^* \leq q < p_1) ROC(q) dq = P(M_1 > M_0, FP^{-1}(p_1) < M_0 \leq TP^{-1}(p_0)) \tag{2}$$

Thus, the partial concordance probability coincides with the partial AUC restricted to the interval that false positive rate less than p_1 and true positive rate greater than p_0 . As proposed by Dodd and Pepe [4], by plugging the empirical distributions

of M_0 and M_1 into (1) and (2), the partial area-under-curve can be estimated by nonparametric U-statistics. The above properties will be extended to multivariate marker case for further analytical developments.

An alternative approach can be adopted by reversing the roles of true and false positive rates to define a function similar to the ROC function:

$$ROC^*(q) = FP[TP^{-1}(q)], \quad q \in (0, 1) \tag{3}$$

By property of composite function, it is seen that

$$ROC^*(q) = ROC^{-1}(q) \tag{4}$$

Clearly, since the mapping $ROC(q)$ is one-to-one, the function $ROC^*(q)$ consists the same amount of information as that of $ROC(q)$. Graphically, $ROC(q)$ and $ROC^*(q)$ are symmetric with respect to the diagonal line which connects points $(0, 0)$ and $(1, 1)$. Thus, $ROC(q) + ROC^*(1 - q) = 1$ and the sum of area under ROC curve and area under ROC^* curve equals 1. In section “Other Types of ROC and WROC Functions”, for multivariate marker model, a function parallel to $ROC^*(q)$ will be introduced and some interesting relationships similar to or different from those of univariate maker case will be explored.

Multivariate Markers: ROC, WROC and AUC

Now consider continuous markers and classification rule in multivariate setting. Suppose \mathbf{M}_0 and \mathbf{M}_1 are independent k -dimensional marker vectors from non-diseased group ($D = 0$) and diseased group ($D = 1$) respectively. Define

$$FP(\mathbf{m}) = P\{\mathbf{M}_0 \in D(\mathbf{m})\} ,$$

$$TP(\mathbf{m}) = P\{\mathbf{M}_1 \in D(\mathbf{m})\} .$$

Let $F_0(\mathbf{m}) = P(M_{01} \leq m_1, M_{02} \leq m_2, \dots, M_{0k} \leq m_k)$ be the cumulative distribution function for non-diseased population, and $F_1(\mathbf{m}) = P(M_{11} \leq m_1, M_{12} \leq m_2, \dots, M_{1k} \leq m_k)$ the cumulative distribution function for diseased population. Define the quantile variable $Q_0 = FP(\mathbf{M}_0)$ and denote by H_0 the distribution function of Q_0 . As an important feature of multivariate markers, in general Q_0 is not uniformly distributed. The distribution of Q_0 depends on the classifier as well as the probability structure of \mathbf{M}_0 , and therefore varies from marker vector to marker vector.

Definition of ROC Function

When marker measurements are multivariate, the function $FP(\mathbf{M}_0)$ is not a one-to-one transformation, which implies that the ROC function for univariate marker, $TP(FP^{-1}(q))$, can not be used for multivariate marker case. Wang and Li [21] considered bivariate marker models and defined an ROC function via generalized inverse set of the quantile function FP , where the ROC function possesses a conditional expectation expression. For multivariate markers, instead of using the generalized inverse set to conceptualize the ROC function, the ROC function is defined as the average of the true positive rate conditioning on the set of marker values with false positive rate q , where the conditional average is calculated subject to the non-diseased population:

$$ROC(q) = E[TP(\mathbf{M}_0) \mid FP(\mathbf{M}_0) = q] \quad (5)$$

There are a few characteristics of $ROC(q)$ in (5), which may or may not be similar to characteristics of the ROC function for univariate marker:

- The value of the ROC function in (5) is bounded between 0 and 1.
- The function $ROC(q)$ may not be an increasing function in q , $0 \leq q \leq 1$.
- If the distributions of \mathbf{M}_0 and \mathbf{M}_1 are the same (i.e., the marker vector is non-predictive for disease), then for each Borel set $D(m_1, m_2, \dots, m_k)$, one has $TP(m_1, m_2, \dots, m_k) = FP(m_1, m_2, \dots, m_k)$. This implies $TP(\mathbf{M}_0) = FP(\mathbf{M}_0)$ with probability one and

$$E[TP(\mathbf{M}_0) \mid FP(\mathbf{M}_0) = q] = q.$$

Thus, if the markers are non-predictive for disease, the ROC function coincides with the diagonal line which connects points (0,0) and (1,1), which is similar to the ROC function for univariate marker.

- When the markers are predictive subject to the classifier $D(m_1, m_2, \dots, m_k)$, it means that $TP(m_1, m_2, \dots, m_k) \geq FP(m_1, m_2, \dots, m_k)$ for each $(m_1, m_2, \dots, m_k) \in R^k$, and this implies $TP(\mathbf{M}_0) \geq FP(\mathbf{M}_0)$ with probability one and

$$ROC(q) = E[TP(\mathbf{M}_0) \mid FP(\mathbf{M}_0) = q] \geq E[FP(\mathbf{M}_0) \mid FP(\mathbf{M}_0) = q] = q,$$

for $0 \leq q \leq 1$. Thus, the ROC function is above the diagonal line if the markers are predictive for disease.

WROC and AUC

In use of the ROC function, a question of interest is whether the function in (5) can be used for comparisons of markers' predictive accuracy at population level. To address the question, we recall that for univariate marker the area under ROC curve is calculated with uniform distribution on q -axis (i.e., FP-axis).

For multivariate markers, the ROC function defined in (5) can be used to compare the performance of true positive rate locally by conditioning on $FP(\mathbf{M}_0) = q$. To evaluate multivariate markers' predictability unconditionally, the evaluation should take into account the distribution of Q_0 besides the use of the conditionally defined ROC function.

Using the probability distribution of Q_0 , the AUC can be naturally defined as the area under ROC curve subject to Lebesgue integration with measure H_0 on q -axis, namely $AUC = \int ROC(q)dH_0(q)$, or equivalently,

$$AUC = \int_0^1 ROC(q) \cdot h_0(q) dq \quad (6)$$

where $h_0(q)$ is the derivative of $H_0(q)$, which is assumed to exist. Define

$$WROC(q) = ROC(q) \cdot h_0(q)$$

as the weighted ROC (*WROC*) function. Note that $WROC(q)$ is the unconditional average of the true positive rate with fixed false positive rate q :

$$WROC(q) = E[TP(\mathbf{M}_0)I(FP(\mathbf{M}_0) = q)]. \quad (7)$$

It is seen that AUC is interpreted as area under *WROC* curve with uniform measure over the unit interval $[0, 1]$. Subsequently, the partial area under *WROC* curve can be defined as

$$AUC(p) = \int_0^p WROC(q) dq, \quad (8)$$

which can be used for comparison of markers in terms of their population-average predictability.

The concordance probability is naturally defined as $CON = P(\mathbf{M}_1 \in D(\mathbf{M}_0))$. Next we prove the equivalence between the concordance probability and the area under *WROC* curve, which is an extension of a property for univariate marker [4]:

$$\begin{aligned} CON &= P(\mathbf{M}_1 \in D(\mathbf{M}_0)) = \int \int I(\mathbf{m}_1 \in D(\mathbf{m}_0)) dF_1(\mathbf{m}_1)dF_0(\mathbf{m}_0) \\ &= \int TP(\mathbf{m}_0) dF_0(\mathbf{m}_0) = \int_0^1 E[TP(\mathbf{M}_0) | Q_0 = q] \cdot h_0(q) dq \\ &= \int_0^1 WROC(q) dq = AUC \end{aligned} \quad (9)$$

With an additional constraint on the false positive rate p , $0 \leq p \leq 1$, the partial concordance probability can be expressed as

$$CON(p) = P(\mathbf{M}_1 \in D(\mathbf{M}_0), FP(\mathbf{M}_0) \leq p),$$

where the full concordance probability corresponds to the special case $p = 1$. The partial concordance probability is

$$\begin{aligned}
 CON(p) &= P(\mathbf{M}_1 \in D(\mathbf{M}_0), FP(\mathbf{M}_0) \leq p) \\
 &= \int \int I(\mathbf{m}_1 \in D(\mathbf{m}_0))I(FP(\mathbf{m}_0) \leq p) dF_1(\mathbf{m}_1)dF_0(\mathbf{m}_0) \\
 &= \int TP(\mathbf{m}_0)I(FP(\mathbf{m}_0) \leq p) dF_0(\mathbf{m}_0) = \int_0^p E[TP(\mathbf{M}_0) | Q_0 = q] \cdot h_0(q) dq \\
 &= \int_0^p WROC(q) dq = AUC(p) \tag{10}
 \end{aligned}$$

The equivalence between $CON(p)$ and $AUC(p)$ is again an extension of the result from univariate marker model to multivariate marker model. Further, with the restrictions that the false positive rate is less than or equal to p and that the true positive rate is greater than q , the formula in (10) can be extended to

$$\begin{aligned}
 CON(p, q) &= P(\mathbf{M}_1 \in D(\mathbf{M}_0), FP(\mathbf{M}_0) \leq p, TP(\mathbf{M}_1) > q) \\
 &= \int \int I(\mathbf{m}_1 \in D(\mathbf{m}_0))I(FP(\mathbf{m}_0) \leq p, TP(\mathbf{m}_1) > q) dF_1(\mathbf{m}_1)dF_0(\mathbf{m}_0) ,
 \end{aligned}$$

which is a useful formula for constructing a U-statistic in estimation of the concordance probability with two-sided constraints. It is also clear that $CON(p, 0) = AUC(p)$.

Nonparametric Estimation

Suppose the observations include independent samples of iid copies of \mathbf{M}_0 and iid copies of \mathbf{M}_1 , where marker vectors are represented by $\{\mathbf{M}_{i,0} : i = 1, \dots, n_0\}$ and $\{\mathbf{M}_{j,1} : j = 1, \dots, n_1\}$, and realization values by $\{\mathbf{m}_{i,0} : i = 1, \dots, n_0\}$ and $\{\mathbf{m}_{j,1} : j = 1, \dots, n_1\}$, respectively from non-diseased and diseased populations. In this section we consider nonparametric approaches for estimation of ROC, WROC, AUC and CON. Denote by $\widehat{TP}, \widehat{FP}, \widehat{F}_1$ and \widehat{F}_0 respectively the empirical distribution of the corresponding function. For those p with $FP(\mathbf{m}_{i,0}) = p$, initially one can use a crude empirical estimate $TP(\mathbf{m}_{i,0})$ to estimate $ROC(p)$. Or, alternatively, we can consider the ROC function in its form as a conditional expectation in (5), $ROC(q) = E[TP(\mathbf{M}_0)|FP(\mathbf{M}_0) = q]$, and construct a kernel average estimate, which can be thought of as a smoothed version of the crude empirical estimate, to estimate $ROC(q)$:

$$\widehat{ROC}(p) = \frac{\int \widehat{TP}(\mathbf{m}_0) \cdot k(\frac{p - \widehat{FP}(\mathbf{m}_0)}{b}) d\widehat{F}_0(\mathbf{m}_0)}{\int k(\frac{p - \widehat{FP}(\mathbf{m}_0)}{b}) d\widehat{F}_0(\mathbf{m}_0)} = \frac{\sum_{i=1}^{n_0} \widehat{TP}(\mathbf{m}_{i,0}) \cdot k(\frac{p - \widehat{FP}(\mathbf{m}_{i,0})}{b})}{\sum_{i=1}^{n_0} k(\frac{p - \widehat{FP}(\mathbf{m}_{i,0})}{b})} ,$$

where the kernel $k(\cdot)$ is a mean zero density function and b is a bandwidth [7].

Note that the ROC function in (5) is defined as the average of true positive rate given a fixed value of the false positive rate, where the calculation of the conditional expectation is through the two one-dimensional variables $TP(\mathbf{M}_0)$ and $FP(\mathbf{M}_0)$. Thus, the ‘curse of dimensionality’ does not occur when the ROC function is estimated nonparametrically. A nonparametric estimator of $WROC(p)$ can be constructed by estimating the derivative of $CON(p)$ in (10) using kernel estimation technique:

$$\widehat{WROC}(p) = \frac{1}{b} \int \widehat{TP}(\mathbf{m}_0) \cdot k\left(\frac{p - \widehat{FP}(\mathbf{m}_0)}{b}\right) d\widehat{F}_0(\mathbf{m}_0) = \frac{1}{n_0 b} \sum_{i=1}^{n_0} \widehat{TP}(\mathbf{m}_{i,0}) \cdot k\left(\frac{p - \widehat{FP}(\mathbf{m}_{i,0})}{b}\right)$$

which is seen to be the same as the product of $\widehat{ROC}(p)$ and the kernel estimate of $h(p)$,

$$\frac{1}{b} \int k\left(\frac{p - \widehat{FP}(\mathbf{m}_0)}{b}\right) d\widehat{F}_0(\mathbf{m}_0) .$$

Based on the equivalence between $AUC(p)$ and $CON(p)$, a nonparametric estimator of $AUC(p)$ can be obtained:

$$\widehat{AUC}(p) = \int \int I(\mathbf{m}_1 \in D(\mathbf{m}_0)) I(\widehat{FP}(\mathbf{m}_0) \leq p) d\widehat{F}_1(\mathbf{m}_1) d\widehat{F}_0(\mathbf{m}_0) \quad (11)$$

With the restriction that the false positive rate is less than or equal to p and the true positive rate greater than q , the formula in (11) can be extended to

$$\begin{aligned} \widehat{CON}(p, q) &= \int I(\mathbf{m}_1 \in D(\mathbf{m}_0)) \cdot I(\widehat{FP}(\mathbf{m}_0) \leq p, \widehat{TP}(\mathbf{m}_1) > q) d\widehat{F}_1(\mathbf{m}_1) d\widehat{F}_0(\mathbf{m}_0) \\ &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(\mathbf{m}_{j,1} \in D(\mathbf{m}_{i,0})) \cdot I(\widehat{FP}(\mathbf{m}_{i,0}) \leq p, \widehat{TP}(\mathbf{m}_{j,1}) > q) , \end{aligned}$$

where the estimator has the form of a U-statistic [12].

Theorem 1. Let $N = n_0 + n_1$. Assume $0 < \lim_{N \rightarrow \infty} n_0/N = \lambda < 1$. Then, for $p, q \in [0, 1]$, (i) $\widehat{CON}(p, q)$ converges to $CON(p, q)$ in probability as $N \rightarrow \infty$, and (ii) $\sqrt{N}\{\widehat{CON}(p, q) - CON(p, q)\} \xrightarrow{d} \text{Normal}(0, \sigma^2)$, where σ^2 is specified in the Appendix.

The asymptotic results require that N be large and $0 < n_0/N = \lambda < 1$. This condition is generally satisfied with random sampling while disease status D could be either random or fixed, which is respectively relevant in prospective and retrospective (case-control) study. In the case D is random, N corresponds to the total sample size and n_0/N converges to $P(D = 0) = \lambda, 0 < \lambda < 1$, with probability 1 and the asymptotic normality holds with the usual interpretation.

Other Types of ROC and WROC Functions

Similar to considerations of using (3) in univariate marker case, for multivariate markers we may want to consider a function with the roles of true and false positive rates reversed. Define $Q_1 = TP(\mathbf{M}_1)$, and let H_1 and h_1 respectively be the distribution function and density function of Q_1 . Then, similar to the structure of $ROC(q)$, where $ROC(q) = E[TP(\mathbf{M}_0) | FP(\mathbf{M}_0) = q]$, for multivariate markers we may define

$$ROC^*(q) = E[FP(\mathbf{M}_1) | TP(\mathbf{M}_1) = q].$$

In general, as a part of the main features which distinguish the univariate and multivariate ROC inferences, the functional transformation $ROC^*(q)$ is not one-to-one and therefore does not have the inverse functional relationship with $ROC(q)$. Further define

$$\overline{ROC}(q) = E[FN(\mathbf{M}_0) | TN(\mathbf{M}_0) = q] \quad \text{and} \quad \overline{ROC}^*(q) = E[TN(\mathbf{M}_1) | FN(\mathbf{M}_1) = q]$$

where $FN(\mathbf{m}) = P(\mathbf{M}_1 \notin D(\mathbf{m}))$ is the false negative rate and $TN(\mathbf{m}) = P(\mathbf{M}_0 \notin D(\mathbf{m}))$ is the true negative rate. The weighted functions corresponding to ROC^* , $\overline{ROC}(q)$ and $\overline{ROC}^*(q)$ can be defined in such ways similar to the $WROC$ function: for $0 < q < 1$,

$$\begin{aligned} WROC(q) &= ROC(q) \cdot h_0(q); \quad WROC^*(q) = ROC^*(q) \cdot h_1(q) \\ \overline{WROC}(q) &= \overline{ROC}(q) \cdot h_0(1-q); \quad \overline{WROC}^*(q) = \overline{ROC}^*(q) \cdot h_1(1-q) \end{aligned}$$

These weighted ROC functions serve to study the performance of predictive accuracy for multivariate markers from different perspectives. For example, $WROC^*(p)$ serves to study the performance of false positive rate with true positive rate controlled at value p . It is shown in the appendix that

$$\begin{aligned} ROC(q) + \overline{ROC}(1-q) &= 1; \quad ROC^*(q) + \overline{ROC}^*(1-q) = 1 \\ WROC(q) + \overline{WROC}(1-q) &= h_0(q); \quad WROC^*(q) + \overline{WROC}^*(1-q) = h_1(q) \end{aligned}$$

Thus, the function ROC provides the same amount of information as \overline{ROC} , and similarly ROC^* is as informative as \overline{ROC}^* . Also, with knowledge of $h(q)$, $WROC(q)$ provides the same amount of information as \overline{WROC} for predictive accuracy, and similar argument applies to the relationship between $WROC^*$ and \overline{WROC}^* . Essentially, the pair-wise relationship can be thought of as the conjugate partnership.

For evaluation based on partial area under curve, subject to either smaller FP ($FP \leq p$) or larger TP ($TP > q$), choices of these weighted ROC functions should be $WROC$ and \overline{WROC}^* so that maximization of area under curve would make sense. These two weighted ROC functions together with their corresponding ROC functions are used in our simulation to study the performance of the proposed criteria and methods for multivariate markers. Note that the partial concordance probability for true negativity is $\overline{CON}^*(p) = P(\mathbf{M}_0 \notin D(\mathbf{M}_1), FN(\mathbf{M}_1) \leq p)$. By similar technique employed in section “WROC and AUC”, it can be proved that this concordance probability coincides with the area under $\overline{WROC}^*(p)$ function, $\overline{CON}^*(p) = \overline{AUC}^*(p)$, and therefore a U-statistic $\widehat{\overline{CON}^*}(p)$ can be constructed to estimate $\overline{CON}^*(p)$.

In case of requiring both $FP \leq p$ and $TP > q$, these ROC or WROC functions cannot be used for evaluation, but $CON(p, q)$ can be used and estimated by the technique described in section “Multivariate Markers: ROC, WROC and AUC”. For estimation of \overline{ROC}^* , \overline{WROC}^* and $\overline{CON}^*(p, q)$, nonparametric estimates can be constructed using methods similar to those for the functions ROC , $WROC$ and $CON(p, q)$. Also, a property similar to Theorem 1 can be established for $\widehat{\overline{CON}^*}(p)$ by the same technique.

Remark. By setting $M_{l1} = M_{l2} = \dots = M_{lk}$, $l = 0, 1$, univariate marker model can be viewed as a degenerated case of multivariate markers. For this degenerated case, the quantile variable $Q_0 = FP(\mathbf{M}_0)$ and $Q_1 = TP(\mathbf{M}_1)$ both follow Uniform[0, 1] distribution, and $\overline{ROC}(q) = FN(TN^{-1}(q))$ and $\overline{ROC}^*(q) = TN(FN^{-1}(q))$. In this case, each of the WROC functions coincides with their counterpart of ROC functions. Further, besides the relationship $ROC(q) + \overline{ROC}(1 - q) = 1$ and $ROC^*(q) + \overline{ROC}^*(1 - q) = 1$, it is seen that $ROC^*(q) = ROC^{-1}(q)$, which implies that each of the four ROC functions provides the same amount of information as the other three functions for predictive accuracy of the marker.

Simulation and Data Example

Simulation

To show the performance of predictive accuracy for multivariate markers, we conduct simulation studies under different scenarios. We compare ROC and WROC curves for multivariate markers under each scenario, along with the weight function $h_0(q)$. We also compare univariate and multivariate marker cases to evaluate the gain and loss by using multiple markers.

Since this paper is a generalization of the bivariate ROC analysis of Wang and Li [21], we take $k \geq 3$ markers for evaluation. For simplicity, we take $k = 3$. Consider the simulation model where (M_{01}, M_{02}, M_{03}) and (M_{11}, M_{12}, M_{13}) follow a multivariate normal distribution. By convention we assume higher marker value

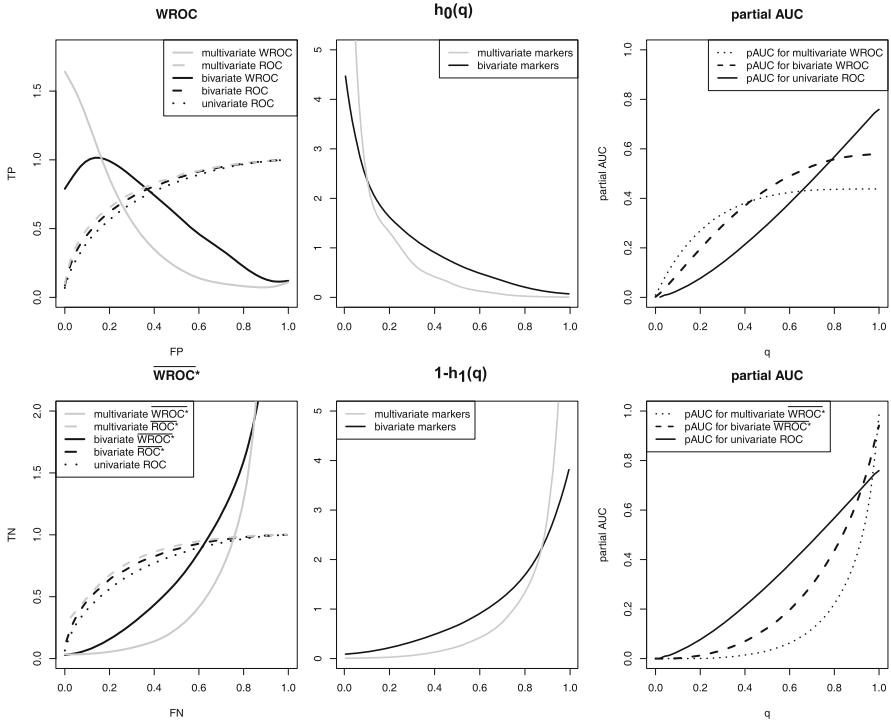


Fig. 1 Simulation for classifier $I(M_1 > m_1, M_2 > m_2, M_3 > m_3)$ with $\rho_0 = \rho_1 = \mathbf{0}$

indicates presence of disease. Let $N_1 = 200$ be the number of diseased individuals and $N_2 = 200$ be the number of non-diseased individuals. We generate data so that (M_{01}, M_{02}, M_{03}) have mean $(0, 0, 0)$ and unit deviations. We generate data so that (M_{11}, M_{12}, M_{13}) have mean $(1, 1, 1)$ and unit deviations. Let $\rho_l = (\rho_{l12}, \rho_{l23}, \rho_{l13})$, $l = 0, 1$, where ρ_{lij} denote the correlation between M_{li} and M_{lj} . We consider different scenarios according to different correlations ρ_l . The ROC analysis for univariate marker is based on data generated from the distributions of M_{11} , bivariate ROC analysis is based on data generated from the distribution of (M_{11}, M_{12}) , and multivariate ROC analysis is based on data generated from the distribution of (M_{11}, M_{12}, M_{13}) .

Figures 1–3 exhibit simulation results when $\rho_0 = \rho_1 = \mathbf{0}, \mathbf{0.5}$ and $\mathbf{1}$ respectively. As discussed in section “Other Types of ROC and WROC Functions”, \overline{WROC} is the conjugate partner of \overline{WROC}^* and $WROC^*$ is the conjugate partner of \overline{WROC} , and with the knowledge of $h_0(q)$ and $h_1(q)$, each of paired-partners provides the same amount of information for prediction as its partner. Choices of these weighted ROC functions should include only $WROC$ and \overline{WROC}^* so that maximization of area under curve makes sense.

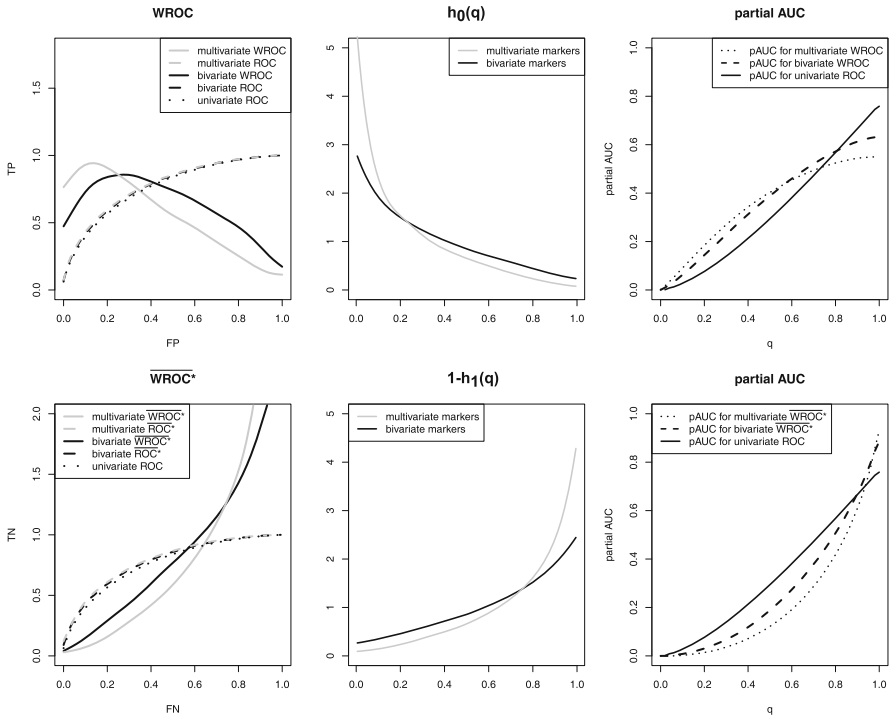


Fig. 2 Simulation for classifier $I(M_1 > m_1, M_2 > m_2, M_3 > m_3)$ with $\rho_0 = \rho_1 = 0.5$

When $\rho_0 = \rho_1 = 0$, the three markers are mutually independent, so the use of all three markers is expected to be more informative than one marker or two markers alone. Figure 1 shows a clear pattern of gain and loss as the number of markers increases. The gain in $WROC(q)$ for small values of q , when compared to univariate ROC curve, is substantial for multivariate ROC curve but only moderate for bivariate ROC curve. Similarly, the loss in $WROC(q)$ for large values of q is substantial for bivariate ROC curve but only moderate for bivariate ROC curve. This phenomenon can partly be explained by the right skewness of the weight function $h_0(q)$: the distribution of FP is uniform in univariate case, but it distributes more probability toward smaller values for bivariate marker case, and the inclusion of the third marker makes the weight function more skewed. By the equivalence between partial concordance probability and partial area under $WROC$ curve, we find that multivariate markers outperform univariate marker and bivariate marker for the region with small FP . The function \overline{WROC}^* for multivariate markers shows the opposite direction of gain and loss, compared to univariate or bivariate marker case. There is loss in $\overline{WROC}^*(q)$ for small values of q (FP) and gain for large values of q , which is due to the left skewness of the weight function $h_1(1 - q)$.

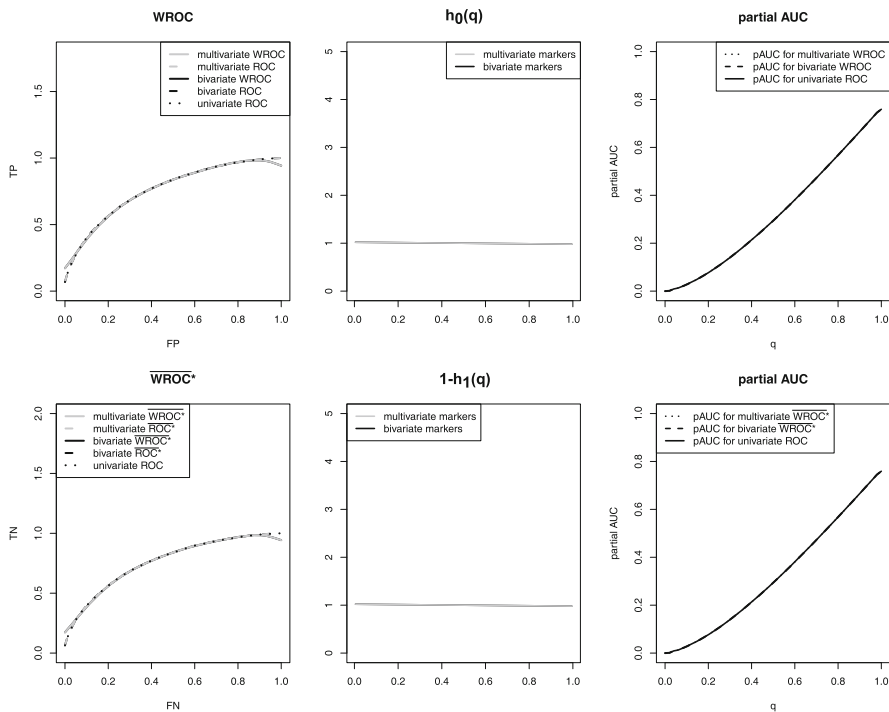


Fig. 3 Simulation for classifier $I(M_1 > m_1, M_2 > m_2, M_3 > m_3)$ with $\rho_0 = \rho_1 = 1$

When $\rho_0 = \rho_1 = 0.5$, the three markers are moderately correlated, similar to the case $\rho_0 = \rho_1 = 0$, the distribution of Q_0 and Q_1 still distribute more probability to small values, so we can observe the same pattern of tradeoff between gain at small FP and loss at large FP .

When $\rho_0 = \rho_1 = 1$, the three markers are identical and they provide the same information as one marker case (or two marker case). The ROC ($WROC$) functions for multivariate case coincides with the ROC function for univariate case (Fig. 3). The univariate case can thus be viewed as a degenerated case of multivariate markers.

A Data Example

We apply the proposed methods to the Alzheimer’s Disease Neuroimaging Initiative (ADNI) data for multivariate ROC analysis. The ADNI study is a research project with research focus on

changes of cognition, function, brain structure and function, and biomarkers in elderly controls, subjects with mild cognitive impairment, and subjects with Alzheimer’s disease

(quoted from <http://adni.loni.ucla.edu/>). The study is supported by the NIH, private pharmaceutical companies, and nonprofit organizations. Enrollment target was 800 participants – 200 normal controls, 400 patients with amnesic MCI, and 200 patients with mild AD – at 58 sites in the United States and Canada. Participants were enrolled on a rolling basis, and evaluated every six months. One of the major goals of the ADNI study is to identify biomarkers that are associated with progression from MCI to AD, and determine which biomarker measures (alone or in combination) are the best predictors of disease progression. Sensitivity and specificity for both cross-sectional and longitudinal diagnostic classification were considered important statistical techniques for assessing biomarkers in disease progression [18].

Investigations of the risk of progressing from MCI to AD dementia have largely focused on measures from the following categories: demographics, cognition, apolipoprotein E (APOE), magnetic resonance imaging (MRI), and cerebrospinal fluid (CSF) data. Demographic variables include age, education and gender. Cognitive measures represent five domains respectively: memory, language, executive function, spatial ability, and attention. Neuroimaging measures include brain volume, ventricular volume, and bilateral hippocampal volumes. The CSF variables include T-tau, A β 42, p-tau181, the ratio of the first two variables, and the ratio of the last two variables.

For this section, we selected three markers, hippocampus volume, memory score and executive function for illustration. To account for censoring, we used a reduced sample data set to create time-independent binary disease outcomes ($D = 0, 1$). We chose the 24th month as the cut-off time to define disease state. Of the 274 subjects who had complete data for the three markers, 49 subjects were loss to follow up before 24 months, so we focused on the 225 subjects who have had follow-up time longer than 24 months: there were 89 failures ($D = 1$) and 136 survivors ($D = 0$) at the 24th month. Let M_1 be hippocampus volume, M_2 be executive function score, and M_3 be memory score. Figure 4 compares the diagnostic performance of three markers (M_1, M_2, M_3), bivariate markers (M_1, M_2), and univariate marker M_1 . If the classifier is $I(M_1 > m_1, M_2 > m_2, M_3 > m_3)$, there is gain for small values of FP and loss for large values of FP . The partial AUC plot indicates that multivariate markers produce higher partial concordance summary than univariate marker when $q < 0.6$, and multivariate markers produce higher partial concordance summary than bivariate marker when $q < 0.3$. In diagnostic testing, it is crucial to maintain the false positive rate to be low to avoid unnecessary monetary costs. Thus, if the prognostic capacity is evaluated in terms of partial AUC, the multivariate marker hippocampus volume, executive function and memory score together would be considered performing much better than hippocampus volume alone.

Without restriction on the false positive rate, the AUC under the multivariate WROC curve is 0.358 (SE: 0.022) and the AUC under the multivariate \overline{WROC}^* is 0.964 (SE: 0.024); the AUC under the bivariate WROC curve is 0.437 (SE: 0.030)

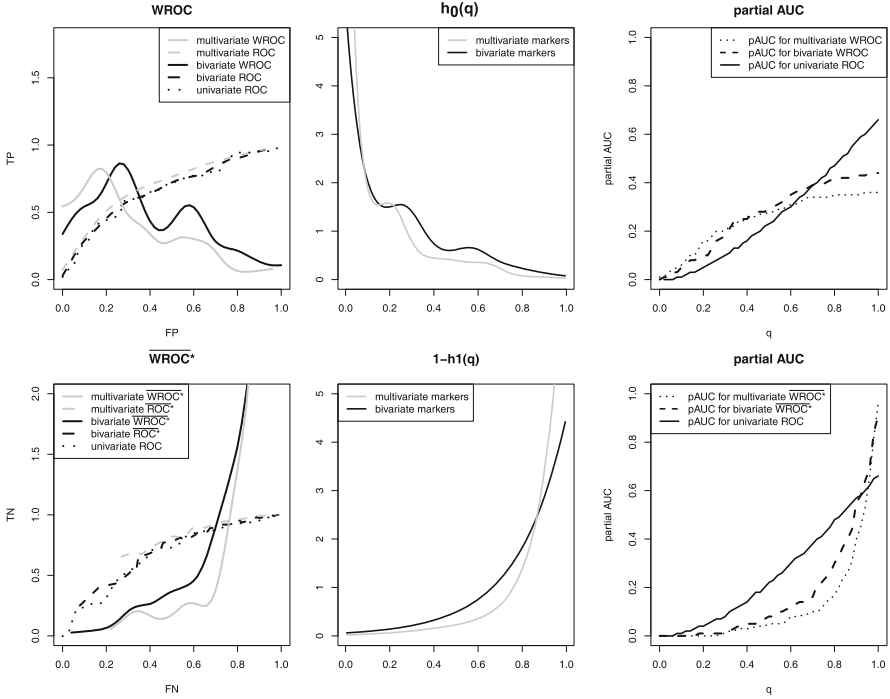


Fig. 4 $(M_1, M_2, M_3) = (\text{hippocampus, executive function, memory})$, with classifier $I(M_1 > m_1, M_2 > m_2, M_3 > m_3)$

and the AUC under the bivariate \overline{WROC}^* is 0.906 (SE: 0.030); the AUC under the univariate ROC curve is 0.658 (SE: 0.040). The bootstrap method was adopted to calculate the standard errors for estimation of AUC.

Discussion

Existing ROC methods to incorporate multiple markers typically consider a composite score based on combined markers by modeling the relationship between the marker vector \mathbf{M} and the binary outcome D [14], where $P(Y = 1|\mathbf{M}) = p(\mathbf{M})$ is used as the optimal score to identify the combination of multiple markers for classifying the disease outcome. In general, by the Neyman-Pearson lemma, the optimality of $p(\mathbf{M})$ is a very general property which holds without dimensionality constraint on \mathbf{M} . In the case that the linear logistic regression model assumption holds, the optimal classification rule, $p(\mathbf{M})$, becomes equivalent to the regression function $\beta\mathbf{M}$ under the logit link. Thus, the optimality property of a one-dimensional

classification score heavily relies on the assumption of logistic regression model. In this paper, we extend tools from univariate marker to multivariate markers for evaluating predictive accuracy of markers under a nonparametric setting based on tree-based classification rules.

The proposed ROC and WROC functions together with the AUC are intended to measure the average performance of and-or classifier among all possible combinations of true positive rate for a given false positive rate for evaluating predictability of markers and comparing curves, and they may not reflect the optimized use of markers for clinical decisions. Although the proposed approach is not designed to achieve optimality as a decision rule such as the one proposed by Jin and Lu [13], our methods and inferential results are much more structural, accessible and workable. The proposed ROC and WROC functions enjoy the advantage of preserving the distributional structures of markers, and the associated summary measures such as AUC or partial AUC serve as very appropriate summary measures to evaluate the performance of and-or classifier among all possible combinations of marker values – this is a feature similar to the univariate marker case. These summary measures are useful in applications, since many biomarker studies (such as the ADNI study and two other Alzheimer’s Disease studies that the authors are currently involved) have research emphasis largely focused on the understanding of predictability of biomarkers in target population, and less emphasis toward optimization of clinical decision rules.

The evaluation takes into account the distributions of quantile variables Q_0 and Q_1 in the diseased and non-diseased populations, which leads to the result of equivalence between AUC and CON, a property similar to the case of univariate marker. We also provide estimation procedures using nonparametric smoothing estimators for the ROC and WROC function, and U-statistic for the AUC. For applications of the proposed analysis, as the ‘curse of dimensionality’ is not a concern for nonparametric estimation of ROC, WROC and other related properties, the usual random split into training sample (for model fitting) and test sample (for creating ROC curve and calculating AUC) would be as proper as it is for univariate marker case, and therefore is advisable.

For future and further research, similar to the considerations for univariate ROC analysis [16, 20], it would be interesting to consider methodology to adjust for covariates such as age, sex or other demographical factors for bivariate or multivariate markers.

Also, given that the disease outcomes typically change with time, it would be interesting to extend the ROC analysis for high-dimensional markers to accommodate time-to-disease information using the ‘survival-tree methodology’ [22], along the lines of extending ROC techniques from binary disease outcome model to right-censored survival data model in univariate marker settings [5, 10, 11, 19].

Appendix

Proof of Theorem 1. Define the kernel function of the U-statistic [12] as

$$h(\mathbf{M}_{0i}, \mathbf{M}_{1j}; FP, TP) = I(\mathbf{M}_{1j} \in D(\mathbf{M}_{0i})) \cdot I(FP(\mathbf{M}_{0i}) \leq p, TP(\mathbf{M}_{1j}) > q).$$

Note that

$$\begin{aligned} \widehat{AUC}(p, q) &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} h(\mathbf{M}_{0i}, \mathbf{M}_{1j}; \widehat{FP}, \widehat{TP}) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} h(\mathbf{M}_{0i}, \mathbf{M}_{1j}; FP, TP) \\ &\quad + \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \{h(\mathbf{M}_{0i}, \mathbf{M}_{1j}; \widehat{FP}, \widehat{TP}) - h(\mathbf{M}_{0i}, \mathbf{M}_{1j}; FP, TP)\} \\ &= I + II \end{aligned}$$

The kernel function in Term I satisfies $E[h^2] < \infty$ and by two-sample U-statistics theory, I converges to $AUC(p, q)$ in probability. Term II can be expressed as

$$II = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(\mathbf{M}_{1j} \in D(\mathbf{M}_{0i})) \{I(\widehat{FP}(\mathbf{M}_{0i}) \leq p, \widehat{TP}(\mathbf{M}_{1j}) > q) - I(FP(\mathbf{M}_{0i}) \leq p, TP(\mathbf{M}_{1j}) > q)\}$$

Note that

$$\begin{aligned} |II| &\leq \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left| I(\widehat{FP}(\mathbf{M}_{0i}) \leq p, \widehat{TP}(\mathbf{M}_{1j}) > q) - I(FP(\mathbf{M}_{0i}) \leq p, TP(\mathbf{M}_{1j}) > q) \right| \\ &\leq \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left| I(\widehat{FP}(\mathbf{M}_{0i}) \leq p) - I(FP(\mathbf{M}_{0i}) \leq p) \right| + \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left| I(\widehat{TP}(\mathbf{M}_{1j}) > q) - I(TP(\mathbf{M}_{1j}) > q) \right| \\ &= \frac{1}{n_0} \sum_{i=1}^{n_0} \left| I(\widehat{FP}(\mathbf{M}_{0i}) \leq p) - I(FP(\mathbf{M}_{0i}) \leq p) \right| + \frac{1}{n_1} \sum_{j=1}^{n_1} \left| I(\widehat{TP}(\mathbf{M}_{1j}) > q) - I(TP(\mathbf{M}_{1j}) > q) \right| \\ &= o_p(n_0^{-1/2}) + o_p(n_1^{-1/2}) = o_p(N^{-1/2}) \end{aligned}$$

The consistency result, (i), in Theorem 1 follows by viewing the fact that term II converges to 0 in probability. To prove (ii), first note that Term I converges in distribution to a normal distribution by U-statistics theory: $\sqrt{N}\{I - AUC(p, q)\} \xrightarrow{d} \text{Normal}(0, \sigma^2)$, where $\sigma^2 = \lambda^{-1} \tau_{1,0} + (1 - \lambda)^{-1} \tau_{0,1}$ with

$$\tau_{1,0} = \text{COV}[h(\mathbf{M}_{01}, \mathbf{M}_{11}), h(\mathbf{M}_{01}, \mathbf{M}_{12})]$$

and

$$\tau_{0,1} = \text{COV}[h(\mathbf{M}_{01}, \mathbf{M}_{11}), h(\mathbf{M}_{02}, \mathbf{M}_{11})].$$

Also,

$$\begin{aligned} \sqrt{N}\{\widehat{AUC}(p, q) - AUC(p, q)\} &= \sqrt{N}\{I - AUC(p, q)\} + \sqrt{N} \cdot II \\ &= \sqrt{N}\{I - AUC(p, q)\} + o_p(1) \xrightarrow{d} \text{Normal}(0, \sigma^2) \end{aligned}$$

Property in section “Other Types of ROC and WROC Functions”.

- (i) $ROC(q) + \overline{ROC}(1 - q) = 1$, and $WROC(q) + \overline{WROC}(1 - q) = h_0(q)$
- (ii) $ROC^*(q) + \overline{ROC}^*(1 - q) = 1$, and $WROC^*(q) + \overline{WROC}^*(1 - q) = h_1(q)$

Proof. Note that

$$\begin{aligned} ROC(q) + \overline{ROC}(1 - q) &= E[TP(\mathbf{M}_0)|FP(\mathbf{M}_0) = q] + E[FN(\mathbf{M}_0)|FP(\mathbf{M}_0) = q] \\ &= E[TP(\mathbf{M}_0) + FN(\mathbf{M}_0)|FP(\mathbf{M}_0) = q] \\ &= E[1 |FP(\mathbf{M}_0) = q] = 1, \end{aligned}$$

and it follows $WROC(q) + \overline{WROC}(1 - q) = ROC(q) \cdot h_0(q) + \overline{ROC}(1 - q)h_0(q) = h_0(q)$, which proved (i). Similar argument can be used to prove (ii).

References

1. Baker, S.G.: Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082–1087 (2000)
2. Baker, S.G., Cook, N.R., Vickers, A., Kramer, B.S.: Using relative utility curves to evaluate risk prediction. *J. R. Stat. Soc. A* **172**, 729–748 (2009)
3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey (1984)
4. Dodd, L., Pepe, M.S.: Partial AUC estimation and regression. *Biometrics* **59**, 614–623 (2003)
5. Etzioni, R., Pepe, M.S., Longton, G., Hu, C., Goodman, G.: Incorporating the time dimension in receiver operating characteristic curves: a prostate cancer case study. *Med. Decis. Making* **19**, 242–251 (1999)
6. Etzioni, R., Kooperberg, C., Pepe, M.S., Smith, R., GANN, P.H.: Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* **4**, 523–538 (2003)
7. Green, P.J., Silverman, B.W.: *Nonparametric Regression and Generalized Linear Models: A Robust Penalty Approach*. Chapman and Hall, London (1994)
8. Gu, W., Pepe, M.S.: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **5**(1), (2009)
9. Hanley, J.A., McNeil, B.J.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983)
10. Heagerty, P.J., Zheng, Y.: Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105 (2005)
11. Heagerty, P.J., Lumley, T., Pepe, M.S.: Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344 (2000)
12. Hoeffding, W.: A class of statistics with asymptotically normal distributions. *Ann. Stat.* **19**, 293–325 (1948)
13. Jin, H., Lu, Y.: ROC region of a regression tree. *Stat. Probab. Lett.* **79**, 936–942 (2009)

14. McIntosh, M.W., Pepe, M.S.: Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657–664 (2002)
15. Metz, C.E., Shen, J.H.: Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Med. Decis. Making* **12**, 60–75 (1992)
16. Pepe, M.S.: Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135 (1998)
17. Pepe, M.S.: *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, New York (2003)
18. Risacher, S.L., Saykin, A.J., West, J.D., Shen, L., Firpi, H.A., McDonald, B.C., Alzheimer's Disease Neuroimaging Initiative (ADNI): baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* **6**, 347–361 (2009)
19. Slate, E.H., Turnbull, B.W.: Models for longitudinal biomarkers of disease onset. *Stat. Med.* **19**, 617–637 (2000)
20. Tosteson, A.N., Begg, C.B.: A general regression methodology for ROC curve estimation. *Med. Decis. Making* **8**, 205–215 (1988)
21. Wang M.-C., Li, S.: Bivariate marker measurements and ROC analysis. *Biometrics* **68**, 1207–1218 (2012)
22. Zhang, H., Crowley, J., Sox, H.C., Olshen, R.A.: Tree-structured statistical methods. *Encycl. Biostat.* **6**, 4561–4573 (1998)

Assessing Discrimination of Risk Prediction Rules in a Clustered Data Setting

Bernard Rosner, Weiliang Qiu, and Mei-Ling Ting Lee

Abstract The AUC (area under ROC curve) is a commonly used metric to assess discrimination of risk prediction rules; however, standard errors of AUC are usually based on the Mann-Whitney U test that assumes independence of sampling units. For ophthalmologic applications, it is desirable to assess risk prediction rules based on eye-specific outcome variables which are generally highly, but not perfectly correlated in fellow eyes [e.g. progression of individual eyes to age-related macular degeneration (AMD)]. In this article, we use the extended Mann-Whitney U test (Rosner and Glynn, *Biometrics* 65:188–197, 2009) for the case where subunits within a cluster may have different progression status and assess discrimination of different prediction rules in this setting. Both data analyses based on progression of AMD and simulation studies show reasonable accuracy of this extended Mann-Whitney U test to assess discrimination of eye-specific risk prediction rules.

Introduction

Age-related macular degeneration (AMD) is the leading cause of irreversible visual impairment and blindness in the United States and other developed countries throughout the world [5]. Evidence is accumulating regarding modifiable factors that may decrease the risk of progression to the advanced forms of AMD.

The paper appeared in volume 19 (2013) of *Lifetime Data Analysis*.

B. Rosner (✉) • W. Qiu

Channing Division of Network Medicine, Brigham and Women's Hospital/Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA

e-mail: stbar@channing.harvard.edu; stwxq@channing.harvard.edu

M.-L.T. Lee

Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA

e-mail: mltlee@umd.edu

The Progression of Age-Related Macular Degeneration Study is a longitudinal study designed to measure multiple risk factors for the onset and progression of AMD. A total of 261 individuals were included in the analyses. Details about the study are provided in Seddon et al. [12]. The average age of the subjects was 72.3 years ($sd = 6.1$, $range = [60, 87]$). Approximately 61% were female. Subjects were followed for 4–6 years. We are interested in assessing whether the prediction of progression to the advanced form of AMD would be improved if we include total fat intake in the prediction model in addition to other AMD risk factors.

A common criterion for assessing discrimination of risk prediction rules is the improvement in the area under the receiver-operating-characteristic curve (AUC) [6, 14]. By using the equivalence between the Mann-Whitney U statistic and AUC [1], one can assess discrimination of risk prediction rules via the Mann-Whitney U test, which is commonly used in nonparametric two-group comparisons when the normality of the underlying distribution is questionable. However, an assumption of the Mann-Whitney U test is that sampling units (e.g. eyes) are independent. For the AMD data set, progression status for the right and left eyes for a subject are correlated.

Moreover, previous works on non-parametric two-sample comparisons mainly focus on the shift alternative in the original data space [2, 3]. However, (a) the meaning of the shift might be different for each underlying null distribution, and (b) the underlying null distribution is usually unknown. Rosner and Glynn [9] proposed a shift alternative in a transformed data space based on the probit transformation and applied their method to assess and compare discrimination of risk prediction rules for a case-control AMD dataset where a subject is a case if either eye has AMD and a control if neither eye has AMD. However, their work requires the assumption that the observations are independent, which is not applicable to clustered data, where the eye is the unit of analysis.

Rosner et al. [10] extended the Wilcoxon rank sum test, which is equivalent to the Mann-Whitney U test, for clustered data for two group comparisons where group membership is defined at the subunit level. In our example, there may be patients where one eye has progressed, while the fellow eye has not progressed.

In this article, we define the extended Mann-Whitney U statistic, which is equivalent to Rosner et al.'s [10] extended Wilcoxon rank sum statistic, and derive its variance under the transformed shift alternative to assess discrimination of risk prediction rules in a clustered data setting with group membership defined at the subunit level. We also derive the variance of the difference of two extended Mann-Whitney U statistics to compare discrimination of different risk prediction rules when applied to the same data set.

The structure of the remaining parts of this article is as follows. In section “Extended Mann-Whitney U Statistic for Clustered Data”, we define the extended Mann-Whitney U statistic and derive its variance. In section “Variance of the Difference of Two Extended Mann-Whitney U Statistics Applied to the Same Data Set”, we derive the variance of the difference of two extended Mann-Whitney U statistics for two prediction rules when applied to the same data set. In section “Multiple Imputation” approaches are described to incorporate

uncertainty in the estimation of regression coefficients used in the prediction rule. In section “Example”, we apply the extended Mann-Whitney U statistic to the AMD data set described in section “Introduction” to assess and compare discrimination of different risk prediction models for a prospective study of age-related macular degeneration (AMD), an important eye disease among the elderly resulting in substantial losses of vision. In section “A Simulation Study”, we conduct a simulation study to evaluate the performance of the extended Mann-Whitney U statistic in terms of bias of the estimator and its variance estimate, coverage probability and power. Section “Discussion” is a discussion. Technique details are shown in the Appendices.

Extended Mann-Whitney U Statistic for Clustered Data

Suppose that there are N independent clusters, where the subunit is the unit of analysis and the i -th cluster has g_i subunits, $i = 1, \dots, N$. We can use the following generalized estimating equations (GEE) model with an exchangeable correlation structure to fit the AMD data since each eye has a progression score and these two scores for a subject are correlated:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_k x_{ijk}, i = 1, \dots, N, j = 1, 2, \quad (1)$$

where $x_{ij\ell}$ is the value of the ℓ -th risk factor for the j -th eye of the i -th subject, $\ell = 1, \dots, k$, $p_{ij} = Pr(\text{progression for the } j\text{-th eye of the } i\text{-th subject} | x_{ij1}, \dots, x_{ijk})$, N is the number of subjects, and k is the number of risk factors. In our example, the initial AMD grade is an eye-specific risk factor, while other risk factors are person-specific.

Denote Z_{ij} as the prediction score for the j -th subunit of the i -th cluster based on the GEE Model (1) given by

$$Z_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \dots + \hat{\beta}_k x_{ijk},$$

where $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are estimated via GEE Model (1). We are interested in testing if the distribution of prediction scores among subunits that progress is the same as that among subunits that do not progress.

To incorporate the information that subunits within a cluster are highly correlated, and that subunits within a cluster might have different progression status, we propose the following extended Mann-Whitney U statistic:

$$\hat{\eta}_c = \frac{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell})(1 - \delta_{ij})\delta_{k\ell}}{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} (1 - \delta_{ij})\delta_{k\ell}}, \quad (2)$$

where $U(Z_{ij} - Z_{k\ell}) = 1$ if $Z_{ij} < Z_{k\ell}$, $= 1/2$ if $Z_{ij} = Z_{k\ell}$, and $= 0$ otherwise, for $i \neq k$, or $j \neq \ell$, and $\delta_{ij} = 1$ if the j -th subunit of the i -th cluster has progressed, $= 0$ if it has not progressed.

The statistic $\hat{\eta}_c$ is the proportion of pairs of subunits where the subunit that did not progress has a lower score than the subunit that did progress. If the proportion is equal to $1/2$, then there is no difference between the location parameter for subunits that progress versus subunits that do not progress. If the proportion is much greater than or smaller than $1/2$, then there exist evidence that the location parameters for the two groups of subunits are different.

We assume that (1) clusters are independent of each other; (2) δ_{ij} are fixed; (3) $Pr(Z_{ij} = Z_{k\ell}) = 0$ for $i \neq k$ or $j \neq \ell$. Furthermore, by definition the probit transformation $H = \Phi^{-1}$ will transform the response variable to a normal distribution. Let

$$H_{ij} \equiv H(F(Z_{ij})), \tag{3}$$

where F is the cumulative distribution of Z_{ij} . In data analysis, F will be estimated by the empirical cumulative distribution \hat{F}_n . We assume that

$$H_{ij} \sim \begin{cases} N(0, 1) & \text{if } \delta_{ij} = 0, \\ N(\mu, 1), \mu \neq 0 & \text{if } \delta_{ij} = 1, \end{cases}$$

and that after the transformation H , the bivariate random vector

$$\begin{pmatrix} H_{i_1 j_1} - H_{k_1 \ell_1} \\ H_{i_2 j_2} - H_{k_2 \ell_2} \end{pmatrix}$$

is bivariate normally distributed for any $i_1, k_1, i_2, k_2, j_1, \ell_1, j_2$, and ℓ_2 , with covariance matrix

$$\text{Cov} \begin{pmatrix} H_{ij} \\ H_{i\ell} \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, j \neq \ell, \text{ and } \text{Cov} \begin{pmatrix} H_{ij} \\ H_{k\ell} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, i \neq k.$$

We are interested in testing the hypotheses:

$$\begin{aligned} H_0 : & H(X) = H(Y) = H(Y_c) \text{ versus} \\ H_a : & H(Y) = H(Y_c) + \mu, \mu \neq 0, \end{aligned}$$

where X is the prediction score of a randomly selected non-progressing subunit, Y is the prediction score of a randomly selected progressing subunit, and Y_c is the counterfactual random variable obtained if each subunit that had progressed actually had not progressed. We refer to H_a as a *probit-shift alternative*. The probit-shift alternative is useful in calculating $\text{Var}(\hat{\eta}_c)$ under H_a in closed-form, which we

will need to (a) obtain confidence limits for η_c ; (b) compare η_c between two risk prediction rules assessed on the same subjects; and (c) compute power for future studies.

We can obtain the expected value of the extended Mann-Whitney U statistic $\hat{\eta}_c$ under the alternative hypothesis:

$$E(\hat{\eta}_c|H_a) = \frac{\theta_c \sum_{i=1}^N c_i d_i + \theta [C \cdot D - \sum_{i=1}^N c_i d_i]}{C \cdot D}, \tag{4}$$

where

$$\theta_c = \Phi\left(\frac{\mu}{\sqrt{2(1-\rho)}}\right), \theta = \Phi\left(\frac{\mu}{\sqrt{2}}\right),$$

and $c_k = \sum_{\ell=1}^{g_k} \delta_{k\ell}$ is the number of progressing subunits for the k -th subject, $d_i = \sum_{j=1}^{g_i} (1 - \delta_{ij})$ is the number of non-progressing subunits for the i -th subject, $C = \sum_{k=1}^N c_k$, $D = \sum_{i=1}^N d_i$. Note that for clustered data, there are two θ 's, whereas in non-clustered data there is only a single θ . θ_c is used for comparison of nonprogressing and progressing subunits (e.g. eyes) within the same cluster (e.g. subject), and θ for comparison of nonprogressing and progressing subunits (e.g. eyes) in different clusters (e.g. different subjects). Under the null hypothesis, i.e., $\mu = 0$, $\theta_c = \theta = 1/2$, and $E(\hat{\eta}_c|H_0) = 1/2$.

To derive the variance of $\hat{\eta}_c$, we rewrite $\hat{\eta}_c$ as

$$\hat{\eta}_c = \frac{A + B}{C \cdot D} \tag{5}$$

where

$$A = \sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} U(Z_{ij} - Z_{i\ell}) (1 - \delta_{ij}) \delta_{i\ell}$$

$$B = \sum_{i=1}^N \sum_{k=1, k \neq i}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell}$$

Then

$$\text{Var}(\hat{\eta}_c|H_a) = \frac{1}{C^2 \cdot D^2} [\text{Var}(A|H_a) + \text{Var}(B|H_a) + 2\text{Cov}(A, B|H_a)]. \tag{6}$$

We can use second order moments of the bivariate normal distribution to derive a closed-form expression for $\text{Var}(\hat{\eta}_c|H_a)$. The complete derivation of $\text{Var}(\hat{\eta}_c|H_a)$ is given in Appendix 1.

Variance of the Difference of Two Extended Mann-Whitney U Statistics Applied to the Same Data Set

Suppose we have two prediction rules (indexed by $t = 1, 2$) applied to the same data set. We can compare the area under ROC curves (AUCs) of two prediction rules by testing the hypothesis $H_0: \eta_c^{(1)} = \eta_c^{(2)}$ versus $H_a: \eta_c^{(1)} \neq \eta_c^{(2)}$, where $\eta_c^{(t)} = \text{AUC}$ for the t -th prediction rule, $t = 1, 2$. We estimate $\eta_c^{(t)}$ by:

$$\hat{\eta}_c^{(t)} = \frac{A^{(t)} + B^{(t)}}{C \cdot D}, t = 1, 2,$$

and

$$A^{(t)} = \sum_{i=k=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} (1 - \delta_{ij}) \delta_{i\ell} U \left(Z_{ij}^{(t)} - Z_{i\ell}^{(t)} \right)$$

and

$$B^{(t)} = \sum_{i=1}^N \sum_{k=1, k \neq i}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_k} (1 - \delta_{ij}) \delta_{k\ell} U \left(Z_{ij}^{(t)} - Z_{k\ell}^{(t)} \right).$$

Since the two prediction rules are applied to the same data set, the two extended Mann-Whitney U statistics are correlated. To calculate the variance of $\hat{\eta}_{c1} - \hat{\eta}_{c2}$, we assume the following correlation structure:

$$\begin{aligned} \rho &= \text{Cov} \left(H_{ij}^{(1)}, H_{ij}^{(2)} \right) = \text{Corr} \left(H_{ij}^{(1)}, H_{ij}^{(2)} \right), \\ \rho_{11} &= \text{Cov} \left(H_{ij_1}^{(1)}, H_{ij_2}^{(1)} \right) = \text{Corr} \left(H_{ij_1}^{(1)}, H_{ij_2}^{(1)} \right), \\ \rho_{22} &= \text{Cov} \left(H_{ij_1}^{(2)}, H_{ij_2}^{(2)} \right) = \text{Corr} \left(H_{ij_1}^{(2)}, H_{ij_2}^{(2)} \right), \\ \rho_{12} &= \text{Cov} \left(H_{ij_1}^{(1)}, H_{ij_2}^{(2)} \right) = \text{Corr} \left(H_{ij_1}^{(1)}, H_{ij_2}^{(2)} \right). \end{aligned}$$

We denote

$$\theta_{c_t} = \Phi \left(\frac{\mu_t}{\sqrt{2(1 - \rho_{tt})}} \right), \theta_t = \Phi \left(\frac{\mu_t}{\sqrt{2}} \right), t = 1, 2.$$

We wish to calculate

$$\text{Var} \left(\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)} \right) = \text{Var} \left(\hat{\eta}_c^{(1)} \right) + \text{Var} \left(\hat{\eta}_c^{(2)} \right) - 2\text{Cov} \left(\hat{\eta}_c^{(1)}, \hat{\eta}_c^{(2)} \right). \quad (7)$$

We already derived the formula for $\text{Var}(\hat{\eta}_c^{(t)})$ in Eq. 6. Hence, we only need to derive $\text{Cov}(\hat{\eta}_c^{(1)}, \hat{\eta}_c^{(2)})$. We can decompose it into 4 parts:

$$\begin{aligned} \text{Cov}(\hat{\eta}_c^{(1)}, \hat{\eta}_c^{(2)}) &= \frac{1}{C^2 D^2} \left[\text{Cov}(A^{(1)}, A^{(2)}) + \text{Cov}(A^{(1)}, B^{(2)}) \right. \\ &\quad \left. + \text{Cov}(B^{(1)}, A^{(2)}) + \text{Cov}(B^{(1)}, B^{(2)}) \right] \end{aligned}$$

The derivation of $\text{Var}(\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)} | H_a)$ in Eq. 7 is given in Appendix 2.

Hence, a large sample test statistic to test the hypotheses $H_0 : \eta_c^{(1)} = \eta_c^{(2)}$ versus $H_a : \eta_c^{(1)} \neq \eta_c^{(2)}$ is given by $Z_{12} = (\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}) / [\widehat{\text{Var}}(\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)})]^{1/2} \sim N(0, 1)$ under H_0 .

Finally, we can calculate the power of the test $H_0 : \Delta = 0$ vs $H_a : \Delta = \Delta_0$, where $\Delta_0 = \eta_c^{(1)} - \eta_c^{(2)}$ using the formula

$$\text{Power} = 1 - \Phi\left(\frac{\sigma_0 Z_{\alpha/2} - \Delta_0}{\sigma_1}\right) + \Phi\left(\frac{-\sigma_0 Z_{\alpha/2} - \Delta_0}{\sigma_1}\right) \quad (8)$$

where $Z_{\alpha/2}$ is the upper $100\alpha/2\%$ percentile of the standard normal distribution $N(0, 1)$, $\sigma_0 = \sqrt{\text{Var}(\hat{\Delta} | H_0)}$ and $\sigma_1 = \sqrt{\text{Var}(\hat{\Delta} | H_a)}$.

Multiple Imputation

When constructing scores, we used estimated regression coefficients from GEE. Hence, we did not account for the variation of these estimates in the calculation of the variance of $\hat{\eta}_c$.

One remedy is to use multiple imputation. Specifically, we randomly generate m random vectors β from the multivariate normal distribution $N(\hat{\beta}, \hat{\Sigma})$, where $\hat{\beta}$ and $\hat{\Sigma}$ are the estimated regression coefficients and their variance-covariance matrix.

Then we obtain m estimates of η_c . Denote these as $\hat{\eta}_{c,i}$, $i = 1, \dots, m$. We then use

$$\hat{\eta}_c^* = \frac{1}{m} \sum_{i=1}^m \hat{\eta}_{c,i}$$

as a final estimate of η_c , which will reflect both between- and within-imputation variance.

We next calculate the variance $\text{Var}(\hat{\eta}_c^*)$ by the following formula [11]

$$\text{Var}(\hat{\eta}_c^*) = \frac{1}{m} \sum_{i=1}^m \text{Var}(\hat{\eta}_{c,i}) + \left(\frac{1 + \frac{1}{m}}{m-1} \right) \sum_{i=1}^m (\hat{\eta}_{c,i} - \hat{\eta}_c^*)^2,$$

where $\text{Var}(\hat{\eta}_{c,i})$ is calculated using Eq. 6.

We will calculate the p-value for a parameter estimate based on multiple imputation by using the method mentioned in Rubin [11] (cf. Appendix 3).

We can obtain the adjusted estimate $\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}$ and its variance and confidence interval using the same approach.

Example

In this section, we apply the proposed method to the AMD data set (with $n = 261$ subjects) mentioned in section “Introduction”. The data set is from the prospective longitudinal study of AMD: the Progression of Age-Related Macular Degeneration Study designed to measure multiple risk factors for the onset and progression of AMD. One modifiable factor is dietary total fat intake. The results of Seddon et al. [12] show that higher total fat intake is associated with an increased risk of progression to the advanced forms of AMD, with an odds ratio (OR) of 2.90 (95% confidence interval, 1.15–7.32) for the highest fat-intake quartile relative to the lowest fat-intake quartile, after controlling for other factors (P trend = 0.01).

The increased risk of the complications of AMD are indicated by drusen, which are yellow deposits under the retina [5]. According to Seddon et al. [12], eyes with extensive small drusen (≥ 15 drusen; with size of drusen $< 63 \mu\text{m}$), nonextensive intermediate drusen (< 20 drusen; with size of drusen $\geq 63 \mu\text{m}$ but $< 125 \mu\text{m}$), or pigment abnormalities associated with AMD were assigned a grade of 2. Eyes with extensive intermediate or large drusen (size of drusen $\geq 125 \mu\text{m}$) were assigned a grade of 3. Eyes with geographic atrophy received a grade of 4. If there was evidence of retinal pigment epithelial detachment or choroidal neovascular membrane, a grade of 5 was assigned. Eyes received a grade of 1 if none of these signs was present. All eyes in our analysis had a grade of ≥ 2 and ≤ 4 at baseline. Each subject contributed two eyes to the analysis. Advanced AMD is defined as grades 4 or 5.

Progression to advanced AMD in an eye over 4–6 years was defined either as progression from a grade of less than 4 at baseline to grades 4 or 5 at any follow-up visit, or progression from grade 4 at baseline to grade 5 at any follow-up visit.

If a set of risk factors has no predictive ability to distinguish progressing versus non-progressing eyes, the value of the extended Mann-Whitney U statistic

$$\hat{\eta}_c = \frac{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell}}{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} (1 - \delta_{ij}) \delta_{k\ell}},$$

will be close to the expected null value $E(\hat{\eta}_c|H_0) = 1/2$. In general, we will use $\hat{\eta}_c$ to assess and compare the discrimination of different risk prediction rules.

If we use the same dataset to both construct a prediction rule and calculate prediction accuracy, the prediction accuracy might be over-estimated. Hence, we randomly split the 261 subjects into two groups. One group with 131 subjects is used as the training set to obtain the estimated regression coefficients $\hat{\beta}$ in the GEE model. The other group with 130 subjects is used as the testing set to obtain the estimates $\hat{\eta}_c^{(1)}$, $\hat{\eta}_c^{(2)}$, and related estimates and p-values. When constructing prediction scores for subjects in the testing set, we used the regression coefficients estimated from the training set. To account for the variability of these estimated regression coefficients, we used the multiple imputation approach mentioned in section “Multiple Imputation”. Hence, for each imputation, we obtain different estimates of η_c in the testing set and use multiple imputation approaches to obtain an overall estimate of η_c that reflects both between and within imputation variation.

Table 1 shows the parameter estimates of two prediction rules based on risk factors: BMI, age, gender, cardiovascular disease (cvd), systolic blood pressure (sbp), current and past smoking (smkcur and smkpst), ln (caloric intake) (lncalor), beta carotene intake (adjbcaro), alcohol intake (alco), initial eye grade (inieye3 and inieye4), protein intake (rprot2, rprot3, and rprot4), and total fat intake (rtotfat2, rtotfat3, and rtotfat4).

The regression coefficients and their standard errors, p-values, and 95% confidence intervals (CIs) were based on the training set (with $n = 131$ subjects). The estimates $\hat{\eta}_c^{(1)}$ and $\hat{\eta}_c^{(2)}$ and their standard errors and p-values were estimated based on the testing set (with $n = 130$ subjects).

In the training set, there are 131 subjects (262 eyes). The number of eyes that progressed was 51. There are 9 subjects where both eyes progressed, 17 subjects where the right eye progressed and the left eye did not progress, 16 subjects where the right eye did not progress and the left eye progressed, and 89 subjects where both eyes did not progress. The odds ratio for progression between fellow eyes is $9 * 89 / (17 * 16) = 2.94$. The correlation between risk scores for fellow eyes was 0.53 in the training set and 0.51 in the testing set.

Table 1 shows that there were significant effects of total fat intake in model 1. The odds ratios and 95% CIs of *rtotfat2* (second quartile), *rtotfat3* (third quartile), and *rtotfat4* (fourth quartile) relative to *rtotfat1* (first quartile) are 1.7 (95% CI: [0.4, 6.7], p-value = 0.48), 4.3 (95% CI: [0.9, 20.1], p-value = 0.06), and 22.4 (95% CI: [4.2, 118.6], p-value < 0.01). However, there was no significant difference between the AUCs for the two prediction rules (multiple-imputation-based estimates: $\hat{\mu}^{(1)} = 0.406$, $\hat{\mu}^{(2)} = 0.381$, $\hat{\theta}^{(1)} = 0.613$, $\hat{\theta}_c^{(1)} = 0.718$, $\hat{\theta}^{(2)} = 0.606$, $\hat{\theta}_c^{(2)} = 0.650$, $AUC_1 = 0.677$ versus $AUC_2 = 0.672$, p-value = 0.93).

Table 1 Comparison of the two prediction rules for the AMD data set based on GEE (number of subjects = 131, number of eyes = 262, number of eyes which progressed = 51)

Training set variable	Prediction rule 1 (model with total fat intake)					Prediction rule 2 (model without total fat intake)		
	Estimate	SE	pval	OR	[95% CI]	Estimate	SE	pval
(Intercept)	-7.95	1.58				-6.15	1.42	
bmi2529	1.72	0.65	0.01			1.72	0.55	0.00
bmi30+	1.86	0.69	0.01			1.43	0.59	0.02
male6069	-0.61	0.80	0.44			-0.15	0.82	0.86
male7079	1.42	0.86	0.10			1.41	0.79	0.07
male80+	0.64	1.18	0.59			0.50	1.22	0.68
feml7079	1.74	0.86	0.04			1.21	0.77	0.11
feml80+	0.82	1.16	0.48			0.27	0.97	0.78
cvd	0.97	0.54	0.07			0.73	0.48	0.13
sysbp _c	-0.02	0.01	0.15			-0.02	0.01	0.14
smkcur	1.58	0.77	0.04			1.46	0.69	0.04
smkpst	0.14	0.50	0.78			0.13	0.50	0.79
ln calor _c	-3.95	1.18	0.00			-1.37	1.00	0.17
adjbcaro _c	0.43	0.27	0.11			0.28	0.31	0.37
alco _c	0.01	0.01	0.42			0.00	0.02	0.92
inিয়ে3	3.15	0.67	0.00			3.19	0.64	0.00
inিয়ে4	2.06	0.67	0.00			2.27	0.73	0.00
rprot2	0.33	0.51	0.52			0.12	0.51	0.81
rprot3	0.53	0.75	0.48			0.54	0.78	0.49
rprot4	-0.51	1.01	0.61			-0.50	0.92	0.58
rtotfat2	0.51	0.71	0.48	1.7	[0.4, 6.7]			
rtotfat3	1.47	0.78	0.06	4.3	[0.9, 20.1]			
rtotfat4	3.11	0.85	0.00	22.4	[4.2, 118.6]			
Testing set	$\hat{\eta}_c^{(1)} = 0.677(sd = 0.050, df = 48.14)$					$\hat{\eta}_c^{(2)} = 0.672(sd = 0.049, df = 38.13)$		
	$\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)} = 0.005(sd = 0.057, df = 7.57), p\text{-value} = 0.93$							
$corr(Z_{t1}, Z_{t2})^a$	0.53					0.51		

bmi2529 = BMI between 25 and 29.9; bmi30+ = BMI greater or equal to 30; male6069 = male aged between 60 and 69; male7079 = male aged between 70 and 79; male80+ = male aged greater than or equal to 80; feml7079 = female aged between 70 and 79; feml80+ = female aged greater than or equal to 80; cvd = cardiovascular disease; sysbp_c = mean-centered systolic blood pressure (mean = 140.17); smkcur = current smoker; smkpast = past smoker; ln calor_c = mean-centered log caloric intake (mean = 7.25); adjbcaro_c = mean-centered calorie-adjusted beta carotene intake ($\mu g/d$, values are expressed as geometric mean after sex-specific energy adjustment) (mean = 8.11); alco_c = mean-centered alcohol intake (g/d), (mean = 6.77); inিয়ে3 = 1 if initial eye grade = 3, = 0 otherwise; inিয়ে4 = 1 if initial eye grade = 4, = 0 otherwise; rprot2 = protein intake (second quartile); rprot3 = protein intake (third quartile); rprot4 = protein intake (fourth quartile); rtotfat2 = total fat intake (second quartile); rtotfat3 = total fat intake (third quartile); rtotfat4 = total fat intake (fourth quartile). The quartiles of protein and total fat intake are sex-specific.

^aEstimated correlation between the scores of the two eyes for the same subject after adjusting for the above covariates. based on training set

A Simulation Study

We conducted a simulation study to evaluate the performance of the extended Mann-Whitney U statistic on comparing two prediction rules. The values of the model parameters were set to be the estimated parameters for the AMD data set (see section “Example”). We assume that all subjects have two subunits ($g = 2$).

Denote

$$H_i = \left(H_{i1}^{(1)}, \dots, H_{ig}^{(1)}, H_{i1}^{(2)}, \dots, H_{ig}^{(2)} \right)^T, i = 1, \dots, N,$$

as the transformed values for the g subunits of the i -th cluster obtained from two prediction rules, where $(H_{i1}^{(t)}, \dots, H_{ig}^{(t)})$ are from the t -th prediction rule. We assume $H_i \sim N(\mu_{ki}, \Sigma)$, where

$$\mu_{ki} = \begin{pmatrix} \mu_{ki1} \\ \mu_{ki2} \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

and $\mu_{kij} = \mu_k \delta_{ij}$ is the mean probit for the j -th subunit in the i -th cluster for the k -th prediction rule, $k = 1, 2, i = 1, \dots, N, j = 1, 2$,

$$\Sigma_{11} = \begin{pmatrix} 1 & \rho_{11} \\ & \ddots \\ \rho_{11} & 1 \end{pmatrix}, \Sigma_{22} = \begin{pmatrix} 1 & \rho_{22} \\ & \ddots \\ \rho_{22} & 1 \end{pmatrix}, \Sigma_{12} = \begin{pmatrix} \rho & \rho_{12} \\ & \ddots \\ \rho_{12} & \rho \end{pmatrix}.$$

where $\delta_{ij} = 1$ if the j -th eye ($j = 1$ means left eye; $j = 2$ means right eye) of the i -th subject has progressed, and $= 0$ if it has not progressed.

To simulate the data, we set $\mu_1 = 0.80, \rho = 0.93, \rho_{11} = 0.59, \rho_{22} = 0.56$, and $\rho_{12} = 0.52$, as were estimated using the whole AMD data set ($n = 261$ subjects). We considered 4 sample sizes: $N = 50, 100, 200$, and 500 .

We generated 4,000 simulated data sets for each scenario. Since δ_{ij} is assumed to be fixed (cf. Assumption 2 in section “Extended Mann-Whitney U Statistic for Clustered Data”), for each of 4,000 simulated data sets in a scenario, we used the same set of $\delta_{ij}, i = 1, \dots, N, j = 1, 2$, where δ_{i1} (indicating left eye’s progression status) and δ_{i2} (indicating right eye’s progression status), $i = 1, \dots, N$, are generated from multinomial distributions with parameters $p_{11} = 0.115$ (proportion where both eyes progressed), $p_{10} = 0.142$ (proportion where only the left eye progressed), $p_{01} = 0.130$ (proportion where only the right eye progressed), and $p_{00} = 0.613$ (proportion where both eyes did not progress), respectively, where

$$\begin{aligned} p_{11} &= Pr(\delta_{i1} = 1 \ \& \ \delta_{i2} = 1), \\ p_{10} &= Pr(\delta_{i1} = 1 \ \& \ \delta_{i2} = 0), \\ p_{01} &= Pr(\delta_{i1} = 0 \ \& \ \delta_{i2} = 1), \\ p_{00} &= Pr(\delta_{i1} = 0 \ \& \ \delta_{i2} = 0), \\ 1 &= p_{11} + p_{10} + p_{01} + p_{00}. \end{aligned}$$

We are interested in testing the null hypothesis $H_0 : \Delta = 0$ versus the alternative hypothesis $H_a : \Delta = \Delta_0$, where

$$\Delta = \eta_c^{(1)} - \eta_c^{(2)}, \quad (9)$$

and

$$\eta_c^{(t)} = \left\{ \theta_c^{(t)} \sum_{i=1}^N c_i d_i + \theta^{(t)} \left[C \cdot D - \sum_{i=1}^N c_i d_i \right] \right\} / (C \cdot D), t = 1, 2.$$

We consider three values (0, 0.025, and 0.05) for Δ_0 . We can obtain μ_2 by solving Eq. 9 using numerical methods, given $\mu_1, \rho_{11}, \rho_{22}, \Delta$, and δ_{ij} , since there is no explicit closed-form expression to express μ_2 as a function of these variables.

We evaluate the performance of the extended Mann-Whitney U statistic in terms of percent bias ($100 \sum_{b=1}^B (\hat{\Delta}_b - \Delta) / \Delta$) where $\Delta \neq 0$, for the b-th simulation and V-statistic ($1/(B-1) \sum_{b=1}^B (\hat{\Delta}_b - \bar{\Delta})^2 / \text{Var}(\hat{\Delta}|H_a)$) of parameter estimates $\hat{\Delta}_b$, and coverage of the confidence interval $[\hat{\Delta}_{L,b}, \hat{\Delta}_{U,b}]$ of the parameter Δ , Type I error rate and power of the hypothesis test $H_0 : \Delta = 0$ versus $H_a : \Delta = \Delta_0$.

The power of the test is computed using Eq. 8 and the Type I error rate is calculated by setting $\Delta_0 = 0$.

The simulation results are shown in Table 2. We can see that the estimate $\hat{\Delta}$ is unbiased as the percent bias is close to zero. The sample variance tends to be slightly smaller than the theoretical variance when $nSubj = 50$ as the V-statistic is smaller than one. The coverage is close to the nominal level 95%, except when $nSubj = 50$ where the procedure slightly over-covers (consistent with the V-statistic). The theoretical and empirical powers are also in close agreement.

Discussion

In this article, we provide methods for assessing discrimination of risk prediction rules for disease progression as characterized by AUC, where the unit of analysis is the subunit (e.g. the eye) within a cluster (e.g., the person) and there is correlation between risk scores for multiple subunits (eyes) in the same cluster (person). The methods are applicable to both balanced (equal cluster size) and unbalanced (unequal cluster size) clustered data. The data analysis and the simulation study show that the proposed test performs well in assessing AUC of risk prediction rules and comparing AUC of competing risk prediction rules estimated from the same subjects.

We assumed that after probit transformation the bivariate random vector $[H_{i_1 j_1} - H_{k_1 \ell_1}, H_{i_2 j_2} - H_{k_2 \ell_2}]$ is bivariate normally distributed. For the AMD data set, we tested the bivariate normality of $(H_{ij}, H_{k\ell})$ using the Shapiro-Wilk test for the

Table 2 Simulation results ($\rho = 0.93, \rho_{11} = 0.59, \rho_{22} = 0.56, \rho_{12} = 0.52$)

$\Delta = 0$					
nSubj	pBias	Coverage	V-stat	Type I error rate	Emp. type I error rate
50	–	96.1	0.84	0.05	0.04
100	–	95.8	0.91	0.05	0.04
200	–	95.1	0.93	0.05	0.05
500	–	94.9	0.98	0.05	0.05
$\Delta = 0.025$					
nSubj	pBias	Coverage	V-stat	Power	Emp. power
50	2.23	96.4	0.85	0.15	0.13
100	0.62	95.6	0.95	0.21	0.20
200	0.05	94.9	0.98	0.45	0.43
500	0.21	95.2	0.99	0.84	0.84
$\Delta = 0.05$					
nSubj	pBias	Coverage	V-stat	Power	Emp. power
50	–0.05	96.4	0.87	0.46	0.45
100	0.27	95.0	1.01	0.72	0.72
200	–0.31	94.9	0.97	0.94	0.94
500	0.42	94.6	1.01	1.00	1.00

pBias the percent bias ($100 \sum_{b=1}^B (\hat{\Delta}_b - \Delta) / \Delta$) where $\Delta \neq 0$ and B is the total number of simulated data sets. *V-stat* the ratio of the empirical variance to the average theoretical variance. *emp. Type I error rate* empirical Type I error rate = the proportion of simulated data sets for which the test statistic $Z = |\hat{\Delta}|/sd(\hat{\Delta}|H_0)$ is greater than the critical value $Z_{\alpha/2}$ given $\Delta = 0$, where $Z_{\alpha/2}$ is the upper $100(\alpha/2)\%$ percentile of a standard normal distribution. *power* theoretical power from Eq. 8. *emp. power* empirical power = the proportion of simulated data sets for which the test statistic $Z = |\hat{\Delta}|/sd(\hat{\Delta}|H_0)$ is greater than the critical value $Z_{\alpha/2}$ given $\Delta > 0$

training set and the p-value = 0.22 indicates we do not reject the bivariate normality assumption. It will be a future research topic to check the robustness of our method to the violation of the bivariate normality assumption.

Another assumption of the extended Mann-Whitney U test is that replicates within a cluster are exchangeable. This is appropriate for the AMD data in which the progression status of the left eye and right eye for a subject are ascertained at one point in time. An interesting extension would be to consider a non-exchangeable within-cluster correlation structure, as might be applicable for longitudinal data.

Obuchowski and McClish [7] have also considered nonparametric analysis of clustered ROC curve data. Inference is based on the statistic $\hat{\theta}_c$ which is equivalent to $\hat{\eta}_c$ in Eq. 2. A U statistic approach is used to calculate $\text{Var}(\hat{\theta}_c)$ based on the quantities

$$V_{10}(Z_{ij}) = \frac{1}{D} \sum_{k=1}^N \sum_{\ell=1}^{d_k} U(Z_{k\ell} - Z_{ij}) \delta_{ij} (1 - \delta_{k\ell}),$$

$$V_{01}(Z_{k\ell}) = \frac{1}{C} \sum_{i=1}^N \sum_{j=1}^{c_i} U(Z_{k\ell} - Z_{ij}) \delta_{ij} (1 - \delta_{k\ell}).$$

An inherent assumption of this approach is that under H_a , $E[U(Z_{ij} - Z_{kl})]$ is the same when $i = k$ or when $i \neq k$. Under the probit shift alternative in the current paper, there are two distinct parameters θ_c and θ when $i = k$ or when $i \neq k$ (see Eq. 4). Hence, variances and covariances need to be computed separately for the components A and B corresponding to θ_c and θ , respectively. Obuchowski and McClish [7] also provide upper bounds on sample size estimates based on variance components in the observed data. In the current manuscript, we provide an explicit power formula (see Eq. 8) as a function of μ , the location parameter of the probit shift under H_a , and $\rho = \text{Corr}(H_{i1}, H_{i2}) =$ correlation between probit scores between two subunits in the same cluster. This power formula has been shown to be accurate in our simulation studies. An advantage of this approach is that power calculations can be performed for various levels of μ and ρ , whereas in Obuchowski and McClish [7] they are based on a specific dataset. An interesting design question is for a fixed total number of subunits $= C + D$, what is the optimal allocation between having many clusters with small cluster size vs having fewer clusters with large cluster size so as to minimize $\text{Var}(\hat{\eta}_c)$ or $\text{Var}[\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}]$.

Li and Zhou [4] provide a unified approach to nonparametric comparison of ROC curves for clustered data. The asymptotic joint distribution of ROC curves defined by two different markers accounting for both between marker and within-cluster (subject) variation is provided. The difference between AUC's corresponding to Δ is estimated from $\int_0^1 D(p)dp$, where

$$D(p) = \text{ROC}^{(1)}(p) - \text{ROC}^{(2)}(p)$$

$$\text{ROC}^{(v)}(p) = 1 - G^{(v)} \left\{ \left(F^{(v)} \right)^{-1} (1 - p) \right\}$$

$F^{(v)}$ and $G^{(v)}$ are the cumulative distribution function of the risk score $Z^{(v)}$ for cases (progressors) and controls (non-progressors), respectively. Unfortunately, there is in general no closed-form expression available for $\text{Var}(\hat{D}(p))$ and Monte-Carlo simulation and numerical integration is used instead. A nice feature of this approach is the avoidance of nonparametric density estimation, which can be computationally challenging in this setting. However, in the current paper, we use the rank-preserving property of the probit transformation to obtain a closed-form expression for $\text{Var}[\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}]$ in Eq. 7 which can be used for any continuous or ordinal scale.

Furthermore, another distinction between our paper and Obuchowski and McClish [7] and Li and Zhou [4] is that our scores are derived from a regression model as a function of one or more risk factors rather than based on a single diagnostic marker as is common in the ROC literature. Since the scores are based on estimated regression coefficients, the variability in estimation needs to be taken into account which we accomplished using multiple imputation methods.

Toledano and Gatsonis [13] propose an ordinal regression model to estimate effects of covariates on the area under the ROC, and compare AUC between

competing risk prediction rules, where parameters in the marginal ordinal regression model are estimated using GEE approaches. Our example is based on a continuous risk score, where there are few patients with exactly the same risk score. It is an open question how the Toledano and Gatsonis’s [13] approach would work with a sparse number of subjects in individual risk categories.

Recently, new measures of performance of prediction models have been proposed, such as net reclassification improvement (NRI) and integrated discrimination improvement (IDI) proposed by Pencina et al. [8]. These new measures offer incremental information over the AUC. Pencina et al. [8] pointed out in their Discussion that they still believe that improvement in the AUC should remain the first criterion, while NRI and IDI should also be taken into consideration. To our knowledge, these new measures have not been extended to handle clustered data yet. Hence, one possible future research area is to extend these new measures for clustered data.

The program that implements the methods in this paper was written in the R language and is available from the authors upon request.

Acknowledgements This work was supported by the National Institutes of Health Grant EY12269 from the National Eye Institute.

Appendix

Appendix 1 Calculating $\text{Var}(\hat{\eta}_c|H_a)$

We can use second order moments of the bivariate normal distribution to derive a closed-form expression for $\text{Var}(\hat{\eta}_c|H_a)$. Specifically,

$$\begin{aligned} \text{Var}(A|H_a) &= \theta_c(1 - \theta_c) \sum_{i=1}^N c_i d_i \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \frac{1}{2} \right) - \theta_c^2 \right] \\ &\cdot \left[\sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - 2 \sum_{i=1}^N c_i d_i \right], \end{aligned}$$

where $\Phi_2(a, b, \rho)$ is the cumulative distribution function of the bivariate normal distribution = $Pr\left(Z_1 \leq a, Z_2 \leq b | (Z_1, Z_2) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)\right)$. For example, the first term in $\text{Var}(A)$ is

$$\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} \text{Var}[U(Z_{ij} - Z_{i\ell})],$$

while the second term represents

$$\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{\ell_1 \neq j}^{g_i} \sum_{\ell_2 \neq \ell_1 \neq j}^{g_i} \text{Cov} [U (Z_{ij} - Z_{i\ell_1}), U (Z_{ij} - Z_{i\ell_2})].$$

Similarly,

$$\begin{aligned} \text{Var}(B|H_a) &= \sum_{i=1}^5 \Delta_{B_i}, \\ \text{Cov}(A, B|H_a) &= \sum_{i=1}^2 \Delta_{AB_i}, \end{aligned}$$

where

$$\begin{aligned} \Delta_{B_1} &= \theta(1 - \theta) \left[C \cdot D - \sum_{i=1}^N c_i d_i \right] \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1 + \rho}{2} \right) - \theta^2 \right] \\ &\cdot \left[D \sum_k c_k^2 + C \sum_i d_i^2 - 2 \cdot C \cdot D - \left(\sum_i c_i^2 d_i + \sum_i c_i d_i^2 - 2 \sum_i c_i d_i \right) \right] \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \rho \right) - \theta^2 \right] \\ &\cdot \left[\sum_i d_i (d_i - 1) \sum_k c_k (c_k - 1) - \sum_i d_i (d_i - 1) c_i (c_i - 1) \right], \\ \Delta_{B_2} &= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2} \right) - \theta^2 \right] \\ &\cdot \left[C^2 D - D \sum_i c_i^2 - 2 \left(C \sum_i c_i d_i - \sum_i c_i^2 d_i \right) \right] \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{\rho}{2} \right) - \theta^2 \right] \cdot \left[\left(\sum_i d_i^2 - D \right) \left(C^2 - \sum_i c_i^2 \right) \right. \\ &\left. - 2 \left(C \cdot \sum_i c_i d_i^2 - C \cdot \sum_i c_i d_i - \sum_i c_i^2 d_i^2 + \sum_i c_i^2 d_i \right) \right], \end{aligned}$$

$$\begin{aligned} \Delta_{B_3} = & \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2} \right) - \theta^2 \right] \\ & \cdot \left[C \left(D^2 - \sum_i d_i^2 \right) - 2 \left(D \sum_i c_i d_i - \sum_i c_i d_i^2 \right) \right] \\ & + \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{\rho}{2} \right) - \theta^2 \right] \left[\left(\sum_i c_i^2 - C \right) \left(D^2 - \sum_i d_i^2 \right) \right. \\ & \left. - 2 \left(D \left(\sum_i c_i^2 d_i - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i d_i^2 \right) \right], \end{aligned}$$

$$\Delta_{B_4} = \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), -\rho \right) - \theta^2 \right] \left[\left(\sum_i c_i d_i \right)^2 - \sum_i c_i^2 d_i^2 \right],$$

$$\begin{aligned} \Delta_{B_5} = & 2 \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), -\frac{\rho}{2} \right) - \theta^2 \right] \\ & \cdot \left[CD \sum_i c_i d_i - C \sum_i c_i d_i^2 - D \sum_i c_i^2 d_i + 2 \sum_i c_i^2 d_i^2 - \left(\sum_i c_i d_i \right)^2 \right], \end{aligned}$$

$$\Delta_{AB_1} = \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta_c), \frac{\sqrt{1-\rho}}{2} \right) - \theta^2 \right] \left[C \sum_i c_i d_i - \sum_i c_i^2 d_i \right],$$

$$\Delta_{AB_2} = \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta_c), \frac{\sqrt{1-\rho}}{2} \right) - \theta^2 \right] \left[D \sum_i c_i d_i - \sum_i c_i d_i^2 \right].$$

Calculating $\text{Var}(A)$

Note that

$$(1 - \delta_{ij}) \delta_{ij} = 0$$

and that

$$\begin{aligned}
 E\{[U(Z_{ij} - Z_{k\ell})](1 - \delta_{ij})\delta_{k\ell}\} &= Pr[U(Z_{ij} - Z_{k\ell}) = 1](1 - \delta_{ij})\delta_{k\ell} \\
 &= Pr[Z_{ij} < Z_{k\ell}](1 - \delta_{ij})\delta_{k\ell} \\
 &= Pr[H(Z_{ij}) < H(Z_{k\ell})](1 - \delta_{ij})\delta_{k\ell} \\
 &= Pr[H(Z_{ij}) - H(Z_{k\ell}) < 0](1 - \delta_{ij})\delta_{k\ell} \\
 &= \begin{cases} \Phi\left(\frac{\mu}{\sqrt{2(1-\rho)}}\right) = \theta_c & \text{if } i = k \\ \Phi\left(\frac{\mu}{\sqrt{2}}\right) = \theta & \text{if } i \neq k \end{cases}
 \end{aligned}$$

We can get

$$\begin{aligned}
 \text{Var}(A) &= \text{Var}\left(\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} U(Z_{ij} - Z_{i\ell})(1 - \delta_{ij})\delta_{i\ell}\right) \\
 &= \sum_{i_1=1}^N \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1}(1 - \delta_{i_1 j_2})\delta_{i_1 \ell_2} \\
 &\quad \text{Cov}(U(Z_{i_1 j_1} - Z_{i_1 \ell_1}), U(Z_{i_1 j_2} - Z_{i_1 \ell_2})) \\
 &= \sum_{i_1=1}^N \sum_{j_1=j_2}^{g_{i_1}} \sum_{\ell_1=\ell_2}^{g_{i_1}} (1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1}(1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1} \\
 &\quad \text{Cov}(U(Z_{i_1 j_1} - Z_{i_1 \ell_1}), U(Z_{i_1 j_1} - Z_{i_1 \ell_1})) \\
 &\quad + \sum_{i_1=1}^N \sum_{j_1 \neq j_2}^{g_{i_1}} \sum_{\ell_1=\ell_2}^{g_{i_1}} (1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1}(1 - \delta_{i_1 j_2})\delta_{i_1 \ell_1} \\
 &\quad \text{Cov}(U(Z_{i_1 j_1} - Z_{i_1 \ell_1}), U(Z_{i_1 j_2} - Z_{i_1 \ell_1})) \\
 &\quad + \sum_{i_1=1}^N \sum_{j_1=j_2}^{g_{i_1}} \sum_{\ell_1 \neq \ell_2}^{g_{i_1}} (1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1}(1 - \delta_{i_1 j_1})\delta_{i_1 \ell_2} \\
 &\quad \text{Cov}(U(Z_{i_1 j_1} - Z_{i_1 \ell_1}), U(Z_{i_1 j_1} - Z_{i_1 \ell_2})) \\
 &\quad + \sum_{i_1=1}^N \sum_{j_1 \neq j_2}^{g_{i_1}} \sum_{\ell_1 \neq \ell_2}^{g_{i_1}} (1 - \delta_{i_1 j_1})\delta_{i_1 \ell_1}(1 - \delta_{i_1 j_2})\delta_{i_1 \ell_2} \\
 &\quad \text{Cov}(U(Z_{i_1 j_1} - Z_{i_1 \ell_1}), U(Z_{i_1 j_2} - Z_{i_1 \ell_2})) \\
 &= \theta_c(1 - \theta_c) \sum_{i_1} c_{i_1} d_{i_1}
 \end{aligned}$$

$$\begin{aligned}
& + \left[\Phi_2 \left(\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \frac{1}{2} \right) - \theta_c^2 \right] \sum_{i_1} c_{i_1} d_{i_1} (d_{i_1} - 1) \\
& + \left[\Phi_2 \left(\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \frac{1}{2} \right) - \theta_c^2 \right] \sum_{i_1} d_{i_1} c_{i_1} (c_{i_1} - 1) \\
& + 0.
\end{aligned}$$

Calculating $\text{Var}(B)$

$$\begin{aligned}
\text{Var}(B) &= \text{Var} \left(\sum_{i=1}^N \sum_{k=1, k \neq i}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell} \right) \\
&= \sum_{i=1}^5 \Delta_{B_i}
\end{aligned}$$

Calculation Δ_{B_1}

$$\begin{aligned}
\Delta_{B_1} &= \sum_{i_1=i_2, i_1 \neq k_1} \sum_{k_1=k_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_1 \ell_2} \\
&\quad \text{Cov} \left(U(Z_{i_1 j_1} - Z_{k_1 \ell_1}), U(Z_{i_1 j_2} - Z_{k_1 \ell_2}) \right) \\
&= \sum_{i_1 \neq k_1} \sum_{j_1=j_2} \sum_{\ell_1=\ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} \text{Cov} \left(U(Z_{i_1 j_1} - Z_{k_1 \ell_1}), U(Z_{i_1 j_1} - Z_{k_1 \ell_1}) \right) \\
&\quad + \sum_{i_1 \neq k_1} \sum_{j_1 \neq j_2} \sum_{\ell_1=\ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_1 \ell_1} \text{Cov} \left(U(Z_{i_1 j_1} - Z_{k_1 \ell_1}), U(Z_{i_1 j_2} - Z_{k_1 \ell_1}) \right) \\
&\quad + \sum_{i_1 \neq k_1} \sum_{j_1=j_2} \sum_{\ell_1 \neq \ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_2} \text{Cov} \left(U(Z_{i_1 j_1} - Z_{k_1 \ell_1}), U(Z_{i_1 j_1} - Z_{k_1 \ell_2}) \right) \\
&\quad + \sum_{i_1 \neq k_1} \sum_{j_1 \neq j_2} \sum_{\ell_1 \neq \ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_1 \ell_2} \text{Cov} \left(U(Z_{i_1 j_1} - Z_{k_1 \ell_1}), U(Z_{i_1 j_2} - Z_{k_1 \ell_2}) \right) \\
&= \theta(1 - \theta) \sum_{i_1 \neq k_1} d_{i_1} c_{k_1} \\
&\quad + \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1+\rho}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1} d_{i_1} (d_{i_1} - 1) c_{k_1} \\
&\quad + \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1+\rho}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1} d_{i_1} c_{k_1} (c_{k_1} - 1) \\
&\quad + \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \rho \right) - \theta^2 \right] \sum_{i_1 \neq k_1} d_{i_1} (d_{i_1} - 1) c_{k_1} (c_{k_1} - 1)
\end{aligned}$$

Calculation Δ_{B_2}

$$\begin{aligned}
\Delta_{B_2} &= \sum_{i_1=i_2, i_1 \neq k_1, k_1 \neq k_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_2 \ell_2} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_1 j_2} - Z_{k_2 \ell_2} \right) \right) \\
&= \sum_{i_1 \neq k_1, i_1 \neq k_2, k_1 \neq k_2} \sum_{j_1=j_2} \sum_{\ell_1} \sum_{\ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_1}) \delta_{k_2 \ell_2} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_1 j_1} - Z_{k_2 \ell_2} \right) \right) \\
&+ \sum_{i_1 \neq k_1, i_1 \neq k_2, k_1 \neq k_2} \sum_{j_1 \neq j_2} \sum_{\ell_1} \sum_{\ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_2 \ell_2} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_1 j_2} - Z_{k_2 \ell_2} \right) \right) \\
&= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1, i_1 \neq k_2, k_1 \neq k_2} d_{i_1} c_{k_1} c_{k_2} \\
&+ \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{\rho}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1, i_1 \neq k_2, k_1 \neq k_2} d_{i_1} (d_{i_1} - 1) c_{k_1} c_{k_2}
\end{aligned}$$

Calculation Δ_{B_3}

$$\begin{aligned}
\Delta_{B_3} &= \sum_{i_1 \neq i_2, i_1 \neq k_1, k_1 = k_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_2} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{k_1 \ell_2} \right) \right) \\
&= \sum_{i_1 \neq k_1, i_1 \neq i_2, k_1 \neq i_2} \sum_{j_1} \sum_{j_2} \sum_{\ell_1=\ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_1} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{k_1 \ell_1} \right) \right) \\
&+ \sum_{i_1 \neq k_1, i_1 \neq i_2, k_1 \neq i_2} \sum_{j_1} \sum_{j_2} \sum_{\ell_1 \neq \ell_2} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_2} \\
&\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{k_1 \ell_2} \right) \right) \\
&= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1, i_1 \neq i_2, k_1 \neq i_2} d_{i_1} d_{i_2} c_{k_1} \\
&+ \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{\rho}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1, i_1 \neq i_2, k_1 \neq i_2} d_{i_1} d_{i_2} c_{k_1} (c_{k_1} - 1)
\end{aligned}$$

Calculation \triangle_{B_4}

$$\begin{aligned}\triangle_{B_4} &= \sum_{i_1=k_2, i_1 \neq k_1} \sum_{k_1 \neq i_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{k_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{i_1 \ell_2} \\ &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{k_1 j_2} - Z_{i_1 \ell_2} \right) \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), -\rho \right) - \theta^2 \right] \sum_{i_1 \neq k_1} d_{i_1} c_{i_1} d_{k_1} c_{k_1}\end{aligned}$$

Calculation \triangle_{B_5}

$$\begin{aligned}\triangle_{B_5} &= 2 \sum_{i_1=k_2, i_1 \neq k_1} \sum_{k_1 \neq i_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\ &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{k_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{i_1 \ell_2} \right) \right) \\ &= 2 \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta), -\frac{\rho}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_1, i_1 \neq i_2, k_1 \neq i_2} d_{i_1} c_{i_1} c_{k_1} d_{i_2}\end{aligned}$$

Calculating $\text{Cov}(A, B)$

$$\begin{aligned}\text{Cov}(A, B) &= \text{Cov} \left(\sum_{i_1=1}^N \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} U \left(Z_{i_1 j_1} - Z_{i_1 \ell_1} \right) (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1}, \right. \\ &\quad \left. \sum_{i_2=1}^N \sum_{k_2=1, k_2 \neq i_2}^N \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_2}} U \left(Z_{i_2 j_2} - Z_{k_2 \ell_2} \right) (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \right) \\ &= \sum_{i=1}^2 \triangle_{AB_i}\end{aligned}$$

Calculating \triangle_{AB_1}

$$\begin{aligned}\triangle_{AB_1} &= \sum_{i_1=i_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_2 \ell_2} \\ &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{i_1 \ell_1} \right), U \left(Z_{i_1 j_2} - Z_{k_2 \ell_2} \right) \right) \\ &= \sum_{i_1 \neq k_2} \sum_{j_1=j_2} \sum_{\ell_1} \sum_{\ell_2} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_1 j_1}) \delta_{k_2 \ell_2} \\ &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{i_1 \ell_1} \right), U \left(Z_{i_1 j_1} - Z_{k_2 \ell_2} \right) \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta_c), \frac{\sqrt{1-\rho}}{2} \right) - \theta^2 \right] \sum_{i_1 \neq k_2} d_{i_1} c_{i_1} c_{k_2}\end{aligned}$$

Calculating \triangle_{AB_2}

$$\begin{aligned}
 \triangle_{AB_2} &= \sum_{i_1=k_2, i_1 \neq i_2, i_2 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\
 &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{i_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{i_1 \ell_2} \right) \right) \\
 &= \sum_{i_1 \neq i_2} \sum_{j_1} \sum_{j_2} \sum_{\ell_1 = \ell_2} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_1} \\
 &\quad \text{Cov} \left(U \left(Z_{i_1 j_1} - Z_{i_1 \ell_1} \right), U \left(Z_{i_2 j_2} - Z_{i_1 \ell_1} \right) \right) \\
 &= \left[\Phi_2 \left(\Phi^{-1}(\theta), \Phi^{-1}(\theta_c), \frac{\sqrt{1-\rho}}{2} \right) - \theta^2 \right] \sum_{i_1 \neq i_2} d_{i_1} c_{i_1} d_{i_2}
 \end{aligned}$$

Appendix 2 Calculating $\text{Cov} \left(\hat{\eta}_c^{(1)}, \hat{\eta}_c^{(2)} \right)$

We can obtain

$$\begin{aligned}
 \text{Cov} \left(A^{(1)}, A^{(2)} \right) &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{\sqrt{(1-\rho_{11})(1-\rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \sum_{i=1}^N c_i d_i \\
 &\quad + \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1-\rho_{11})(1-\rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \\
 &\quad \cdot \left[\sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - 2 \sum_{i=1}^N c_i d_i \right], \\
 \text{Cov} \left(A^{(1)}, B^{(2)} \right) &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1-\rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \\
 &\quad \cdot \left[(C+D) \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i - \sum_{i=1}^N c_i d_i^2 \right], \\
 \text{Cov} \left(B^{(1)}, A^{(2)} \right) &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1-\rho_{22})}} \right) - \theta_1 \theta_{c_2} \right] \\
 &\quad \cdot \left[(C+D) \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i - \sum_{i=1}^N c_i d_i^2 \right], \\
 \text{Cov} \left(B^{(1)}, B^{(2)} \right) &= \sum_{t=1}^9 s_t,
 \end{aligned}$$

where

$$\begin{aligned}
s_1 &= [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho) - \theta_1 \theta_2] \left[C \cdot D - \sum_{i=1}^N c_i d_i \right] \\
s_2 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho + \rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \\
&\quad \cdot \left[2 \sum_{i=1}^N c_i d_i + C \sum_{i=1}^N d_i^2 + D \sum_{i=1}^N c_i^2 - \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i^2 d_i - 2CD \right] \\
s_3 &= [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho_{12}) - \theta_1 \theta_2] \left\{ \left[\sum_{i=1}^N c_i^2 \right] \left[\sum_{i=1}^N d_i^2 \right] \right. \\
&\quad \left. - C \sum_{i=1}^N d_i^2 - D \sum_{i=1}^N c_i^2 + CD - \sum_{i=1}^N c_i^2 d_i^2 + \sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i d_i \right\} \\
s_4 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2 \right] \cdot \left\{ C \left[D^2 - \sum_{i=1}^N d_i^2 \right] - 2 \left[D \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i d_i^2 \right] \right\} \\
s_5 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \cdot \left\{ \left(\sum_i c_i^2 - C \right) \left(D^2 - \sum_i d_i^2 \right) \right. \\
&\quad \left. - 2 \left[D \left(\sum_i c_i^2 d_i - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i d_i^2 \right] \right\} \\
s_6 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2 \right] \left\{ D \left[C^2 - \sum_{i=1}^N c_i^2 \right] - 2 \left[C \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i \right] \right\} \\
s_7 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \left\{ \left(\sum_i d_i^2 - D \right) \left(C^2 - \sum_i c_i^2 \right) \right. \\
&\quad \left. - 2 \left[C \left(\sum_i c_i d_i^2 - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i^2 d_i \right] \right\} \\
s_8 &= [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\rho_{12}) - \theta_1 \theta_2] \left[\left(\sum_{i=1}^N c_i d_i \right)^2 - \sum_{i=1}^N c_i^2 d_i^2 \right] \\
s_9 &= 2 \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \\
&\quad \cdot \left[CD \sum_{i=1}^N c_i d_i - C \sum_{i=1}^N c_i d_i^2 - D \sum_{i=1}^N c_i^2 d_i - \left(\sum_{i=1}^N c_i d_i \right)^2 + 2 \sum_{i=1}^N c_i^2 d_i^2 \right]
\end{aligned}$$

Calculating $\text{Cov}(A^{(1)}, A^{(2)})$

$$\begin{aligned} & \text{Cov}(A^{(1)}, A^{(2)}) \\ &= \text{Cov} \left[\sum_{i_1=1}^N \sum_{j_1}^{g_{i_1}} \sum_{\ell_1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} U(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)}), \sum_{i_2=1}^N \sum_{j_2}^{g_{i_2}} \sum_{\ell_2}^{g_{i_2}} (1 - \delta_{i_2 j_2}) \delta_{i_2 \ell_2} U(Z_{i_2 j_2}^{(2)} - Z_{i_2 \ell_2}^{(2)}) \right] \end{aligned}$$

Denote

$$\Delta_{A_1 A_2, 1} = \sum_{i_1=i_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_2}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_2 \ell_2} \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)}), U(Z_{i_2 j_2}^{(2)} - Z_{i_2 \ell_2}^{(2)}) \right)$$

and

$$\begin{aligned} \Delta_{A_1 A_2, 2} &= \sum_{i_1 \neq i_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_2}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_2 \ell_2} \\ & \quad \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)}), U(Z_{i_2 j_2}^{(2)} - Z_{i_2 \ell_2}^{(2)}) \right) \end{aligned}$$

Then

$$\text{Cov}(A^{(1)}, A^{(2)}) = \Delta_{A_1 A_2, 1} + \Delta_{A_1 A_2, 2}.$$

We can get

$$\Delta_{A_1 A_2, 2} = 0$$

since i_1 and i_2 are two different subjects and hence are independent.

Now we calculate $\Delta_{A_1 A_2, 1}$.

$$\begin{aligned} & \Delta_{A_1 A_2, 1} \\ &= \sum_i \sum_{j_1=j_2=j} \sum_{\ell_1=\ell_2=\ell} (1 - \delta_{ij}) \delta_{i\ell} \text{Cov} \left(U(Z_{ij}^{(1)} - Z_{i\ell}^{(1)}), U(Z_{ij}^{(2)} - Z_{i\ell}^{(2)}) \right) \\ & \quad + \sum_i \sum_{j_1 \neq j_2} \sum_{\ell_1=\ell_2=\ell} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \delta_{i\ell} \text{Cov} \left(U(Z_{ij_1}^{(1)} - Z_{i\ell}^{(1)}), U(Z_{ij_2}^{(2)} - Z_{i\ell}^{(2)}) \right) \\ & \quad + \sum_i \sum_{j_1=j_2=j} \sum_{\ell_1 \neq \ell_2} (1 - \delta_{ij}) \delta_{i\ell_1} \delta_{i\ell_2} \text{Cov} \left(U(Z_{ij}^{(1)} - Z_{i\ell_1}^{(1)}), U(Z_{ij}^{(2)} - Z_{i\ell_2}^{(2)}) \right) \\ & \quad + \sum_i \sum_{j_1 \neq j_2} \sum_{\ell_1 \neq \ell_2} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \delta_{i\ell_1} \delta_{i\ell_2} \text{Cov} \left(U(Z_{ij_1}^{(1)} - Z_{i\ell_1}^{(1)}), U(Z_{ij_2}^{(2)} - Z_{i\ell_2}^{(2)}) \right). \end{aligned}$$

Note that

$$\begin{aligned} & \text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{il}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{il}^{(2)}\right)\right) \\ &= E\left[U\left(Z_{ij_1}^{(1)} - Z_{il_1}^{(1)}\right)U\left(Z_{ij_2}^{(2)} - Z_{il_2}^{(2)}\right)\right] - E\left[U\left(Z_{ij_1}^{(1)} - Z_{il_1}^{(1)}\right)\right]E\left[U\left(Z_{ij_2}^{(2)} - Z_{il_2}^{(2)}\right)\right] \\ &= Pr\left(H_{ij}^{(1)} - H_{il}^{(1)} < 0 \& H_{ij}^{(2)} - H_{il}^{(2)} < 0\right) - Pr\left(H_{ij}^{(1)} - H_{il}^{(1)} < 0\right)Pr\left(H_{ij}^{(2)} - H_{il}^{(2)} < 0\right) \end{aligned}$$

and

$$\begin{pmatrix} H_{ij}^{(1)} - H_{il}^{(1)} \\ H_{ij}^{(2)} - H_{il}^{(2)} \end{pmatrix} \sim N\left(\begin{pmatrix} -\mu_1 \\ -\mu_2 \end{pmatrix}, \begin{pmatrix} 2(1 - \rho_{11}) & \sigma_{12} \\ \sigma_{12} & 2(1 - \rho_{22}) \end{pmatrix}\right)$$

where

$$\begin{aligned} \sigma_{12} &= \text{Cov}\left(H_{ij}^{(1)} - H_{il}^{(1)}, H_{ij}^{(2)} - H_{il}^{(2)}\right) \\ &= \text{Cov}\left(H_{ij}^{(1)}, H_{ij}^{(2)}\right) - \text{Cov}\left(H_{ij}^{(1)}, H_{il}^{(2)}\right) - \text{Cov}\left(H_{il}^{(1)}, H_{ij}^{(2)}\right) + \text{Cov}\left(H_{il}^{(1)}, H_{il}^{(2)}\right) \\ &= \rho - \rho_{12} - \rho_{12} + \rho \\ &= 2(\rho - \rho_{12}) \end{aligned}$$

Hence we can get

$$\begin{aligned} & \text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{il}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{il}^{(2)}\right)\right) \\ &= \Phi_2\left(\frac{\mu_1}{\sqrt{2(1 - \rho_{11})}}, \frac{\mu_2}{\sqrt{2(1 - \rho_{22})}}, \frac{\rho - \rho_{12}}{\sqrt{(1 - \rho_{11})(1 - \rho_{22})}}\right) \\ &\quad - \Phi\left(\frac{\mu_1}{\sqrt{2(1 - \rho_{11})}}\right)\Phi\left(\frac{\mu_2}{\sqrt{2(1 - \rho_{22})}}\right) \\ &= \Phi_2\left(\theta_{c_1}^{-1}, \theta_{c_2}^{-1}, \frac{\rho - \rho_{12}}{\sqrt{(1 - \rho_{11})(1 - \rho_{22})}}\right) - \theta_{c_1}\theta_{c_2}. \end{aligned}$$

Similarly, we can get

$$\begin{aligned} & \text{Cov}\left(U\left(Z_{ij_1}^{(1)} - Z_{il}^{(1)}\right), U\left(Z_{ij_2}^{(2)} - Z_{il}^{(2)}\right)\right) \\ &= \Phi_2\left(\Phi^{-1}\left(\theta_{c_1}\right), \Phi^{-1}\left(\theta_{c_2}\right), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})(1 - \rho_{22})}}\right) - \theta_{c_1}\theta_{c_2} \end{aligned}$$

and

$$\begin{aligned} & \text{Cov} \left(U \left(Z_{ij}^{(1)} - Z_{i\ell_1}^{(1)} \right), U \left(Z_{ij}^{(2)} - Z_{i\ell_2}^{(2)} \right) \right) \\ &= \Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})(1 - \rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \end{aligned}$$

and

$$\text{Cov} \left(U \left(Z_{ij_1}^{(1)} - Z_{i\ell_1}^{(1)} \right), U \left(Z_{ij_2}^{(2)} - Z_{i\ell_2}^{(2)} \right) \right) = 0.$$

Hence, we can get

$$\begin{aligned} \Delta_{A_1 A_2, 1} &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{\sqrt{(1 - \rho_{11})(1 - \rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \sum_{i=1}^N c_i d_i \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})(1 - \rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \\ &\quad \left[\sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - 2 \sum_{i=1}^N c_i d_i \right] \end{aligned}$$

Therefore

$$\begin{aligned} & \text{Cov} \left(A^{(1)}, A^{(2)} \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{\sqrt{(1 - \rho_{11})(1 - \rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \sum_{i=1}^N c_i d_i \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})(1 - \rho_{22})}} \right) - \theta_{c_1} \theta_{c_2} \right] \\ &\quad \left[\sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - 2 \sum_{i=1}^N c_i d_i \right] \tag{10} \end{aligned}$$

Calculating $\text{Cov} \left(A^{(1)}, B^{(2)} \right)$

$$\begin{aligned} & \text{Cov} \left(A^{(1)}, B^{(2)} \right) \\ &= \text{Cov} \left[\sum_{i_1=1}^N \sum_{j_1}^{g_{i_1}} \sum_{\ell_1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), \right. \\ &\quad \left. \sum_{i_2=1}^N \sum_{k_2=1, k_2 \neq i_2}^{g_{i_2}} \sum_{j_2}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_2}} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} U \left(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right] \end{aligned}$$

Denote

$$\begin{aligned} & \Delta_{A_1 B_2, 1} \\ = & \sum_{i_1 \neq i_2, i_1 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \\ & \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{A_1 B_2, 2} \\ = & \sum_{i_1=i_2, i_1 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_2 \ell_2} \\ & \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{A_1 B_2, 3} \\ = & \sum_{i_1 \neq i_2, i_1 = k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{i_1}} \sum_{j_2=1}^{i_2} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\ & \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)} \right) \right) \end{aligned}$$

Then

$$\text{Cov} \left(A^{(1)}, B^{(2)} \right) = \Delta_{A_1 B_2, 1} + \Delta_{A_1 B_2, 2} + \Delta_{A_1 B_2, 3}.$$

We can easily get

$$\Delta_{A_1 B_2, 1} = 0$$

since subjects i_1 , i_2 , and k_2 are different subjects and hence are independent.

We can get

$$\begin{aligned} & \Delta_{A_1 B_2, 2} \\ = & \sum_{i \neq k} \sum_{j_1 = j_2 = j} \sum_{\ell_1=1}^{g_i} \sum_{\ell_2=1}^{g_k} (1 - \delta_{ij}) \delta_{i \ell_1} \delta_{k \ell_2} \text{Cov} \left(U \left(Z_{ij}^{(1)} - Z_{i \ell_1}^{(1)} \right), U \left(Z_{ij}^{(2)} - Z_{k \ell_2}^{(2)} \right) \right) \\ & + \sum_{i \neq k} \sum_{j_1 \neq j_2} \sum_{\ell_1=1}^{g_i} \sum_{\ell_2=1}^{g_k} (1 - \delta_{ij_1}) (1 - \delta_{ij_2}) \delta_{i \ell_1} \delta_{k \ell_2} \\ & \text{Cov} \left(U \left(Z_{ij_1}^{(1)} - Z_{i \ell_1}^{(1)} \right), U \left(Z_{ij_2}^{(2)} - Z_{k \ell_2}^{(2)} \right) \right) \end{aligned}$$

We can get

$$\begin{aligned} & \text{Cov} \left(U \left(Z_{ij}^{(1)} - Z_{il_1}^{(1)} \right), U \left(Z_{ij}^{(2)} - Z_{kl_2}^{(2)} \right) \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \end{aligned}$$

and

$$\text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{k \ell_2}^{(2)} \right) \right) = 0$$

Hence

$$\begin{aligned} & \Delta_{A_1 B_2, 2} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \sum_{i \neq k} c_i d_i c_k \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \left[C \cdot \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i \right] \end{aligned}$$

Now we calculate $\Delta_{A_1 B_2, 3}$.

$$\begin{aligned} & \Delta_{A_1 B_2, 3} \\ &= \sum_{i_1 \neq i_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1=\ell_2=\ell}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell} (1 - \delta_{i_2 j_2}) \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell}^{(2)} \right) \right) \\ & \quad + \sum_{i_1 \neq i_2} \sum_{j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1 \neq \ell_2}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{i_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)} \right) \right) \end{aligned}$$

We can get

$$\begin{aligned} & \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell}^{(2)} \right) \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \end{aligned}$$

and

$$\text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{i_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_2 \ell_2}^{(2)} \right) \right) = 0$$

Hence

$$\begin{aligned} & \triangle_{A_1 B_2, 3} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \sum_{i_1 \neq i_2} c_{i_1} d_{i_1} d_{i_2} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \left[D \cdot \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i d_i^2 \right] \end{aligned}$$

Therefore

$$\begin{aligned} & \text{Cov} \left(A^{(1)}, B^{(2)} \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_{c_1}), \Phi^{-1}(\theta_2), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{11})}} \right) - \theta_{c_1} \theta_2 \right] \quad (11) \\ & \cdot \left[(C + D) \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i - \sum_{i=1}^N c_i d_i^2 \right] \end{aligned}$$

Calculating $\text{Cov} \left(B^{(1)}, A^{(2)} \right)$

By symmetry to $\text{Cov} \left(A^{(1)}, B^{(2)} \right)$, we can get

$$\begin{aligned} & \text{Cov} \left(B^{(1)}, A^{(2)} \right) \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_{c_2}), \frac{\rho - \rho_{12}}{2\sqrt{(1 - \rho_{22})}} \right) - \theta_1 \theta_{c_2} \right] \quad (12) \\ & \cdot \left[(C + D) \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i - \sum_{i=1}^N c_i d_i^2 \right]. \end{aligned}$$

Calculating $\text{Cov}(B^{(1)}, B^{(2)})$

$$\begin{aligned} & \text{Cov}(B^{(1)}, B^{(2)}) \\ = & \text{Cov} \left[\sum_{i_1=1}^N \sum_{k_1=1, k_1 \neq i_1}^N \sum_{j_1}^{g_{i_1}} \sum_{\ell_1}^{g_{k_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} U(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}), \right. \\ & \left. \sum_{i_2=1}^N \sum_{k_2=1, k_2 \neq i_2}^N \sum_{j_2}^{g_{i_2}} \sum_{\ell_2}^{g_{k_2}} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} U(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)}) \right] \end{aligned}$$

Denote

$$\begin{aligned} & \Delta_{B_1 B_2, 0} \\ = & \sum_S \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \\ & \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}), U(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)}) \right) \end{aligned}$$

where S is the set that i_1, k_1, i_2, k_2 are all not equal, and

$$\begin{aligned} & \Delta_{B_1 B_2, 1} \\ = & \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 = i_2, k_1 = k_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{k_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_2} \\ & \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}), U(Z_{i_1 j_2}^{(2)} - Z_{k_1 \ell_2}^{(2)}) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 2} \\ = & \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 \neq i_2, k_1 = k_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_2} \\ & \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}), U(Z_{i_2 j_2}^{(2)} - Z_{k_1 \ell_2}^{(2)}) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 3} \\ = & \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 = i_2, k_1 \neq k_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \\ & \text{Cov} \left(U(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}), U(Z_{i_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)}) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 4} \\ &= \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 \neq i_2, k_1 \neq k_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 5} \\ &= \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 = k_2, k_1 = i_2}} \sum_{j_1=1}^{g_{k_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_1}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{k_1 j_2}) \delta_{i_1 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{k_1 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 6} \\ &= \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 \neq k_2, k_1 = i_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{k_1 j_2}) \delta_{k_2 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{k_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 7} \\ &= \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 = k_2, k_1 \neq i_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \Delta_{B_1 B_2, 8} \\ &= \sum_{\substack{i_1 \neq k_1, i_2 \neq k_2 \\ i_1 \neq k_2, k_1 \neq i_2}} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_2 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_2 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

Then

$$\text{Cov}\left(B^{(1)}, B^{(2)}\right) = \sum_{t=0}^8 \Delta_{B_1 B_2, t}.$$

We can easily to get

$$\Delta_{B_1 B_2, 0} = \Delta_{B_1 B_2, 4} = \Delta_{B_1 B_2, 8} = 0.$$

Now we calculate $\Delta_{B_1 B_2, 1}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 1} \\ = & \sum_{i \neq k} \sum_{j_1=j_2=j}^{g_{i_1}} \sum_{\ell_1=\ell_2=\ell}^{g_{k_1}} (1 - \delta_{ij}) \delta_{k\ell} \text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{k\ell}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{k\ell}^{(2)}\right)\right) \\ & + \sum_{i \neq k} \sum_{j_1 \neq j_2}^{g_{i_1}} \sum_{\ell_1=\ell_2=\ell}^{g_{k_1}} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \delta_{k\ell} \text{Cov}\left(U\left(Z_{ij_1}^{(1)} - Z_{k\ell}^{(1)}\right), U\left(Z_{ij_2}^{(2)} - Z_{k\ell}^{(2)}\right)\right) \\ & + \sum_{i \neq k} \sum_{j_1=j_2=j}^{g_{i_1}} \sum_{\ell_1 \neq \ell_2}^{g_{k_1}} (1 - \delta_{ij}) \delta_{k\ell_1} \delta_{k\ell_2} \text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{k\ell_1}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{k\ell_2}^{(2)}\right)\right) \\ & + \sum_{i \neq k} \sum_{j_1 \neq j_2}^{g_{i_1}} \sum_{\ell_1 \neq \ell_2}^{g_{k_1}} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \delta_{k\ell_1} \delta_{k\ell_2} \text{Cov}\left(U\left(Z_{ij_1}^{(1)} - Z_{k\ell_1}^{(1)}\right), U\left(Z_{ij_2}^{(2)} - Z_{k\ell_2}^{(2)}\right)\right) \end{aligned}$$

We can get

$$\text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{k\ell}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{k\ell}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho\right) - \theta_1 \theta_2$$

and

$$\text{Cov}\left(U\left(Z_{ij_1}^{(1)} - Z_{k\ell}^{(1)}\right), U\left(Z_{ij_2}^{(2)} - Z_{k\ell}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho + \rho_{12}}{2}\right) - \theta_1 \theta_2$$

and

$$\text{Cov}\left(U\left(Z_{ij}^{(1)} - Z_{k\ell_1}^{(1)}\right), U\left(Z_{ij}^{(2)} - Z_{k\ell_2}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho + \rho_{12}}{2}\right) - \theta_1 \theta_2$$

and

$$\text{Cov} \left(U \left(Z_{ij_1}^{(1)} - Z_{k\ell_1}^{(1)} \right), U \left(Z_{ij_2}^{(2)} - Z_{k\ell_2}^{(2)} \right) \right) = \Phi_2 \left(\Phi^{-1} \left(\theta_1 \right), \Phi^{-1} \left(\theta_2 \right), \rho_{12} \right) - \theta_1 \theta_2$$

Note that

$$\begin{aligned} & \sum_{i \neq k} \sum_{\ell=1}^{g_{k_1}} \delta_{k\ell} \sum_{j_1 \neq j_2}^{g_{i_1}} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \\ &= \sum_{i \neq k} c_k [d_i^2 - d_i] \\ &= \sum_{i=1}^N \sum_{k=1}^N c_k [d_i^2 - d_i] - \sum_{i=k=1}^N c_i [d_i^2 - d_i] \\ &= C \left[\sum_{i=1}^N d_i^2 - D \right] - \sum_{i=1}^N c_i d_i^2 + \sum_{i=1}^N c_i d_i \end{aligned}$$

Similarly, we can get

$$\begin{aligned} & \sum_{i \neq k} \sum_{j=1}^{g_{i_1}} (1 - \delta_{ij}) \sum_{\ell_1 \neq \ell_2}^{g_{k_1}} \delta_{k\ell_1} \delta_{k\ell_2} \\ &= D \left[\sum_{i=1}^N c_i^2 - C \right] - \sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i \end{aligned}$$

and

$$\begin{aligned} & \sum_{i \neq k} \sum_{j_1 \neq j_2}^{g_{i_1}} (1 - \delta_{ij_1})(1 - \delta_{ij_2}) \sum_{\ell_1 \neq \ell_2}^{g_{k_1}} \delta_{k\ell_1} \delta_{k\ell_2} \\ &= \sum_i d_i (d_i - 1) \sum_k c_k (c_k - 1) - \sum_i c_i d_i (c_i - 1) (d_i - 1) \end{aligned}$$

and

$$\sum_{i \neq k} \sum_{j=1}^{g_{i_1}} \sum_{\ell=1}^{g_{k_1}} (1 - \delta_{ij}) \delta_{k\ell} = CD - \sum_i c_i d_i$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 1} \\ &= [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho) - \theta_1 \theta_2] \left[C \cdot D - \sum_{i=1}^N c_i d_i \right] \\ &+ \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho + \rho_{12}}{2}\right) - \theta_1 \theta_2 \right] \left[2 \sum_{i=1}^N c_i d_i + C \sum_{i=1}^N d_i^2 + D \sum_{i=1}^N c_i^2 - \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i^2 d_i - 2CD \right] \\ &+ [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho_{12}) - \theta_1 \theta_2] \left\{ \left[\sum_{i=1}^N c_i^2 \right] \left[\sum_{i=1}^N d_i^2 \right] \right. \\ &- C \sum_{i=1}^N d_i^2 - D \sum_{i=1}^N c_i^2 + CD - \sum_{i=1}^N c_i^2 d_i^2 \\ &\left. + \sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i d_i \right\} \end{aligned}$$

Now we calculate $\Delta_{B_1 B_2, 2}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 2} \\ &= \sum_{i_1 \neq k_1 \neq i_2, j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1=\ell_2=\ell}^{g_{k_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell} \text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{k_1 \ell}^{(2)}\right)\right) \\ &+ \sum_{i_1 \neq k_1 \neq i_2, j_1=1}^{g_{i_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_1 \neq \ell_2}^{g_{k_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{k_1 \ell_2} \text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{k_1 \ell_2}^{(2)}\right)\right) \end{aligned}$$

We can get

$$\text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{k_1 \ell}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2}\right) - \theta_1 \theta_2$$

and

$$\text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{k_1 \ell_2}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2}\right) - \theta_1 \theta_2$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 2} \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2}\right) - \theta_1 \theta_2 \right] \sum_{i_1 \neq i_2 \neq k_1} c_{k_1} d_{i_1} d_{i_2} \\ &+ \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2}\right) - \theta_1 \theta_2 \right] \sum_{i_1 \neq i_2 \neq k_1} d_{i_1} d_{i_2} (c_{k_1}^2 - c_{k_1}) \end{aligned}$$

Note that

$$\begin{aligned}
 & \sum_{i_1 \neq i_2 \neq k_1} c_{k_1} d_{i_1} d_{i_2} \\
 &= \sum_{i_1 \neq i_2} d_{i_1} d_{i_2} \left[\sum_{k_1} c_{k_1} - c_{i_1} - c_{i_2} \right] \\
 &= C \sum_{i_1 \neq i_2} d_{i_1} d_{i_2} - \sum_{i_1 \neq i_2} d_{i_1} c_{i_1} d_{i_2} - \sum_{i_1 \neq i_2} d_{i_1} d_{i_2} c_{i_2} \\
 &= C \left[D^2 - \sum_{i=1}^N d_i^2 \right] - 2 \left[D \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i d_i^2 \right]
 \end{aligned}$$

and

$$\begin{aligned}
 & \sum_{i_1 \neq i_2 \neq k_1} d_{i_1} d_{i_2} (c_{k_1}^2 - c_{k_1}) \\
 &= \sum_{i_1 \neq i_2} d_{i_1} d_{i_2} \left[\sum_{k_1} c_{k_1} (c_{k_1} - 1) - c_{i_1} (c_{i_1} - 1) - c_{i_2} (c_{i_2} - 1) \right] \\
 &= \left(\sum_i c_i^2 - C \right) \left(D^2 - \sum_i d_i^2 \right) - 2 \left[D \left(\sum_i c_i^2 d_i - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i d_i^2 \right]
 \end{aligned}$$

Hence

$$\begin{aligned}
 & \triangle_{B_1 B_2, 2} \\
 &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2 \right] \left\{ C \left[D^2 - \sum_{i=1}^N d_i^2 \right] - 2 \left[D \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i d_i^2 \right] \right\} \\
 &+ \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \left\{ \left(\sum_i c_i^2 - C \right) \left(D^2 - \sum_i d_i^2 \right) \right. \\
 &\left. - 2 \left[D \left(\sum_i c_i^2 d_i - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i d_i^2 \right] \right\}
 \end{aligned}$$

Now we calculate $\Delta_{B_1 B_2, 3}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 3} \\ &= \sum_{i_1 \neq k_1 \neq k_2} \sum_{j_1=j_2=j}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j}) \delta_{k_1 \ell_1} \delta_{k_2 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_1 j}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \\ & \quad + \sum_{i_1 \neq k_1 \neq k_2} \sum_{j_1 \neq j_2}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_1 j_2}) \delta_{k_2 \ell_2} \\ & \quad \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

We can get

$$\text{Cov} \left(U \left(Z_{i_1 j}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_1 j}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) = \Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2$$

and

$$\text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{i_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) = \Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 3} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2 \right] \sum_{i_1 \neq k_1 \neq k_2} d_{i_1} c_{k_1} c_{k_2} \\ & \quad + \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \sum_{i_1 \neq k_1 \neq k_2} c_{k_1} c_{k_2} (d_{i_1}^2 - d_{i_1}) \end{aligned}$$

Note that

$$\begin{aligned} & \sum_{i_1 \neq k_1 \neq k_2} d_{i_1} c_{k_1} c_{k_2} \\ &= \sum_{k_1 \neq k_2} c_{k_1} c_{k_2} \left[\sum_{i_1} d_{i_1} - d_{k_1} - d_{k_2} \right] \\ &= D \sum_{k_1 \neq k_2} c_{k_1} c_{k_2} - \sum_{k_1 \neq k_2} c_{k_1} c_{k_2} d_{k_1} - \sum_{k_1 \neq k_2} c_{k_1} c_{k_2} d_{k_2} \\ &= D \left[C^2 - \sum_{i=1}^N c_i^2 \right] - 2 \left[C \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i \right] \end{aligned}$$

and

$$\begin{aligned} & \sum_{k_1 \neq k_2 \neq i_1} c_{k_1} c_{k_2} (d_{i_1}^2 - d_{i_1}) \\ &= \sum_{k_1 \neq k_2} c_{k_1} c_{k_2} \left[\sum_{i_1} d_{i_1} (d_{i_1} - 1) - d_{k_1} (d_{k_1} - 1) - d_{k_2} (d_{k_2} - 1) \right] \\ &= \left(\sum_i d_i^2 - D \right) \left(C^2 - \sum_i c_i^2 \right) - 2 \left[C \left(\sum_i c_i d_i^2 - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i^2 d_i \right] \end{aligned}$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 3} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2} \right) - \theta_1 \theta_2 \right] \left\{ D \left[C^2 - \sum_{i=1}^N c_i^2 \right] - 2 \left[C \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i \right] \right\} \\ &+ \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2} \right) - \theta_1 \theta_2 \right] \left\{ \left(\sum_i d_i^2 - D \right) \left(C^2 - \sum_i c_i^2 \right) \right. \\ &\left. - 2 \left[C \left(\sum_i c_i d_i^2 - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i^2 d_i \right] \right\} \end{aligned}$$

Now we calculate $\Delta_{B_1 B_2, 5}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 5} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\rho_{12} \right) - \theta_1 \theta_2 \right] \sum_{i_1 \neq k_1} c_{i_1} d_{i_1} c_{k_1} d_{k_1} \\ &= \left[\Phi_2 \left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\rho_{12} \right) - \theta_1 \theta_2 \right] \left[\left(\sum_{i=1}^N c_i d_i \right)^2 - \sum_{i=1}^N c_i^2 d_i^2 \right]. \end{aligned}$$

Now we calculate $\Delta_{B_1 B_2, 6}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 6} \\ &= \sum_{i_1 \neq k_1 \neq k_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{k_1}} \sum_{\ell_2=1}^{g_{k_2}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{k_1 j_2}) \delta_{k_2 \ell_2} \\ & \text{Cov} \left(U \left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)} \right), U \left(Z_{k_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)} \right) \right) \end{aligned}$$

We can get

$$\text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}\right), U\left(Z_{k_1 j_2}^{(2)} - Z_{k_2 \ell_2}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 6} \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2\right] \sum_{i_1 \neq k_1 \neq k_2} d_i c_{k_1} d_{k_1} c_{k_2} \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2\right] \left[CD \sum_{i=1}^N c_i d_i - C \sum_{i=1}^N c_i d_i^2 - D \sum_{i=1}^N c_i^2 d_i - \left(\sum_{i=1}^N c_i d_i\right)^2 + 2 \sum_{i=1}^N c_i^2 d_i^2\right]. \end{aligned}$$

Now we calculate $\Delta_{B_1 B_2, 7}$.

$$\begin{aligned} & \Delta_{B_1 B_2, 7} \\ &= \sum_{i_1 \neq k_1 \neq i_2} \sum_{j_1=1}^{g_{i_1}} \sum_{\ell_1=1}^{g_{k_1}} \sum_{j_2=1}^{g_{i_2}} \sum_{\ell_2=1}^{g_{i_1}} (1 - \delta_{i_1 j_1}) \delta_{k_1 \ell_1} (1 - \delta_{i_2 j_2}) \delta_{i_1 \ell_2} \\ & \quad \text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)}\right)\right). \end{aligned}$$

We can get

$$\text{Cov}\left(U\left(Z_{i_1 j_1}^{(1)} - Z_{k_1 \ell_1}^{(1)}\right), U\left(Z_{i_2 j_2}^{(2)} - Z_{i_1 \ell_2}^{(2)}\right)\right) = \Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2$$

Hence

$$\begin{aligned} & \Delta_{B_1 B_2, 7} \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2\right] \sum_{i_1 \neq k_1 \neq i_2} d_i c_{k_1} d_{i_2} c_{i_1} \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}\right) - \theta_1 \theta_2\right] \left[CD \sum_{i=1}^N c_i d_i - C \sum_{i=1}^N c_i d_i^2 - D \sum_{i=1}^N c_i^2 d_i^2 - \left(\sum_{i=1}^N c_i d_i\right)^2 + 2 \sum_{i=1}^N c_i^2 d_i^2\right] \end{aligned}$$

Therefore, we can get

$$\begin{aligned} & \text{Cov}\left(B^{(1)}, B^{(2)}\right) \\ &= \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho\right) - \theta_1 \theta_2\right] \left[C \cdot D - \sum_{i=1}^N c_i d_i\right] \\ & \quad + \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho + \rho_{12}}{2}\right) - \theta_1 \theta_2\right] \left[2 \sum_{i=1}^N c_i d_i + C \sum_{i=1}^N d_i^2 + D \sum_{i=1}^N c_i^2 - \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i^2 d_i - 2CD\right] \\ & \quad + \left[\Phi_2\left(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \rho_{12}\right) - \theta_1 \theta_2\right] \left\{ \left[\sum_{i=1}^N c_i^2\right] \left[\sum_{i=1}^N d_i^2\right] \right\} \end{aligned}$$

$$\begin{aligned}
 & -C \sum_{i=1}^N d_i^2 - D \sum_{i=1}^N c_i^2 + CD - \sum_{i=1}^N c_i^2 d_i^2 \\
 & + \sum_{i=1}^N c_i^2 d_i + \sum_{i=1}^N c_i d_i^2 - \sum_{i=1}^N c_i d_i \left. \right\} \\
 & + [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2}) - \theta_1 \theta_2] \left\{ C \left[D^2 - \sum_{i=1}^N d_i^2 \right] - 2 \left[D \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i d_i^2 \right] \right\} \\
 & + [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2}) - \theta_1 \theta_2] \left\{ \left(\sum_i c_i^2 - C \right) \left(D^2 - \sum_i d_i^2 \right) \right. \\
 & \left. - 2 \left[D \left(\sum_i c_i^2 d_i - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i d_i^2 \right] \right\} \\
 & + [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho}{2}) - \theta_1 \theta_2] \left\{ D \left[C^2 - \sum_{i=1}^N c_i^2 \right] - 2 \left[C \sum_{i=1}^N c_i d_i - \sum_{i=1}^N c_i^2 d_i \right] \right\} \\
 & + [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), \frac{\rho_{12}}{2}) - \theta_1 \theta_2] \left\{ \left(\sum_i d_i^2 - D \right) \left(C^2 - \sum_i c_i^2 \right) \right. \\
 & \left. - 2 \left[C \left(\sum_i c_i d_i^2 - \sum_i c_i d_i \right) - \sum_i c_i^2 d_i^2 + \sum_i c_i^2 d_i \right] \right\} \\
 & + [\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\rho_{12}) - \theta_1 \theta_2] \left[\left(\sum_{i=1}^N c_i d_i \right)^2 - \sum_{i=1}^N c_i^2 d_i^2 \right] \\
 & + 2 \left[\Phi_2(\Phi^{-1}(\theta_1), \Phi^{-1}(\theta_2), -\frac{\rho_{12}}{2}) - \theta_1 \theta_2 \right] \left[CD \sum_{i=1}^N c_i d_i - C \sum_{i=1}^N c_i d_i^2 - D \sum_{i=1}^N c_i^2 d_i - \left(\sum_{i=1}^N c_i d_i \right)^2 + 2 \sum_{i=1}^N c_i^2 d_i^2 \right]
 \end{aligned}$$

(13)

Appendix 3 Calculating p-Value Based on Multiple Imputation

Rubin [11] mentioned that the p-value for a parameter estimate $\hat{\eta}$ obtained by using multiple imputation with m imputations is obtained by using t-statistic $\hat{\eta} / \sqrt{\text{Var}(\hat{\eta})}$ with degrees of freedom

$$df = (m - 1) \left[1 + \frac{mW}{(m + 1)B} \right]^2,$$

where

$$\begin{aligned}
 W &= \frac{1}{m} \sum_{i=1}^m \text{Var}(\hat{\eta}_i) \\
 B &= \left(1 + \frac{1}{m} \right) \frac{1}{(m - 1)} \sum_{i=1}^m (\hat{\eta}_i - \hat{\eta})^2 \\
 \hat{\eta} &= \frac{1}{m} \sum_{i=1}^m \hat{\eta}_i.
 \end{aligned}$$

and

$$\text{Var}(\hat{\eta}) = W + B.$$

$\hat{\eta}_i$ is the parameter estimated based on the i -th imputation.

References

1. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. *Diagn. Radiol.* **143**, 29–36 (1982)
2. Hodges, J.L., Jr., Lehmann, E.L.: The efficiency of some nonparametric competitors of the t test. *Ann. Math. Stat.* **27**, 324–335 (1956)
3. Kendall, M.G., Stuart, A.: *The Advanced Theory of Statistics*. Hafner, New York (1969)
4. Li, G., Zhou, K.: A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *J. Am. Stat. Assoc.* **103**(482), 705–713 (2008)
5. National Eye Institute.: Age-related macular degeneration. http://www.nei.nih.gov/health/maculardegen/armd_facts.asp (2011). Retrieved on Mar 2011
6. Obuchowski, N.A.: Receiver operating characteristic curves and their use in radiology. *Radiology* **229**, 3–8 (2003)
7. Obuchowski, N.A., McClish, D.K.: Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Stat. Med.* **16**(13), 1529–1542 (1997)
8. Pencina, M.J., D'Agostino, R.B. Sr., D'Agostino, R.B. Jr., Vasan, R.S.: Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Stat. Med.* **27**, 157–172 (2008)
9. Rosner, B., Glynn, R.J.: Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models. *Biometrics* **65**, 188–197 (2009)
10. Rosner, B., Glynn, R.J., Lee, M.T.: Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. *Biometrics* **62**, 1251–1259 (2006)
11. Rubin, D.B.: *Multiple Imputation for Non-response in Surveys*. Wiley, New York (1987)
12. Seddon, J.M., Cote, J., Rosner, B.: Progression of age-related macular degeneration. *Arch. Ophthalmol.* **121**, 1728–1737 (2003)
13. Toledano, A.Y., Gatsonis, C.A.: GEEs for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* **55**, 488–496 (1996)
14. Zou, K.H., O'Malley, A.J., Mauri, L.: Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* **115**, 654–657 (2007)

Time-Dependent AUC with Right-Censored Data: A Survey

Paul Blanche, Aurélien Latouche, and Vivian Viallon

Abstract The ROC curve and the corresponding AUC are popular tools for the evaluation of diagnostic tests. They have been recently extended to assess prognostic markers and predictive models. However, due to the many particularities of time-to-event outcomes, various definitions and estimators have been proposed in the literature. This review article aims at presenting the ones that accommodate to right-censoring, which is common when evaluating such prognostic markers.

Introduction

In the medical literature, a variety of general criteria have been used to assess diagnostic tests [14, 24]. Among them, the ROC curve and the area under it – the AUC – are popular tools, originally aimed at evaluating the discriminant power of continuous diagnostic tests. In this simple situation, the outcome status D is a binary variable (typically, $D = 1$ for cases and $D = 0$ for controls) and the ROC curve for a continuous diagnostic test X plots the true positive rate, or sensitivity, $\text{TPR}(c) = \mathbb{P}(X > c | D = 1)$ against the false positive rate, or one minus the specificity, $\text{FPR}(c) = \mathbb{P}(X > c | D = 0)$, when making threshold c vary. The AUC, which is the area under this curve, is a commonly used summary measure

P. Blanche (✉)

University of Bordeaux, ISPED and INSERM U897, Bordeaux, France
e-mail: Paul.Blanche@isped.u-bordeaux2.fr

A. Latouche

Conservatoire national des arts et métiers, Paris, France
e-mail: aurelien.latouche@cnam.fr

V. Viallon

UMRESTTE (Univ. Lyon 1 and IFSTTAR), Bron, France
e-mail: viallon@math.univ-lyon1.fr

of the information contained in the sequences $(\text{TPR})_{c \in \mathbf{R}}$ and $(\text{FPR})_{c \in \mathbf{R}}$. As such, it inherits some of the properties of true and false positive rates. In particular, it is not affected by the disease prevalence (unlike positive and negative predictive values) and it can be evaluated from random samples of cases and controls. It also has a nice interpretation since it corresponds to the probability that the marker value of a randomly selected case exceeds that of a randomly selected control.

The extension of the AUC (and other evaluation criteria) to the setting of prognostic markers has raised several issues. In particular, when evaluating such markers, the outcome status typically changes over time: in a cohort study for instance, patients are disease-free when entering the study and may develop the disease during the study. This leads to three major differences with the evaluation of diagnostic tests. First, this time-dependent outcome status (which may be defined in several ways, as will be seen in section “Time-Dependent AUCs in the Standard Setting”) naturally implies that sensitivity, specificity, ROC curves, their AUC values and, more generally, any extension of the criteria used in the diagnostic setting, are functions of time as well. Second, the time-to-event, i.e. the time between the entry in the study and the disease onset, is usually *censored* and not fully observed, requiring dedicated inference. Third, the time lag between the entry in the study and the disease onset also leads to two further refinements: (i) the marker can be repeatedly measured over time and (ii) competing events (in addition to censoring) may be observed between the entry and the putative disease onset.

These particularities has led to the development of numerous methods aimed at estimating the time-dependent AUC for prognostic markers. In this paper, we review those that accommodate to right-censoring. Some notations are introduced in the following section “Notations”. Then, in section “Time-Dependent AUCs in the Standard Setting” we will present several definitions and estimators of the time-dependent AUC in the “standard” setting of a baseline marker and univariate survival data. Section “Time-Dependent ROC Curve and AUCs with Longitudinal Marker” will cover the case of longitudinal markers which corresponds to the marker being repeatedly measured over time, while we will discuss the competing events setting in section “Time-Dependent AUC and Competing Risks”. Finally, concluding remarks will be given in section “Discussion”.

Notations

Let T_i and C_i denote survival and censoring times for subject i , $i = 1, \dots, n$. We further let $Z_i = \min(T_i, C_i)$ and $\delta_i = 1(T_i \leq C_i)$ denote the observed time and the status indicator respectively. We will denote by $D_i(t)$ the time-dependent outcome status for subject i at time t , $t \geq 0$. Several definitions for $D_i(t)$ will be given hereafter, but we will always have $D_i(t) = 1$ if subject i is considered as a case at time t and $D_i(t) = 0$ if subject i is considered as a control at time t .

We will denote by X the marker under study, which can be a single biological marker or several biological markers combined into a predictive model (in this case, it is assumed throughout this article that the predictive model has been constructed on an independent data set; otherwise sub-sampling techniques are needed [21]). Without loss of generality, we will suppose that larger values of X are associated with greater risks (otherwise, X can be recoded to achieve this). We will denote by g and G^{-1} the probability density function and the quantile function of marker X . In section “Time-Dependent AUCs in the Standard Setting”, we assume that marker X is measured once at $t = 0$, and we will denote by X_i the marker value for subject i . In section “Time-Dependent ROC Curve and AUCs with Longitudinal Marker”, which treats the longitudinal setting, the marker is measured repeatedly over time, and we will denote by $X_i(s)$ the marker value at time s for subject i .

Time-Dependent AUCs in the Standard Setting

Definitions of time-dependent ROC curves, $\text{ROC}(t)$, follow from definitions of usual ROC curves and thus rely on first defining time-dependent true and false positive rates. For any threshold c , these two functions of time are defined as $\text{TPR}(c, t) = \mathbb{P}(X > c | D(t) = 1)$ and $\text{FPR}(c, t) = \mathbb{P}(X > c | D(t) = 0)$. $\text{ROC}(t)$ then simply plots $\text{TPR}(c, t)$ against $\text{FPR}(c, t)$ making threshold c vary. The time-dependent AUC at time t is then defined as the area under this curve,

$$\text{AUC}(t) = \int_{-\infty}^{\infty} \text{TPR}(c, t) \left| \frac{\partial \text{FPR}(c, t)}{\partial c} \right| dc. \quad (1)$$

As a matter of fact, these definitions deeply rely on that of the outcome status at time t , $D(t)$. Heagerty and Zheng [17] described several definitions of *cases* and *controls* in this survival outcome setting. According to Heagerty and Zheng’s terminology and still denoting by T_i survival time for subject i , cases are said to be *incident* if $T_i = t$ is used to define cases at time t , and *cumulative* if $T_i \leq t$ is used instead. Similarly, depending on whether $T_i > \tau$ for a large time $\tau > t$ or $T_i > t$ is used for defining controls at time t , they are said to be *static* or *dynamic* controls. Depending on the definition retained for cases and controls at time t , four definitions of the time-dependent AUC value may be put forward. In the following paragraphs, we will present formulas and estimators for the most commonly used ones and will discuss their respective interests.

The Cumulative Dynamic AUC: $\text{AUC}^{\mathbb{C}, \mathbb{D}}(t)$

The setting of cumulative cases and dynamic controls may be regarded as the most natural choice for planning enrollment criteria in clinical trials or when specific

evaluation times are of particular interest. It simply corresponds to defining cases at time t as subjects who experienced the event prior to time t , and controls at time t as patients who were still event-free at time t . In other words, it corresponds to setting $D_i(t) = \mathbb{I}(T_i \leq t)$. Cumulative true positive rates and dynamic false positive rates are then respectively defined as

$$\text{TPR}^{\text{C}}(c, t) = \mathbb{P}(X > c | T \leq t) \quad \text{and} \quad \text{FPR}^{\text{D}}(c, t) = \mathbb{P}(X > c | T > t). \quad (2)$$

The *cumulative/dynamic* AUC at time t is then obtained by using these definitions of true and false positive rates in (1). Usually, however, $\mathbb{I}(T \leq t)$ is not observed for all subjects due to the presence of censoring before time t and simple contingency tables can therefore not be used to return estimates of $\text{TPR}^{\text{C}}(c, t)$, $\text{FPR}^{\text{D}}(c, t)$ and $\text{AUC}^{\text{C,D}}(t)$. To handle censoring, Bayes’ theorem can be used to rewrite $\text{AUC}^{\text{C,D}}(t)$ as a function of the conditional survival function $\mathbb{P}(T > t | X = x)$ (see section “Methods Based on Primary Estimates of $\mathbb{P}(T > t | X = x)$ ” below). Other approaches rely on so-called Inverse Probability of Censoring Weighted (IPCW) estimates (see section “Methods Based on IPCW Estimators” below). Before describing these two approaches in more details below, we shall add that Chambless and Diao [5] developed an alternative method – which will not be described here – based on an idea similar to the one used to derive the Kaplan-Meier estimator of the cumulative distribution function in the presence of censoring. Among all these methods, only those relying on primary estimates of $\mathbb{P}(T > t | X = x)$ (and a recent extension of IPCW estimates proposed in [2]) may account for the dependence between censoring and the marker (since they basically only assume that T and C are independent given X and not that T and C are independent). We refer the reader to [2, 19, 37] for empirical comparisons and illustrations of these various methods.

Methods Based on Primary Estimates of $\mathbb{P}(T > t | X = x)$

Bayes’ theorem yields the following expressions for $\text{TPR}^{\text{C}}(c, t)$ and $\text{FPR}^{\text{D}}(c, t)$

$$\text{TPR}^{\text{C}}(c, t) = \frac{\int_c^\infty \mathbb{P}(T \leq t | X = x)g(x)dx}{\mathbb{P}(T \leq t)}, \quad \text{FPR}^{\text{D}}(c, t) = \frac{\int_c^\infty \mathbb{P}(T > t | X = x)g(x)dx}{\mathbb{P}(T > t)}.$$

From (1), it readily follows that

$$\text{AUC}^{\text{C,D}}(t) = \int_{-\infty}^\infty \int_c^\infty \frac{\mathbb{P}(T \leq t | X = x)\mathbb{P}(T > t | X = c)}{\mathbb{P}(T \leq t)\mathbb{P}(T > t)}g(x)g(c)dxdc. \quad (3)$$

Since $\mathbb{P}(T > t) = \int_{-\infty}^\infty \mathbb{P}(T > t | X = x)g(x)dx$, any estimator $\widehat{S}_n(t|x)$ of the conditional survival function $\mathbb{P}(T > t | X = x)$ yields an estimator of $\text{AUC}^{\text{C,D}}(t)$:

$$\widehat{\text{AUC}}^{\text{C,D}}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]\mathbb{I}(X_i > X_j)}{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]}.$$

In [5], the authors suggested to use a Cox model to derive estimates $\widehat{S}_n(t|x)$, while one of the methods described in Heagerty et al. [18] reduces to using the conditional Kaplan-Meier estimator as in [1]. Some theoretical results for these two methods can be found in [32] and [4, 7, 20] respectively.

We shall add that Viallon and Latouche [37] related $\text{AUC}^{\text{C,D}}(t)$ to the quantity $\mathbb{P}(T \leq t|X = G^{-1}(q)) - \text{a time-dependent version of the predictiveness curve}$:

$$\text{AUC}^{\text{C,D}}(t) = \frac{\int_0^1 q \mathbb{P}(T \leq t|X = G^{-1}(q)) dq - [\mathbb{P}(T \leq t)]^2/2}{\mathbb{P}(T > t)\mathbb{P}(T \leq t)}.$$

This confirms that most standard statistical summaries of predictability and discrimination can be derived from the predictiveness curve, as pointed out in [14, 15].

Methods Based on IPCW Estimators

In [19] and [36], the authors independently suggested to use IPCW-type estimates:

$$\widehat{\text{TPR}}^{\text{C}}(c,t) = \frac{\sum_{i=1}^n \mathbb{I}(X_i > c, Z_i \leq t) \frac{\delta_i}{n\widehat{S}_C(Z_i)}}{\sum_{i=1}^n \mathbb{I}(Z_i \leq t) \frac{\delta_i}{n\widehat{S}_C(Z_i)}}, \quad \widehat{\text{FPR}}^{\text{D}}(c,t) = \frac{\sum_{i=1}^n \mathbb{I}(X_i > c, Z_i > t)}{\sum_{i=1}^n \mathbb{I}(Z_i > t)},$$

where $\widehat{S}_C(\cdot)$ is the Kaplan-Meier estimator of the survival function of the censoring time C . The expression of the false positive rate estimator is more compact because weights all equal $1/(n\widehat{S}_C(t))$ under the assumption of independence between C and X , and then vanish. $\widehat{\text{FPR}}^{\text{D}}(c,t)$ corresponds to 1 minus the empirical distribution function of X among individuals for whom $Z_i > t$. In the absence of censoring before time t , $\widehat{\text{TPR}}^{\text{C}}(c,t)$ also reduces to the usual empirical version of $\text{TPR}^{\text{C}}(c,t)$, i.e., 1 minus the empirical distribution function of X among individuals for whom $T_i \leq t$.

It can be shown (see [19, 30]) that an estimator of $\text{AUC}^{\text{C,D}}(t)$ is then given by

$$\widehat{\text{AUC}}^{\text{C,D}}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{I}(Z_i \leq t) \mathbb{I}(Z_j > t) \mathbb{I}(X_i > X_j) \frac{\delta_i}{\widehat{S}_C(Z_i)\widehat{S}_C(t)}}{n^2 \widehat{S}(t) [1 - \widehat{S}(t)]},$$

where $\widehat{S}(t)$ is the Kaplan-Meier estimator of $\mathbb{P}(T > t)$.

Theoretical guarantees for these estimators can be found in [19] and [36]. These estimators are in a sense more flexible than those presented in section “Methods Based on Primary Estimates of $\mathbb{P}(T > t|X = x)$ ” above: they are model-free and they do not rely on any bandwidth selection (unlike the estimator of Heagerty et al. [18] for instance, which is based on a local version of the Kaplan-Meier estimator). However, when censoring may depend on marker X , quantities like the conditional survival function of C given the marker X , $S_C(\cdot|X)$, have to be estimated [2], which implies either to work under some (semi-)parametric model or the selection of some parameter if nonparametric estimation is preferred.

The Incident Static AUC: $AUC^{\mathbb{I},\mathbb{S}}(t)$

Using the dynamic definition of controls, the control group varies with time, and so does the x -axis of the corresponding ROC curves: in situations where trends over time are of particular interest, this renders their interpretation more difficult (since such trends may be partly due to changing control groups). Moreover, the group of static controls is interesting in that it tries to mimic the group of individuals who never develop the disease, which can be seen as an ideal control group in some situations. In particular, patients with preclinical diseases are eliminated from the control group as far as possible, if τ is large enough.

Regarding cases, the incident definition has several advantages over the dynamic definition [25]. The cumulative TPR does not distinguish between events that occur early versus late, and it shows redundant information over time (since early events are also included in the cumulative TPR for late evaluation times). Moreover, as pointed out by Cai et al. [3], the cumulative TPR can be computed from the incident TPR when the distribution of the event time is known.

Putting all this together, several authors have proposed estimators of the time-dependent ROC curve relying on the incident definition of cases and static definition of controls. Standard numerical integration techniques are then used to compute an estimate for $AUC^{\mathbb{I},\mathbb{S}}(t)$ from the estimators of the ROC curve.

Incident true positive rates and static false positive rates are defined as

$$TPR^{\mathbb{I}}(c,t) = \mathbb{P}(X > c|T = t) \quad \text{and} \quad FPR_{\tau}^{\mathbb{S}}(c) = \mathbb{P}(X > c|T > \tau). \tag{4}$$

Applying Bayes' theorem, they can further be rewritten (see, e.g., [32]) as

$$TPR^{\mathbb{I}}(c,t) = \frac{\int_c^{\infty} f(t|x)g(x)dx}{\int_{-\infty}^{\infty} f(t|x)g(x)dx} \quad \text{and} \quad FPR_{\tau}^{\mathbb{S}}(c) = \frac{\int_c^{\infty} \mathbb{P}(T > \tau|X = x)g(x)dx}{\int_{-\infty}^{\infty} \mathbb{P}(T > \tau|X = x)g(x)dx},$$

where $f(t|x) = \partial\mathbb{P}(T \leq t|X = x)/\partial t$ is the conditional density function of T given $X = x$.

Under a standard Cox model of the form $\lambda(t;X) = \lambda_0(t) \exp(\beta X)$ – here $\lambda(t;X)$ stands for the conditional hazard rate of T given X while λ_0 is the unspecified baseline hazard rate – Song and Zhou [32] deduced that

$$TPR^{\mathbb{I}}(c,t) = \frac{\int_c^{\infty} \exp(\beta x) \exp\{-\Lambda_0(t) \exp(\beta x)\}g(x)dx}{\int_{-\infty}^{\infty} \exp(\beta x) \exp\{-\Lambda_0(t) \exp(\beta x)\}g(x)dx}$$

$$FPR_{\tau}^{\mathbb{S}}(c) = \frac{\int_c^{\infty} \exp\{-\Lambda_0(\tau) \exp(\beta x)\}g(x)dx}{\int_{-\infty}^{\infty} \exp\{-\Lambda_0(\tau) \exp(\beta x)\}g(x)dx},$$

where $\Lambda_0(t) = \int_{-\infty}^t \lambda_0(u)du$ is the cumulative baseline hazard function. Estimation of $TPR^{\mathbb{I}}(c,t)$ and $FPR_{\tau}^{\mathbb{S}}(c)$ can then be achieved by plug-in methods. We shall add that Song and Zhou actually considered a slightly more general set-up where additional covariates can be accounted for.

In [17], Heagerty and Zheng adopted a slightly different approach. To estimate $\text{TPR}^{\mathbb{I}}(c, t)$, they used a (possibly time-varying-coefficients) Cox model of the form $\lambda(t; X) = \lambda_0(t) \exp(\beta(t)X)$ in combination with the fact that the distribution of $X \cdot \exp(\beta X)$ for subjects in the risk set at time t is equal to the conditional distribution of X given $T = t$ (see, e.g., [38]). Setting $R(t) = \{i : Z_i \geq t\}$, this leads to

$$\widehat{\text{TPR}}^{\mathbb{I}}(c, t) = \frac{\sum_{i \in R(t)} \mathbb{I}(X_i > c) \exp\{\beta(t)X_i\}}{\sum_{i \in R(t)} \exp\{\beta(t)X_i\}}.$$

As for the estimation of $\text{FPR}_{\tau}^{\mathbb{S}}(c)$, they proposed a model-free approach using the empirical distribution function for marker values among the control set $S_{\tau} := \{i : Z_i > \tau\}$. Namely, denoting by n_{τ} the cardinality of S_{τ} , they proposed

$$\widehat{\text{FPR}}_{\tau}^{\mathbb{S}}(c) = \frac{1}{n_{\tau}} \sum_{i \in S_{\tau}} \mathbb{I}(X_i > c),$$

which is $\widehat{\text{FPR}}^{\mathbb{D}}(c, t)$ of section “Methods Based on IPCW Estimators”, except τ is used instead of t .

Cai et al. [3] proposed another approach in the context of longitudinal markers; it will be described in more details in section “Time-Dependent ROC Curve and AUCs with Longitudinal Marker”. In addition, two non parametric approaches were recently proposed (see [33] and [29]).

Note also that estimators for the time-dependant incident/dynamic AUC, $\text{AUC}^{\mathbb{I}, \mathbb{D}}(t)$, can be obtained by simply replacing τ by t in the definitions of $\widehat{\text{FPR}}_{\tau}^{\mathbb{S}}(c)$ above [17]. A global accuracy measure has further been derived from the definition of $\text{AUC}^{\mathbb{I}, \mathbb{D}}(t)$, which is particularly appealing when no a priori time t is identified and/or when trends over time are not of interest [17].

Time-Dependent ROC Curve and AUCs with Longitudinal Marker

In this section, we review extensions of the above estimators for longitudinally collected subject measurements. For instance some authors would assess the discrimination performance of CD4 counts repeatedly measured every week on time from seroconversion to progression to AIDS [41]. Therefore, time-dependent sensitivities and specificities have been extended to deal with the fact that (i) the time at which marker X is measured can vary and (ii) marker can be repeatedly measured on the same subject. Let s denote the timing of marker measurement and $X(s)$ the marker value at time s . For $t \geq s$, Zheng and Heagerty [41] extended *cumulative/dynamic* definitions

$$\text{TPR}^{\mathbb{C}}(c, s, t) = \mathbb{P}(X(s) > c | T \in [s, t]), \quad \text{FPR}^{\mathbb{D}}(c, s, t) = \mathbb{P}(X(s) > c | T > t).$$

For a fixed time $\tau \geq s$, Zheng and Heagerty [39] extended *incident/static* definitions:

$$\text{TPR}^{\mathbb{I}}(c, s, t) = \mathbb{P}(X(s) > c | T = t) \quad \text{and} \quad \text{FPR}^{\mathbb{S}}(c, s, \tau) = \mathbb{P}(X(s) > c | T > \tau).$$

Although other approaches have been proposed to estimate these quantities, we only review estimators that deal with censored data here. We should also mention that in this longitudinal context, estimators of the AUC are obtained by numerically integrating estimators of the ROC curve.

Cumulative Dynamic Estimators with Longitudinal Marker

Rizopoulos [27] recently proposed a joint modeling approach. The marker trajectory is modeled by usual linear mixed model for longitudinal data, and a parametric proportional hazard is used to model the time-to-event given the marker trajectory. The two submodels are linked with shared random effects to capture the intra-subject correlation. Standard maximum likelihood estimation is used to fit the joint model. Then, $\text{TPR}^{\mathbb{C}}$ and $\text{FPR}^{\mathbb{D}}$ are computed from the estimated parameters and Monte Carlo simulations are used to make inference. As this approach is fully parametric, its main advantage is its efficiency. This approach also allows censoring to depend on the marker [35]. The counterpart is that the parametric model must be carefully chosen, and checking model fit is not straightforward.

A more flexible methodology was proposed in [41], with fewer parametric assumptions. Setting $T^* = T - s$, the “residual failure time”, and $t^* = t - s$, they rewrote

$$\text{TPR}^{\mathbb{C}}(c, s, t^*) = \frac{1 - F_{X|s}(c) - S(c, t^*|s)}{1 - S(t^*|s)}, \quad \text{FPR}^{\mathbb{D}}(c, s, t^*) = 1 - \frac{S(c, t^*|s)}{S(t^*|s)},$$

with $F_{X|s}(c) = \mathbb{P}(X(s) < c | s, T^* > 0)$ the conditional distribution of marker given measurement time, $S(t^*|s) = \mathbb{P}(T^* > t^* | s, T^* > 0)$ the survival probability for individuals who survived beyond s and $S(c, t^*|s) = \mathbb{P}(X(s) > c, T^* > t^* | s, T^* > 0)$. They proposed to estimate $F_{X|s}(c)$ with the semiparametric estimator proposed by Heagerty and Pepe [16]. Therefore, only the location and scale of the conditional distribution of marker given measurement time are parametrized. To estimate the survival terms $S(c, t^*|s)$ and $S(t^*|s)$, they proposed the use of a “partly conditional” hazard function to model the residual failure time $T^* = T - s$. For subject i at measurement time s_{ik} , this function is modeled by

$$\lambda_{ik}(t^* | X_i(s_{ik}), 0 \leq s_{ik} \leq T_i) = \lambda_0(t^*) \exp [\beta(t^*)X_i(s_{ik}) + \alpha^T f(s_{ik})]$$

where $f(s)$ are vectors of spline basis functions evaluated at measurement time s , and $\lambda_0(t^*)$ is left unspecified. Estimators of $\beta(\cdot)$ and α have been previously proposed [40]. As this approach is semiparametric, its main advantage is its flexibility. However, by contrast to the approach of Rizopoulos [27], this one is less efficient and does not allow marker-dependent censoring.

Incident Static Estimators with Longitudinal Marker

Several authors consider the incident/static definition of AUC [3, 12, 31, 34]. However, censored data are only accounted for by Cai et al. [3] who proposed to model

$$\begin{aligned} \text{TPR}^{\mathbb{I}}(c, s, t) &= g_D(\eta_\alpha(t, s) + h(c)), \quad t \leq \tau \\ \text{FPR}^{\mathbb{S}}(c, s, \tau) &= g_\tau(\xi_a(s) + d(c)) \end{aligned}$$

where g_D and g_τ are specified inverse link functions and $h(\cdot)$ and $d(\cdot)$ are unspecified baseline functions of threshold c . The dependence in time is parametrically modeled by $\eta_\alpha(t, s) = \alpha^T \eta(t, s)$ and $\xi_a(s) = a^T \xi(s)$ where $\eta(t, s)$ and $\xi(s)$ are vectors of polynomial or spline basis functions. These models are semiparametric with respect to the marker distribution in cases and nonparametric in regards to controls. As pointed out in [25], this model is very flexible as it does not specify any distributional form for the distribution of the marker given the event-time, but only model the effect of time-to-event on the marker distribution with a parametric form. Model estimation is performed by solving some estimating equations and large sample theory was established allowing a resampling method to construct confidence bands and make inference [3]. Interestingly, the authors of [3] also showed that covariates can easily be included in $\text{TPR}^{\mathbb{I}}$ and $\text{FPR}^{\mathbb{S}}$, enabling to directly quantify how performances of the marker vary with these covariates.

Time-Dependent AUC and Competing Risks

We now consider the setting where a subject might experience multiple type of failures: in this section, we review extensions of time-dependent AUCs to competing risks. For example, we may want to assess the discrimination of a given score on death from prostate cancer with death from other causes acting as a competing event.

For the sake of simplicity, we will assume there are only two competing events, and we let $\delta_i = j$ denote that subject i experienced the competing event of type j ($j = 1, 2$, with $j = 1$ for the event of interest). The observed data consists of a failure time and a failure type (Z_i, δ_i) with $\delta_i = 0$ denoting a censored observation.

In their review paper [25], Pepet et al. sketched most potential extensions and introduced event-specific sensitivity and specificity. They also highlighted that the crucial point was to determine whether patients experiencing a competing event should be treated as a control when evaluating the discrimination of the marker under study with respect to the event of interest. More precisely, two settings can be considered.

First, if marker X is potentially discriminatory for both the event of interest and the competing event, then both event specific AUCs should be considered

simultaneously [13, 28]. For illustration, in the cumulative/dynamic setting, cases at time t can be stratified according to the event type, $\text{Case}_1 = \{i : T_i \leq t, \delta_i = 1\}$ and $\text{Case}_2 = \{i : T_i \leq t, \delta_i = 2\}$, and controls at time t are the event-free group at time t , $\text{Control} = \{i : T_i > t\}$. Following these lines Saha and Hearnerty [28] proposed event specific versions of (2)

$$\text{TPR}_j^{\mathbb{C}}(c, t) = \mathbb{P}(X > c | T \leq t, \delta = j), \quad \text{FPR}^{\mathbb{D}}(c, t) = \mathbb{P}(X > c | T > t, \delta \in \{1, 2\}). \quad (5)$$

Estimation follows from [18], using the conditional cumulative incidence associated to competing event j , $\mathbb{P}(T \leq t, \delta = j | X)$, instead of the conditional survival function of $\mathbb{P}(T \leq t | X)$. In the context of renal transplantation, Foucher et al. [13] considered a slight modification of definitions (5), where controls can also be “event specific”. In addition, an extension of the incident/dynamic AUC to the competing events setting was proposed by Saha and Hearnerty [28].

The other option is to consider both event-free patients and patients with the competing event [21, 42] as controls. For instance, dynamic controls at time t can be defined as the group $\{i : T_i > t\} \cup \{i : T_i \leq t, \delta_i = 2\}$. This leads to the estimation of only one ROC curve, for the event of interest. In [42], Zheng et al. based their approach on initial estimates of the conditional cumulative incidence function for the event of interest. Their initial method provides consistent estimators if the proportional hazard assumption holds for each cause specific hazard. To relax this assumption a smooth estimator was also proposed. Another approach was described in [21], which follows the lines of DeLong et al. [9]. However, the suitability of this method to deal with censored data is not established.

We shall add that, as pointed out in [28, 42], employing a direct regression model for the conditional cumulative incidence would lead to a simpler estimation of the cumulative/dynamic AUC and a less convoluted interpretation of the marker effect. However, the extension to the setting of a longitudinal marker [8] as well as the evaluation of a risk score (which is usually built with a cause-specific hazard approach) would not be straightforward.

Discussion

While the AUC is uniquely defined in the context of the evaluation of diagnostic tests, its extension to prognostic markers has led to the development of a variety of definitions: these definitions vary according to the underlying definitions of cases (incident or cumulative) and controls (static or dynamic), and also depend on the study characteristics (the marker can be measured only once or repeatedly and competing events may be considered, or not). Regarding the choice of the retained definition for cases and controls, no clear guidance has really emerged in the literature. It seems however that the cumulative/dynamic definition may be more appropriate for clinical decisions making (enrollment in clinical trials for instance) while the incident/static definition may be more appropriate for “pure” evaluation

of the marker (if interpretation of trends of AUC values over time is of particular interest for instance). Once this definition has been chosen, appropriate estimators are available, depending on various assumptions (independence of the marker and censoring, proportional hazards, . . .), and we presented most of them in this review article.

For the sake of brevity, we were not able to cover some interesting extensions of time-dependent AUCs. In particular, covariate specific time-dependent ROC curves and AUCs have been studied in order to adjust the discrimination of a marker for external covariates (age, gender, . . .). We refer the reader to [19, 32] for the standard setting, [3] for the longitudinal setting and [42] for the competing events setting. In addition, some authors advocate that not the entire ROC curve is of interest and the area under only a portion of it should be computed, leading to the so-called *partial AUC* [11]. In the context of prognostic markers, Hung and Chiang [20] proposed a nonparametric estimator of the cumulative/dynamic time-dependent version of the partial AUC. Other interesting extensions include diverse censoring patterns [22] (only right-censoring was considered in this review) and the combination of results from multiple studies [4] which is particularly useful in genomic studies.

Another closely related topic is the evaluation of the added predictive ability of a new marker: for instance, we may wonder how better a risk score would be if we added some biological markers (SNPs, genes, . . .). We refer the reader to the works in [6, 10, 23, 26] for some insights, noticing though that most of these works do not cover the right-censored setting considered in our review.

References

1. Akritas, M.: Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Stat.* **22**, 1299–1327 (1994)
2. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* **55**(5), 687–704. doi: 10.1002/bimj.201200045 (2013)
3. Cai, T., Pepe, M.S., Zheng, Y., Lumley, T., Jenny, N.S.: The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182–197 (2006)
4. Cai, T., Gerds, T.A., Zheng, Y., Chen, J.: Robust prediction of t-year survival with data from multiple studies. *Biometrics* **67**, 436–444 (2011)
5. Chambless, L.E., Diao, G.: Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat. Med.* **25**, 3474–3486 (2006)
6. Chambless, L.E., Cummiskey, C.P., Cui, G.: Several methods to assess improvement in risk prediction models: extension to survival analysis. *Stat. Med.* **30**(1), 22–38 (2011)
7. Chiang, C.T., Hung, H.: Nonparametric estimation for time-dependent AUC. *J. Stat. Plan. Inference* **140**(5), 1162–1174 (2010)
8. Cortese, G., Andersen, P.K.: Competing risks and time-dependent covariates. *Biom. J.* **52**(1), 138–158 (2010)
9. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845 (1988)

10. Demler, O.V., Pencina, M.J., D'Agostino, R.B. Sr.: Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. *Stat. Med.* **30**(12), 1410–1418 (2011)
11. Dodd, L.E., Pepe, M.S.: Partial AUC estimation and regression. *Biometrics* **59**(3), 614–623 (2003)
12. Etzioni, R., Pepe, M., Longton, G., Hu, C., Goodman, G.: Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med. Decis. Mak.* **19**(3), 242–251 (1999)
13. Foucher, Y., Giral, M., Soullou, J.P., Daures, J.P.: Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Stat. Med.* **29**, 3079–3087 (2010)
14. Gail, M.H., Pfeiffer, R.M.: On criteria for evaluating models of absolute risk. *Biostatistics* **6**(2), 227–239 (2005)
15. Gu, W., Pepe, M.S.: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **5**, 27–49 (2009)
16. Heagerty, P.J., Pepe, M.S.: Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *J. R. Stat. Soc. C (Appl. Stat.)* **48**(4), 533–551 (1999)
17. Heagerty, P.J., Zheng, Y.: Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105 (2005)
18. Heagerty, P.J., Lumley, T., Pepe, M.S.: Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344 (2000)
19. Hung, H., Chiang, C.T.: Estimation methods for time-dependent AUC models with survival data. *Can. J. Stat.* **38**(1), 8–26 (2010)
20. Hung, H., Chiang, C.T.: Nonparametric methodology for the time-dependent partial area under the ROC curve. *J. Stat. Plan. Inference* **141**(12), 3829–3838 (2011)
21. Lee, M., Cronin, K.A., Gail, M.H., Feuer, E.J.: Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Stat. Med.* **31**(5), 489–500 (2012)
22. Li, J., Ma, S.: Time-dependent ROC analysis under diverse censoring patterns. *Stat. Med.* **30**, 1266–1277 (2011)
23. Pencina, M.J., D'Agostino, R.B. Sr., D'Agostino, R.B. Jr., Vasan, R.S.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**(2), 157–172 (2008)
24. Pepe, M.S.: *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford (2004)
25. Pepe, M.S., Zheng, Y., Jin, Y., Huang, Y., Parikh, C.R., Levy, W.C.: Evaluating the ROC performance of markers for future events. *Lifetime Data Anal.* **14**, 86–113 (2008)
26. Pepe, M.S., Kerr, K.F., Longton, G., Wang, Z.: Testing for improvement in prediction model performance. Preprint available at <http://www.bepress.com/uwbiostat/paper379/> (2011)
27. Rizopoulos, D.: Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829 (2011)
28. Saha, P., Heagerty, P.J.: Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* **66**, 999–1011 (2010)
29. Saha-Chaudhuri, P., Heagerty, P.J.: Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics* **14**, 42–59 (2013)
30. Satten, G.A., Datta, S.: The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Am. Stat.* **55**(3), 207–210 (2001)
31. Slate, E.H., Turnbull, B.W.: Statistical models for longitudinal biomarkers of disease onset. *Stat. Med.* **19**(4), 617–637 (2000)
32. Song, X., Zhou, X.H.: A semi-parametric approach for the covariate specific ROC curve with survival outcome. *Stat. Sin.* **18**, 947–965 (2008)
33. Song, X., Zhou, X.H., Ma, S.: Nonparametric receiver operating characteristic-based evaluation for survival outcomes. *Stat. Med.* **31**(23), 2660–2675 (2012)

34. Subtil, F., Pouteil-Noble, C., Toussaint, S., Villar, E., Rabilloud, M.: A simple modeling-free method provides accurate estimates of sensitivity and specificity of longitudinal disease biomarkers. *Methods Inf. Med.* **48**, 299–305 (2009)
35. Tsiatis, A.A., Davidian, M.: Joint modeling of longitudinal and time-to-event data: an overview. *Stat. Sin.* **14**(3), 809–834 (2004)
36. Uno, H., Cai, T., Tian, L., Wei, L.J.: Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**(478), 527–537 (2007)
37. Viallon, V., Latouche, A.: Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve. *Biom. J.* **53**, 217–236 (2011)
38. Xu, R., O’Quigley, J.: Proportional hazards estimate of the conditional survival function. *J. R. Stat. Soc. B (Stat. Methodol.)* **62**(4), 667–680 (2000)
39. Zheng, Y., Heagerty, P.J.: Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**(4), 615–632 (2004)
40. Zheng, Y., Heagerty, P.J.: Partly conditional survival models for longitudinal data. *Biometrics* **61**(2), 379–391 (2005)
41. Zheng, Y., Heagerty, P.J.: Prospective accuracy for longitudinal markers. *Biometrics* **63**, 332–341 (2007)
42. Zheng, Y., Cai, T., Jin, Y., Feng, Z.: Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* **68**(2), 388–396 (2012)

Subgroup Specific Incremental Value of New Markers for Risk Prediction

Q. Zhou, Y. Zheng, and T. Cai

Abstract In many clinical applications, understanding when measurement of new markers is necessary to provide added accuracy to existing prediction tools could lead to more cost effective disease management. Many statistical tools for evaluating the incremental value of the novel markers over the routine clinical risk factors have been developed in recent years. However, most existing literature focuses primarily on global assessment. Since the incremental values of new markers often vary across subgroups, it would be of great interest to identify subgroups for which the new markers are most/least useful in improving risk prediction. In this paper we provide novel statistical procedures for *systematically* identifying potential traditional-marker based subgroups in whom it might be beneficial to apply a new model with measurements of both the novel and traditional markers. We consider various conditional time-dependent accuracy parameters for censored failure time outcome to assess the subgroup-specific incremental values. We provide nonparametric kernel-based estimation procedures to calculate the proposed parameters. Simultaneous interval estimation procedures are provided to account for sampling variation and adjust for multiple testing. Simulation studies suggest that our proposed procedures work well in finite samples. The proposed procedures

The paper appeared in volume 19 (2013) of Lifetime Data Analysis.

Q. Zhou (✉)

Simon Fraser University, Burnaby, BC V5A 1S6, Canada

e-mail: qmzhou@stat.sfu.ca

Y. Zheng

Fred Hutchinson Cancer Research Center, Seattle, Washington, DC 98109, USA

e-mail: yzheng@fhcrc.org

T. Cai

Harvard University, Boston, MA 02115, USA

e-mail: tcai@hsph.harvard.edu

are applied to the Framingham Offspring Study to examine the added value of an inflammation marker, C-reactive protein, on top of the traditional Framingham Risk Score for predicting 10-year risk of cardiovascular disease.

Introduction

Risk models have been applied in medical practice for prediction of long-term incidence or progression of many chronic diseases such as cardiovascular disease (CVD) and cancer. With the advancement in science and technology, a wide range of biological and genomic markers have now become available to assist in risk prediction. However, due to the potential financial and medical costs associated with measuring these markers, their ability in improving the prediction of disease outcomes and treatment response over existing risk models needs to be rigorously accessed.

Effective statistical tools for evaluating the incremental value (IncV) of the novel markers over the routine clinical risk factors are crucial in the field of outcome prediction. Many of newly discovered markers, while promising and strongly associated with clinical outcomes, may have limited capacity in improving risk prediction over and above routine clinical variables [43, 48]. For example, on top of traditional risk variables from the Framingham risk score (FRS) [52], the inflammation biomarker, C-Reactive Protein (CRP), was shown to provide modest prognostic information [3, 9, 37] while a genetic risk score consisting of 101 single nucleotide polymorphisms was reported as not useful [31]. In a recent paper, Wang et al. [50] concluded that almost all new contemporary biomarkers for prevention of coronary heart disease (CHD) added rather moderate *overall* predictive values to the FRS.

One possible explanation for the minimal improvement at the population average level is that the new markers may only be useful for certain subpopulations. For example, while much debate about the clinical utility of CRP remains, there is empirical evidence that CRP may substantially improve the prediction for subjects at intermediate risk [36]. Such finding, if valid, would be extremely useful in clinical practice, since identifying the subgroups where markers can provide valuable improvement in prediction will not only lead to more informed clinical decisions but also reduced cost and effort compared to measuring novel markers on the entire population. However, to ensure the validity of such claims and more precisely pinpoint such specific subgroups, rigorous and systematic analytical tools for IncV evaluation are needed.

To quantify the *global* IncV of new markers for risk prediction, various approaches have been advocated. For example with the most popular one being focused on a comparison of summary measures of accuracy under a conventional and new models respectively [4, 23, 44]. Excellent discussions on the choices of different accuracy measures can be found in Gail and Pfeiffer [20]. However, these measures quantify the overall IncV of new markers averaged over the

entire study population and do not provide information on how the IncV may vary across different groups of subjects. If there are pre-defined subgroups, these measures could be estimated for each of the subgroups. However, in practice, it is often unclear how to optimally select subgroups for comparisons and ad-hoc subgroup analyses without careful planning and execution may lead to invalid results [35, 39, 51]. Furthermore, it is vitally important to adjust for multiple comparisons when conducting any subgroup analysis. Thus, an important question is how to *systematically* identify the potential subgroups who would benefit from the additional markers properly adjusting for multiple comparisons. There is a paucity of statistical literature on approaches for identifying such subgroups [14]. Tian et al. [41] proposed an inference procedure to estimate the incremental values in absolute prediction error of new markers in various subgroups of patients classified by the conventional markers. However, their method does not incorporate censoring. In addition, the subgroups in their paper were defined as groups of subjects whose conventional risk scores lie in different pre-assigned intervals. However, how to determine the length of intervals could be an issue. Uno et al. [46] proposed estimation procedures for the conditional quantiles of the improvement in the predicted risk separately for the cases and the controls. However, they did not provide procedures for determining which subgroups should be recommended to have the new markers measured. Furthermore, no procedures were provided to account for the sampling variation or control overall type I error which is particularly important in subgroup analysis.

In this paper, we propose systematic approaches to analyzing censored event time data for identifying subgroups of patients for whom the new markers have the most or least IncV. We consider two common accuracy measures, the partial area under the ROC curve (pAUC) and the integrated discrimination improvement (IDI) index. Compared with the standard C-statistic, for many applications, the pAUC is often advocated as a better summary measure [5, 15, 17], since clinical interests often lie only in a specific range of the false positive rates (FPRs) or true positive rates (TPRs). For example, the region with low FPR is of more concern for disease screening [1]; while the region with high TPR is of more concern for the prognosis of serious disease [24]. However, the ROC curve does not capture certain aspects of the predicted absolute risk, since it is scale invariant. Many model performance measures, including the reclassification table [8], Net Reclassification Improvement (NRI) and Integrated Discrimination Improvement (IDI) [33], Proportion of case followed (PCF) and Proportion needed to follow-up (PNF) [34], have been proposed recently to overcome the limitation of the ROC curve method. Many of these measures, such as the reclassification table, NRI, PCF and PNF, rely on pre-specified clinically meaningful risk or quantile threshold values which may not be available for most diseases. For illustration purposes, we focus primarily on pAUC and IDI in this paper but note that our procedures can be easily extended to accommodate other accuracy measures.

The rest of paper is organized as follows. In section “Methods”, we present our proposed non-parametric estimation procedure for subgroup-specific IncV of new markers and along with their corresponding interval estimation procedures.

In particular, resampling based simultaneous interval estimation procedures are provided as convenient and effective tools to control for multiple comparisons. We describe results from our simulation studies in section “Simulation Studies” and the analyses of the Framingham Offspring Study using our proposed procedures in section “Example: The Framingham Offspring Study”. Concluding remarks are given in section “Concluding Remarks”. All the technical details are included in the appendices.

Methods

Risk Modeling with and Without New Markers

Let X denote a set of conventional markers and let Z denote a set of new markers. Due to censoring, for the event time T^\dagger , one can only observe $T = \min(T^\dagger, C)$, $\Delta = I(T^\dagger \leq C)$, where C is the censoring time, which is assumed to be independent of T^\dagger conditional on (X, Z) . See below for more discussions about censoring assumptions. Furthermore, define $Y^\dagger = I(T^\dagger \leq t_0)$, where t_0 is the prediction time of clinical interest, and $Y = I(T \leq t_0)$. Let $\mathcal{P}_1(X) = pr(Y^\dagger = 1|X)$ and $\mathcal{P}_2(X, Z) = pr(Y^\dagger = 1|X, Z)$ be the true conditional risk of developing the event by time t_0 conditional on X only and (X, Z) , respectively. Suppose a data set for analysis consists of n independent realizations of (T, Δ, X, Z) , $\{(T_i, \Delta_i, X_i, Z_i)\}$. Although Y^\dagger and the conditional risk functions depend on t_0 , we suppress t_0 from the notation throughout for the ease of presentation. From the Neyman-Pearson Lemma and similar arguments as given in McIntosh and Pepe [28], it is not difficult to show that $\mathcal{P}_1(X)$ achieves the optimal ROC curve for predicting Y^\dagger based on X only. Similarly, $\mathcal{P}_2(X, Z)$ is the optimal score for prediction Y^\dagger given (X, Z) .

To estimate $\mathcal{P}_1(X)$ and $\mathcal{P}_2(X, Z)$, one may consider a fully non-parametrical approach [27]. However, in practice, such nonparametric estimates may perform poorly when the dimension of X or Z is not small due to the curse of dimensionality [38]. An alternative feasible way is approximate $\mathcal{P}_1(\cdot)$ and $\mathcal{P}_2(\cdot)$ by imposing simple working models

$$pr(Y^\dagger = 1 | X) = g_1(\beta'V), \quad pr(Y^\dagger = 1 | X, Z) = g_2(\gamma'W), \tag{1}$$

where V , a $p \times 1$ vector, is a function of X , W , a $q \times 1$ vector, is a function of X and Z , β and γ are vectors of unknown regression parameters, and g_1 and g_2 are known, smooth, increasing functions. An estimator of β and γ can be obtained respectively by solving the following inverse probability weighted (IPW) estimating equations as given in Uno et al. [44]:

$$\sum_{i=1}^n \hat{\omega}_i V_i \{Y_i - g_1(\beta'V_i)\} = 0, \quad \sum_{i=1}^n \hat{\omega}_i W_i \{Y_i - g_2(\gamma'W_i)\} = 0. \tag{2}$$

where $\hat{\omega}_i = \Delta_i I(T_i \leq t_0) / \hat{G}_{X_i, Z_i}(T_i) + I(T_i > t_0) / \hat{G}_{X_i, Z_i}(t_0)$, and $\hat{G}_{X, Z}(t)$ is a root-n consistent estimator of $G_{X, Z}(t) = pr(C \geq t | X, Z)$. This IPW estimator may be justified heuristically using the argument that $E\{\hat{\omega}_i I(Y_i = y) | T_i^\dagger, X_i, Z_i\} \approx E\{I(Y_i^\dagger = y) | T_i^\dagger, X_i, Z_i\}$, for $y = 0, 1$. Let $\hat{\beta}$ and $\hat{\gamma}$ be the resulting estimator of β and γ , respectively. For a subject with $X = x, Z = z$ whose $V = v$ and $W = w$, the risk is estimated by $\hat{p}_1(x) = g_1(\hat{\beta}^\dagger v)$ based on X alone and by $\hat{p}_2(x, z) = g_2(\hat{\gamma}^\dagger w)$ based on X and Z . It has been previously shown in Uno et al. [44] that regardless of the adequacy of the working model (1), $\hat{\theta} = (\hat{\beta}^\dagger, \hat{\gamma}^\dagger)'$ converges in probability to a deterministic vector $\theta_0 = (\beta_0^\dagger, \gamma_0^\dagger)'$ as $n \rightarrow \infty$. Let $\bar{p}_1(x) = g_1(\beta_0^\dagger v)$ and $\bar{p}_2(x, z) = g_2(\gamma_0^\dagger w)$. When the models in (1) are correctly specified, $\bar{p}_1(x) = \mathcal{P}_1(x)$ and $\bar{p}_2(x, z) = \mathcal{P}_2(x, z)$. To obtain a valid estimator $\hat{G}_{X, Z}(\cdot)$, one may impose a semi-parametric model, such as the proportional hazards (PH) model [10], for $G_{X, Z}(t)$ and obtain $\hat{G}_{X, Z}(t)$ as $\exp\{-\hat{\Lambda}_0(t) \exp(\hat{\gamma}_c^\dagger W_c)\}$, where W_c is a function of (X, Z) , $\hat{\gamma}_c$ is the maximum partial likelihood estimator and $\hat{\Lambda}_0(t)$ is the Breslow's estimator. When the censoring is independent of both T and (X, Z) , one may obtain $\hat{G}_{X, Z}(\cdot)$ simply as the Kaplan-Meier estimator. It is important to note that if the models in (1) only hold for a given t_0 and the dimension of (X, Z) is not small, root-n consistent estimators of β and γ may not exist without imposing additional modeling assumptions about $G_{X, Z}(t)$ due to the curse of dimensionality [38].

Subgroup Specific Incremental Values

For illustration purposes, we consider two accuracy measures, the pAUC and the IDI index. We first define both concepts in the context of evaluating a risk score/model. Suppose that we use $\bar{p}_2(X, Z)$ as a risk score for classifying the event status Y^\dagger , and without loss of generality, we assume that a higher value of $\bar{p}_2(X, Z)$ is associated with a higher risk and refer to the two states, $Y^\dagger = 1$ and $Y^\dagger = 0$, as “diseased” and “disease-free” or “cases” and “controls”. The discrimination capacity of $\bar{p}_2(X, Z)$ can be quantified based on the ROC curve, which is a plot of the true positive rate (TPR) function, $\mathcal{S}_1(c) \equiv pr\{\bar{p}_2(X, Z) \geq c | Y^\dagger = 1\}$, against the false positive rate (FPR) function, $\mathcal{S}_0(c) \equiv pr\{\bar{p}_2(X, Z) \geq c | Y^\dagger = 0\}$. The ROC curve, $ROC(u) = \mathcal{S}_1\{\mathcal{S}_0^{-1}(u)\}$, describes the inherent capacity of distinguishing “cases” from “controls”. The pAUC with a restricted region of FPR, say $FPR \leq f$, is given by $pAUC_f = \int_0^f ROC(u) du$, for $f \in [0, 1]$. The IDI index, is simply $IDI = \int_0^1 \mathcal{S}_1(c) dc - \int_0^1 \mathcal{S}_0(c) dc$.

To evaluate how the IncV of Z may vary across subgroups defined by X , we define new conditional pAUC and IDI index. We propose to use $\bar{p}_1(x)$ as a scoring system for grouping subjects with potentially similar initial risk estimates and create subgroups $\mathcal{U}_s = \{X : \bar{p}_1(X) = s\}$. Then we evaluate the IncV of Z for each \mathcal{U}_s based on how well $\bar{p}_2(X, Z)$ can further discriminate subjects within \mathcal{U}_s with $Y^\dagger = 1$ from those with $Y^\dagger = 0$. Suppose $\bar{p}_2(X, Z)$ is used to classify Y^\dagger for subjects in \mathcal{U}_s .

The TPR and FPR of the classification rule $\bar{p}_2(X, Z) \geq c$ given \mathcal{U}_s are $\mathcal{S}_1(c; s)$ and $\mathcal{S}_0(c; s)$, respectively, where

$$\mathcal{S}_y(c; s) = pr \{ \bar{p}_2(X, Z) \geq c | Y^\dagger = y, \bar{p}_1(X) = s \}, \quad \text{and } c \in (0, 1), y = 0, 1.$$

Conditional on $\bar{p}_1(X) = s$, the ROC curve of $\bar{p}_2(X, Z)$ is $ROC(u; s) = \mathcal{S}_1\{\mathcal{S}_0^{-1}(u; s); s\}$, for $u \in [0, 1]$. The conditional pAUC is given by $pAUC_f(s) = \int_0^f ROC(u; s) du$, $f \in [0, 1]$. Note that $f = 1$ yields the conditional $AUC(s)$. If Z is non-informative for \mathcal{U}_s , the corresponding ROC curve would be a diagonal line, and we expect that $pAUC_s = f^2/2$, which is the area under a diagonal line. Thus, the subgroup \mathcal{U}_s specific IncV of Z with respect to (w.r.t.) the pAUC is given by $pAUC_f(s) - f^2/2$. The IDI index conditional on $\bar{p}_1(X) = s$ is given by

$$\begin{aligned} IDI(s) &= \int_0^1 \mathcal{S}_1(c; s) dc - \int_0^1 \mathcal{S}_0(c; s) dc \\ &= E \{ \bar{p}_2(X, Z) | Y^\dagger = 1, \bar{p}_1(X) = s \} - E \{ \bar{p}_2(X, Z) | Y^\dagger = 0, \bar{p}_1(X) = s \}. \end{aligned} \tag{3}$$

If Z is non-informative for this subgroup \mathcal{U}_s , the conditional IDI index would be 0, and therefore, the subgroup \mathcal{U}_s specific IncV of Z w.r.t. the IDI index is $IDI(s)$. Based on these subgroup-specific IncVs, we are able to identify the set of s such that Z is useful to improve the prediction accuracy for \mathcal{U}_s , which is referred to as the *effective subpopulation* \mathcal{U}^* in our paper. Specifically, the effective subpopulation w.r.t. pAUC is defined as $\mathcal{U}^* = \{X : pAUC_f(\bar{p}_1(X)) - f^2/2 > c_1\}$; the effective subpopulation w.r.t. IDI index are defined as $\mathcal{U}^* = \{X : IDI(\bar{p}_1(X)) > c_2\}$, where c_1 and c_2 are some possibly data dependent threshold values. For the subjects in the effective subpopulation, measurement of new markers would provide added accuracy to the conventional risk model.

Inference About Subgroup-Specific Incremental Values

We first discuss the estimation for the conditional TPR and FPR functions $\{\mathcal{S}_y(c; s), y = 0, 1\}$ since both $pAUC_f(s)$ and $IDI(s)$ are simple functionals of these two functions. Let $\hat{p}_{1i} = \hat{p}_1(X_i) = g_1(\hat{\beta}'V_i)$ and $\hat{p}_{2i} = \hat{p}_2(X_i, Z_i) = g_2(\hat{\gamma}'W_i)$. To obtain a consistent estimator of $\mathcal{S}_y(c; s)$, since $\mathcal{S}_y(c; s)$ is between 0 and 1, we consider a non-parametric local likelihood estimation method [42] along with IPW accounting for censoring. Specifically, we obtain $\{\hat{a}_{y, h_y}(c; s), \hat{b}_{y, h_y}(c; s)\}$ as the solution to the following IPW local likelihood score equation,

$$\sum_{i: Y_i=y} \hat{\omega}_i \left[h_y^{-1} \hat{\mathcal{E}}_{1i}(s) \right] K_{h_y} \left\{ \hat{\mathcal{E}}_{1i}(s) \right\} \left[I(\hat{p}_{2i} \geq c) - g \left\{ a + b \hat{\mathcal{E}}_{1i}(s) \right\} \right] = 0, \tag{4}$$

where $\hat{\mathcal{E}}_{1i}(s) = \phi(\hat{\rho}_{1i}) - \phi(s)$, $g(x) = \exp(x) / \{1 + \exp(x)\}$, $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a known smooth symmetric kernel density function with a bounded support, and the bandwidth $h_y > 0$ is assumed to be $O(n^\nu)$, for $1/5 < \nu < 1/2$, and $\phi(\cdot)$ is a known, non-decreasing transformation function that can potentially be helpful in improving the performance of the smoothed estimator [30, 49]. Then, $\mathcal{S}_y(c; s)$ can be estimated by $\hat{\mathcal{S}}_{y, h_y}(c; s) = g\{\hat{a}_{y, h_y}(c; s)\}$ for $y = 0, 1$. In the section ‘‘Asymptotic Expansions for $\hat{\mathcal{S}}_y(c; s)$ ’’ in Appendix 1, we show that $\hat{\mathcal{S}}_{y, h_y}(c; s) - \mathcal{S}_y(c; s) \rightarrow 0$ in probability as $n \rightarrow \infty$, uniformly in $c \in [0, 1]$ and $s \in \mathcal{I}_{h_y} \equiv [\phi^{-1}(\rho_l + h_y), \phi^{-1}(\rho_u - h_y)]$, where $[\rho_l, \rho_u]$ is a subset of the support of $\phi\{g_1(\beta_0^*V)\}$ and β_0 is the limit of $\hat{\beta}$. As a special case, by setting b in (4) to 0, one may obtain a local constant estimator,

$$\hat{\mathcal{S}}_{y, h_y}(c; s) = \frac{\sum_{i=1}^n \hat{\omega}_i I(Y_i = y) K_{h_y}\{\hat{\mathcal{E}}_{1i}(s)\} I(\hat{\rho}_{2i} \geq c)}{\sum_{i=1}^n \hat{\omega}_i I(Y_i = y) K_{h_y}\{\hat{\mathcal{E}}_{1i}(s)\}}, y = 0, 1.$$

Inference procedures for $\text{pAUC}_f(s)$ Based on $\hat{\mathcal{S}}_{y, h_y}(c; s)$, $\text{pAUC}_f(s)$ can be estimated as

$$\widehat{\text{pAUC}}_f(s) = \int_0^f \widehat{\text{ROC}}(u; s) du = \int_{\hat{\mathcal{S}}_{0, h_0}^{-1}(f; s)}^\infty \hat{\mathcal{S}}_{1, h_1}(c; s) d\{1 - \hat{\mathcal{S}}_{0, h_0}(c; s)\}.$$

where $\widehat{\text{ROC}}(u; s) = \hat{\mathcal{S}}_{1, h_1}\{\hat{\mathcal{S}}_{0, h_0}^{-1}(u; s); s\}$ and (h_0, h_1) is the pair of optimal bandwidths for estimating $\mathcal{S}_0(c; s)$ and $\mathcal{S}_1(c; s)$, respectively. In the section ‘‘Uniform Consistency of $\widehat{\text{pAUC}}_f(s)$ ’’ in Appendix 1, we show that $\widehat{\text{pAUC}}_f(s)$ is uniformly consistent for $\text{pAUC}_f(s)$.

As a special case, when both X and Z are univariate, the ROC curve of $\bar{p}_2(X, Z)$ conditional on $\bar{p}_1(X)$ is equivalent to the ROC curve of Z conditional on X since the ROC curve is scale invariant. A simple local constant IPW estimator of $\mathcal{S}_y(z; x)$ is given by

$$\hat{\mathcal{S}}_{y, h_y}(z; x) = \frac{\sum_{i=1}^n \hat{\omega}_i I(Y_i = y) K_{h_y}(X_i - x) I(Z_i \geq z)}{\sum_{i=1}^n \hat{\omega}_i I(Y_i = y) K_{h_y}(X_i - x)}.$$

The resulting estimator of $\text{pAUC}_f(x)$ is

$$\begin{aligned} \widehat{\text{pAUC}}_f(x) &= \int_0^f \hat{\mathcal{S}}_{1, h_1}\{\hat{\mathcal{S}}_{0, h_0}^{-1}(u; x); x\} du \\ &= \frac{\sum_{i=1}^n \hat{\omega}_i Y_i K_{h_1}(X_i - x) U_i(x; f, h_0)}{\sum_{i=1}^n \hat{\omega}_i Y_i K_{h_1}(X_i - x)}, \end{aligned}$$

where $U_i(x; f, h_0) = f - \min\{f, \hat{\mathcal{S}}_{0, h_0}(Z_i; x)\}$ is the estimated truncated placement value proposed by Cai and Dodd [5].

It is difficult to directly estimate the variance of $\mathcal{W}_{\widehat{\text{pAUC}}_f}(s) = \sqrt{nh_1}\{\widehat{\text{pAUC}}_f(s) - \text{pAUC}_f(s)\}$ since it involves unknown derivative functions. We propose a perturbation resampling method to approximate the distribution of $\mathcal{W}_{\widehat{\text{pAUC}}_f}(s)$. This method has been widely used in survival analyses (see for example, [6, 25, 29]). To be specific, let $\Xi = \{\xi_i, i = 1, \dots, n\}$ be n independent positive random variables following a known distribution with mean 1 and variance 1, and Ξ is independent of the data. For each set of Ξ , we first obtain β^* and γ^* , as the respective solutions to

$$\sum_{i=1}^n \omega_i^* V_i \{Y_i - g_1(\beta^* V_i)\} \xi_i = 0, \quad \text{and} \quad \sum_{i=1}^n \omega_i^* W_i \{Y_i - g_2(\gamma^* W_i)\} \xi_i = 0,$$

where $\omega_i^* = \Delta I(T_i \leq t_0) / G_{X_i, Z_i}^*(T_i) + I(T_i > t_0) / G_{X_i, Z_i}^*(t_0)$ and $G_{X, Z}^*$ is the perturbed estimator of $G_{X, Z}(\cdot)$ with Ξ being the weights. Let $p_{1i}^* = g_1(\beta^* V_i)$, $\mathcal{E}_{1i}^*(s) = \phi(p_{1i}^*) - \phi(s)$, $p_{2i}^* = g_2(\gamma^* W_i)$, and $M_i^*(c) = I(p_{2i}^* \geq c)$. Subsequently, we obtain the perturbed counterpart of $\mathcal{S}_y(c; s)$ as $\mathcal{S}_{y, h_y}^*(c; s) = g\{a_{y, h_y}^*(c; s)\}$, where $a_{y, h_y}^*(c; s)$ is the solution to the perturbed score equation

$$\sum_{i=1}^n \omega_i^* I(Y_i = y) \left[h_y^{-1} \frac{1}{\mathcal{E}_{1i}^*(s)} \right] K_{h_y} \{ \mathcal{E}_{1i}^*(s) \} [M_i^*(c) - g\{a + b \mathcal{E}_{1i}^*(s)\}] \xi_i = 0.$$

Then, the perturbed pAUC is given by, $\text{pAUC}_f^*(s) = \int_0^f \text{ROC}^*(u; s) du$, where

$$\text{ROC}^*(u; s) = \mathcal{S}_{1, h_1}^* \left\{ \mathcal{S}_{0, h_0}^{*-1}(u; s); s \right\}.$$

In the section ‘‘Asymptotic Distribution of $\mathcal{W}_{\widehat{\text{pAUC}}_f}(s)$ ’’ in Appendix 1, we show that the unconditional distribution of $\mathcal{W}_{\widehat{\text{pAUC}}_f}(s)$ can be approximated by the conditional distribution of

$$\mathcal{W}_{\widehat{\text{pAUC}}_f}^*(s) = \sqrt{nh_1} \left\{ \text{pAUC}_f^*(s) - \widehat{\text{pAUC}}_f(s) \right\}, \tag{5}$$

given the data. With the above resampling method, for any fixed $s \in \mathcal{S}_{h_1}$, one may obtain a variance estimator of $\mathcal{W}_{\widehat{\text{pAUC}}_f}(s)$, $\hat{\sigma}_f^2(s)$, based on the empirical variance of B realizations from (5). For any fixed $s \in \mathcal{S}_{h_1}$ and $\alpha \in (0, 1)$, a pointwise $100(1 - \alpha)\%$ confidence interval (CI) for $\text{pAUC}_f(s)$ can be constructed via $\widehat{\text{pAUC}}_f(s) \pm (nh_1)^{-1/2} c_\alpha \hat{\sigma}_f(s)$, where c_α is the $100(1 - \alpha)$ th percentile of the standard normal distribution.

Inference for IDI(s) Based on $\mathcal{S}_{y, h_y}^*(c; s)$, we may obtain plug-in estimators for $\text{IS}(s) = \int_0^1 \mathcal{S}_1(c; s) dc$ and $\text{IP}(s) = \int_0^1 \mathcal{S}_0(c; s) dc$ respectively as

$$\widehat{\text{IS}}(s) = \int_0^1 \hat{\mathcal{S}}_{1, h_1}(c; s) dc, \quad \text{and} \quad \widehat{\text{IP}}(s) = \int_0^1 \hat{\mathcal{S}}_{0, h_0}(c; s) dc.$$

Thus, $\text{IDI}(s)$ can be estimated by $\widehat{\text{IDI}}(s) = \widehat{\text{IS}}(s) - \widehat{\text{IP}}(s)$. Similar to the derivations given in the Appendix 1 for $\widehat{\text{pAUC}}_f(s)$, the asymptotic results for $\widehat{\mathcal{S}}_{y,h_y}(c;s)$ can be directly used to establish the consistency and asymptotic normality for $\widehat{\text{IDI}}(s)$. In addition, the unconditional distribution of $\widehat{\mathcal{W}}_{\text{IDI}}(s) = \sqrt{nh_1}\{\widehat{\text{IDI}}(s) - \text{IDI}(s)\}$ can be approximated by the conditional distribution of $\mathcal{W}_{\text{IDI}}^*(s) = \sqrt{nh_1}\{\text{IDI}^*(s) - \widehat{\text{IDI}}(s)\}$, given the data, where $\text{IDI}^*(s) = \int_0^1 \mathcal{S}_{1,h_1}^*(c;s)dc - \int_0^1 \mathcal{S}_{0,h_0}(c;s)dc$ and $\mathcal{S}_{y,h_y}^*(c;s)$ is the perturbed counterpart of $\widehat{\mathcal{S}}_{y,h_y}(c;s)$, $y = 0, 1$. The pointwise confidence intervals for any fixed $s \in \mathcal{S}_{h_1}$ are constructed in a similar way as the inference for $\text{pAUC}_f(s)$. As a special case, a kernel local constant estimator of $\text{IDI}(s)$ is given by

$$\widehat{\text{IDI}}(s) = \frac{\sum_{i=1}^n \widehat{\omega}_i Y_i K_{h_1}\{\widehat{\mathcal{E}}_{1i}(s)\} \widehat{p}_{2i}}{\sum_{i=1}^n \widehat{\omega}_i Y_i K_{h_1}\{\widehat{\mathcal{E}}_{1i}(s)\}} - \frac{\sum_{i=1}^n \widehat{\omega}_i (1 - Y_i) K_{h_0}\{\widehat{\mathcal{E}}_{1i}(s)\} \widehat{p}_{2i}}{\sum_{i=1}^n \widehat{\omega}_i (1 - Y_i) K_{h_0}\{\widehat{\mathcal{E}}_{1i}(s)\}},$$

with the perturbed counterpart given by

$$\text{IDI}^*(s) = \frac{\sum_{i=1}^n \omega_i^* Y_i K_{h_1}\{\mathcal{E}_{1i}^*(s)\} p_{2i}^* \xi_i}{\sum_{i=1}^n \omega_i^* Y_i K_{h_1}\{\mathcal{E}_{1i}^*(s)\} \xi_i} - \frac{\sum_{i=1}^n \omega_i^* (1 - Y_i) K_{h_0}\{\mathcal{E}_{1i}^*(s)\} p_{2i}^* \xi_i}{\sum_{i=1}^n \omega_i^* (1 - Y_i) K_{h_0}\{\mathcal{E}_{1i}^*(s)\} \xi_i}.$$

Selection of the optimal bandwidths for $\text{pAUC}_f(s)$ and $\text{IDI}(s)$ is illustrated in the Appendix 2.

Identifying the effective subpopulation To identify the effective subpopulation, one may simultaneously assess the subgroup-specific IncV w.r.t. a certain accuracy measure, denoted by $\mathbb{A}(s)$, for example $\text{pAUC}_f(s) - f^2/2$ or $\text{IDI}(s)$, over a range of s values by constructing simultaneous CI for $\{\mathbb{A}(s), s \in \mathcal{S}_{h_1}\}$. Unfortunately, the distribution of $\widehat{\mathcal{W}}_{\mathbb{A}}(s)$ does not converge as a process in s , as $n \rightarrow \infty$. Thus, we cannot apply the standard large sample theory for stochastic processes to approximate the distribution of $\widehat{\mathcal{W}}_{\mathbb{A}}(s)$. Nevertheless, by the strong approximation argument and extreme value limit theorem [2], we show in the section ‘‘Asymptotic distribution of $\widehat{\mathcal{W}}_{\text{pAUC}_f}(s)$ ’’ in Appendix 1 that a standardized version of the sup-statistic $\Gamma = \sup_{s \in \mathcal{S}_{h_1}} |\widehat{\mathcal{W}}_{\mathbb{A}}(s) / \widehat{\sigma}_{\mathbb{A}}(s)|$ converges in distribution to a proper random variable, where $\widehat{\sigma}_{\mathbb{A}}^2$ denotes the variance estimator of $\widehat{\mathcal{W}}_{\mathbb{A}}(s)$. In practice, for large n , one can approximate the distribution of Γ based on realizations of $\Gamma^* = \sup_{s \in \mathcal{S}_{h_1}} |\mathcal{W}_{\mathbb{A}}^*(s) / \widehat{\sigma}_{\mathbb{A}}(s)|$, where $\mathcal{W}_{\mathbb{A}}^*$ is the perturbed counterpart of $\widehat{\mathcal{W}}_{\mathbb{A}}$. Therefore, a $100(1 - \alpha)\%$ simultaneous CI for $\mathbb{A}(s)$ can be obtained as $\widehat{\mathbb{A}}(s) \pm (nh_1)^{-1/2} d_\alpha \widehat{\sigma}_{\mathbb{A}}(s)$, where d_α is the empirical $100(1 - \alpha)$ th quantile of Γ^* . Thus, to account for sampling variation and multiple testing, the effective subpopulation is chosen as $\{X : \widehat{\mathbb{A}}(\widehat{p}_1(X)) > (nh_1)^{-1/2} d_\alpha \widehat{\sigma}_{\mathbb{A}}(\widehat{p}_1(X))\}$ in real data analyses.

Test for heterogeneous IncV Another question of interest is whether the subgroup-specific IncV $\mathbb{A}(s)$, for example $\text{pAUC}_f(s)$, is constant across different values of s over a certain interval $[s_l, s_u]$. We define the average IncV over $[s_l, s_u]$ as

$$\mathbb{A}_{[s_l, s_u]} = \frac{\int_{s_l}^{s_u} \mathbb{A}(s) d\mathcal{F}(s)}{[\mathcal{F}(s_u) - \mathcal{F}(s_l)]}$$

where $\mathcal{F}(s) = pr\{\bar{p}_1(X) \leq s\}$, and we define the relative subgroup-specific IncV over $[s_l, s_u]$ as $\mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}(s) = \mathbb{A}(s) - \mathbb{A}_{[s_l, s_u]}$. The point estimate of $\mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}(s)$ is given by

$$\hat{\mathbb{D}}_{\mathbb{A}_{[s_l, s_u]}}(s) = \hat{\mathbb{A}}(s) - \hat{\mathbb{A}}_{[s_l, s_u]}, \quad \hat{\mathbb{A}}_{[s_l, s_u]} = \frac{n^{-1} \sum_{i=1}^n \hat{\mathbb{A}}(\hat{p}_{1i}) I(\hat{p}_{1i} \in [s_l, s_u])}{\hat{\mathcal{F}}(s_u) - \hat{\mathcal{F}}(s_l)}$$

where $\hat{\mathcal{F}}(s) = n^{-1} \sum_{i=1}^n I\{\hat{p}_{1i} \leq s\}$. In addition, the unconditional distribution of $\mathcal{W}_{\mathbb{D}}(s) = \sqrt{nh_1}\{\hat{\mathbb{D}}_{\mathbb{A}_{[s_l, s_u]}}(s) - \mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}(s)\}$ can be approximated by the conditional distribution of $\mathcal{W}_{\mathbb{D}}^*(s) = \sqrt{nh_1}\{\mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}^*(s) - \hat{\mathbb{D}}_{\mathbb{A}_{[s_l, s_u]}}(s)\}$ given the data, where $\mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}^*(s) = \mathbb{A}^*(s) - \mathbb{A}_{[s_l, s_u]}^*$ with $\mathbb{A}^*(s)$ as the perturbed counterpart of $\hat{\mathbb{A}}(s)$ and $\mathbb{A}_{[s_l, s_u]}^* = n^{-1} \sum_{i=1}^n \mathbb{A}^*(\hat{p}_{1i}) I(\hat{p}_{1i} \in [s_l, s_u]) / [\hat{\mathcal{F}}(s_u) - \hat{\mathcal{F}}(s_l)]$. The variance estimator $\hat{\sigma}_{\mathbb{D}}^2$ of $\mathcal{W}_{\mathbb{D}}(s)$ can be obtained from realizations of $\mathcal{W}_{\mathbb{D}}^*(s)$.

If the subgroup-specific IncV of Z is constant over $[s_l, s_u]$, i.e., $\mathbb{A}(s) \equiv c_0$ for $s \in [s_l, s_u]$, $\mathbb{A}_{[s_l, s_u]} = c_0$ and $\mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}(s) = 0$ for $s \in [s_l, s_u]$. Testing whether the subgroup specific IncV is constant over $[s_l, s_u]$ is equivalent to testing the null hypothesis $H_0 : \mathbb{D}_{\mathbb{A}_{[s_l, s_u]}}(s) = 0$ for $s \in [s_l, s_u]$. To adjust for multiple testing, we consider the standard version of the sup-statistic $\Gamma_{\mathbb{D}} = \sup_{s \in [s_l, s_u]} |\mathcal{W}_{\mathbb{D}}^{(0)}(s) / \hat{\sigma}_{\mathbb{D}}(s)|$, where $\mathcal{W}_{\mathbb{D}}^{(0)}(s) = \sqrt{nh_1} \hat{\mathbb{D}}_{\mathbb{A}_{[s_l, s_u]}}(s)$ is the statistic $\mathcal{W}_{\mathbb{D}}(s)$ under the null hypothesis H_0 . One may approximate the distribution of $\Gamma_{\mathbb{D}}$ based on realizations of $\Gamma_{\mathbb{D}}^* = \sup_{s \in [s_l, s_u]} |\mathcal{W}_{\mathbb{D}}^*(s) / \hat{\sigma}_{\mathbb{D}}(s)|$. The empirical p -value for testing the null hypothesis H_0 can be obtained by $B^{-1} \sum_{b=1}^B I\{\Gamma_{\mathbb{D}}^{*(b)} > \Gamma_{\mathbb{D}}\}$, where $\{\Gamma_{\mathbb{D}}^{*(b)}, b = 1, \dots, B\}$ are B realizations of $\Gamma_{\mathbb{D}}^*$.

Simulation Studies

To examine the finite sample properties of the proposed estimation procedure, we conduct a simulation study where the conventional marker X and the new marker Z are both univariate and jointly generated from a bivariate normal distribution

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim BVN \left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Z \rho_{XZ} \\ \sigma_X \sigma_Z \rho_{XZ} & \sigma_Z^2 \end{bmatrix} \right).$$

In this simulation study, $\mu_X = \mu_Z = 0$, $\sigma_X = 2$ and $\sigma_Z = 0.5$, and $\rho_{XZ} = 0.01$. The failure time T , given the markers X and Z , is generated from an accelerated failure time model with a log-normal distribution for T , i.e., $\log T = h(X, Z) + \varepsilon$, where ε is a normal random variable with mean 0 and standard deviation $\sigma_T = 1.5$. In this simulation study, $h(X, Z)$ is a linear model, i.e., $h(X, Z) = -\beta_X X - \beta_Z Z - \beta_{XZ} XZ$. We consider a practical situation where the new marker Z may make a major contribution to the underlying mechanism in contrast with the conventional marker

X , although it may not be measured routinely. Thus, in this simulation study, we set $\beta_X = 0.01$ and $\beta_Z = \beta_{XZ} = 1$. The censoring time C is generated from an exponential distribution with rate c_0^{-1} . A value of $c_0 \approx 20$ is chosen such that roughly 80% of the failure time is censored. A time point $t_0 \approx 0.2$ is set such that the proportion of the “cases” in the sample is approximately 20%.

We investigate the kernel local constant estimator for the conditional pAUC_f with $f = 0.1$ representing a low FPR region and $f = 1$ representing the standard AUC. Since Z and $\log T$ are jointly normal conditional on $X = x$, it is straightforward to calculate the true values of $\text{pAUC}_f(x)$. We consider a relatively smaller sample size 1000, a moderate sample size of 5000 and a relatively larger sample size of 10000. Both of the pAUC with $\text{FPR} \leq 0.1$, i.e., $\text{pAUC}_{0.1}(x)$ and the full AUC are estimated at a sequence of values of X . For ease of computation, the pair of the bandwidths (h_0, h_1) for constructing the nonparametric estimate was fixed at (i) for the full AUC, (2.531, 2.102) for $n = 1000$, (1.905, 1.534) for $n = 5000$, and (1.640, 1.380) for $n = 10000$; (ii) for the $\text{pAUC}_{0.1}$, (2.377, 2.432) for $n = 1000$, (1.843, 2.361) for $n = 5000$, and (1.507, 2.085) for $n = 10000$. Here, $(h_0^{\text{opt}}, h_1^{\text{opt}})$ were chosen as the average of the bandwidths selected from 10 independent simulated datasets using the two-stage of five-fold cross-validation method described in the Appendix 2 and $n^{-0.1}$ was multiplied to h_y^{opt} to yield the final bandwidths used for simulation. In addition, the kernel function $K(\cdot)$ was chosen as the Epanechnikov kernel. Here, since we assume that the censoring time C is independent of both T and (X, Z) , $G_{X,Z}(t) = G(t)$ is estimated by a Kaplan-Meier estimator.

The performance of the point estimates and pointwise 95% confidence intervals obtained by the resampling method was assessed from 1000 independent replicates. For all of these scenarios, the nonparametric estimators have substantially small biases, the estimated standard errors are close to their empirical counterparts, and empirical coverage levels are close to the nominal level. In Fig. 1, we summarize the performance of the point and interval estimates for $\text{pAUC}_{0.1}$ with sample size 10000. For this scenario, the empirical coverage probabilities of the 95% pointwise confidence intervals range from 92.9 to 95.4%. The empirical coverage levels of the 95% simultaneous confidence bands for the standard AUC are 93.2% for $n = 1000$, 93.3% for $n = 5000$, and 94.5% for $n = 10000$; the empirical coverage levels of the 95% simultaneous confidence bands for the $\text{pAUC}_{0.1}$ are 93.3% for $n = 1000$, 93.4% for $n = 5000$, and 92.5% for $n = 10000$.

Example: The Framingham Offspring Study

The Framingham Offspring Study was established in 1971 with 5124 participants who were monitored prospectively on epidemiological and genetic risk factors of CVD. Here, we use data from 1687 female participants of which 261 have either died or experienced a CVD event by the end of follow-up period, and the 10-year event rate is 6%. The Framingham risk model, based on several clinical risk factors including age, systolic blood pressure (SBP), diastolic blood pressure (DBP), total

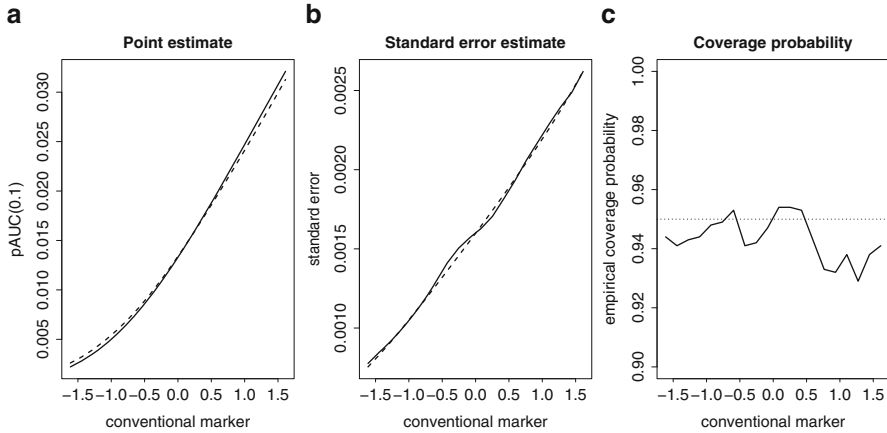


Fig. 1 Performance of the point estimates, the standard error estimates and pointwise confidence intervals for $pAUC_{0.1}$ with sample size 10000: (a) the true $pAUC_{0.1}(x)$ (solid) and the average point estimates (dashed) over 1000 replicates; (b) the empirical standard error estimates (solid) and the average of the estimated errors (dashed) based on the resampling procedure; and (c) the empirical coverage levels of the pointwise 95% confidence intervals obtained from the resampling procedures

cholesterol(TC), high-density lipoprotein (HDL) cholesterol, current smoking status and diabetes, is widely used in clinical settings but only with moderate accuracy for predicting the 10-year risk of CVD [9]. The Framing risk score (FRS) is constructed as the weighted average of the risk factors in the Framingham risk model using β -coefficients given in Table 6 of [52]. The risk estimates are obtained from the FRS through the transformation $1 - \exp\{-\exp(\cdot)\}$. The density plot of the risk estimates obtained from the FRS is shown in Fig. 2a. The overall gain in C-statistic by adding the CRP on top of FRS is 0.002 (from 0.776 to 0.778, with 95% CI $(-0.005, 0.01)$). Note that a log transformation is applied on the CRP throughout the analysis. According to the Framingham risk model [52] and the risk threshold values employed by the Adult Treatment Panel III (ATP III) of the National Cholesterol Education Program [18], these 1687 female participants may be classified into three risk groups: 1462 as low risk ($<10\%$); 193 as intermediate risk (between 10 and 20%); 32 as high risk ($>20\%$). The IncVs w.r.t. C-statistic are 0.00057 (with 95% CI $(-0.012, 0.013)$) for the low risk group; 0.037 (with 95% CI $(-0.054, 0.13)$) for the intermediate risk group; 0.034 (with 95% CI $(-0.097, 0.16)$) for the high risk group. Note that the low risk group consists of about 87% of the entire cohort. Now we further classify the 1462 patients of the low risk group into 10 finer subgroups with the length of the risk interval for each subgroup being 0.01, for example, 0–0.01, 0.01–0.02, and etc. The IncVs w.r.t. C-statistics for these 10 subgroups of low risk as well as the intermediate and high risk groups with their 95% CIs are shown in Fig. 2b. This suggests that adding CRP on top of FRS may be most useful for the risk groups around 5%, which is also referred to as the intermedium low risk group in some literature.

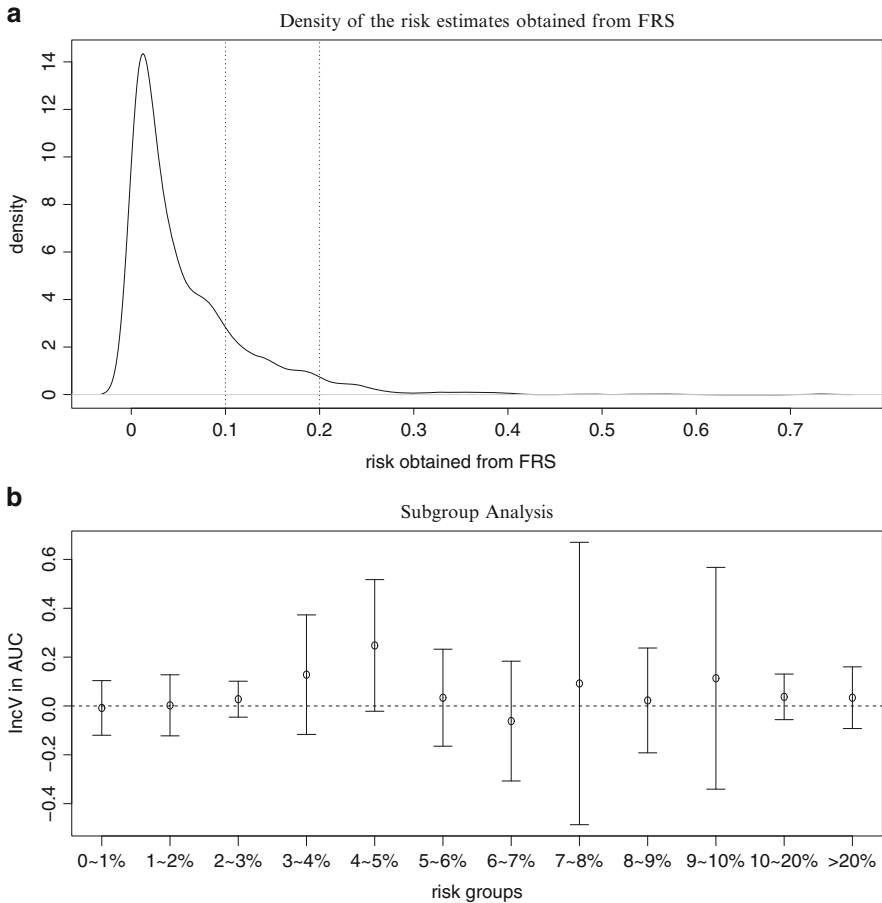


Fig. 2 (a) The density estimates of the 10-year event risk calculated from the Framingham risk score. (b) The IncVs w.r.t. C-statistics for the 10 subgroups of low risk as well as the intermediate and high risk groups with their 95% CIs

First, we investigate the IncV of the CRP over the FRS w.r.t. AUC, $pAUC_{0,1}$ and IDI in predicting the 10-year risk of CVD events among subgroups defined by the FRS. For the purpose of kernel smoothing, the transformation function $\phi(\cdot)$ in the local likelihood score equation (4) is $\phi(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$, where $\mu_X = -3.74$ is the sample mean of the FRS and $\sigma_X = 1.35$ is the sample standard deviation of the FRS, and $\Phi(x)$ is the cumulative distribution function of a standard normal distribution. Here we use local kernel constant estimates with Epanechnikov kernel. The optimal bandwidths (h_0^{opt}, h_1^{opt}) in ϕ -scale are chosen via a 10-fold cross validation procedure: (0.117, 0.393) for the standard AUC, (0.264, 0.721) for $pAUC_{0,1}$, and (0.018, 0.273) for IDI. The point estimates along with the 95% pointwise and simultaneous confidence intervals for the subgroup-specific IncV

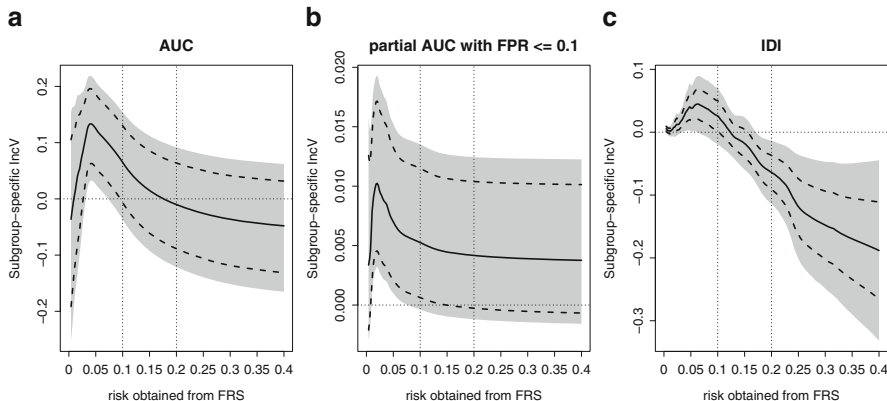


Fig. 3 The point estimates (*solid line*), and its 95% pointwise confidence intervals (*dashed lines*) and the 95% simultaneous confidence bands (*dark shaded region*) for (a) the subgroup-specific IncV with respect to AUC, $AUC(x) - 1/2$; (b) the subgroup-specific IncV with respect to $pAUC_{0.1}$, $pAUC_{0.1}(x) - 0.1^2/2$; (c) the subgroup-specific IncV with respect to IDI. The two vertical dotted lines represent the risk category cut-offs, 10% and 20%, from left to right

w.r.t. AUC, $pAUC_{0.1}$ and IDI are shown in Fig. 3. The point estimate for IDI is obtained via a cross-validation procedure to correct for biases due to overfitting. Based on the pointwise CIs of the subgroup-specific IncV w.r.t. AUC, the addition of CRP appears to improve risk prediction for subjects with the FRS risk ranging from 0.028 to 0.096. The corresponding range is 0.008–0.148 when based on the CIs for the subgroup-specific IncV w.r.t. $pAUC_{0.1}$; 0.004–0.102 when based on the CIs for the subgroup-specific IncV w.r.t. IDI. After controlling for the overall type I error, inclusion of CRP may significantly improve discrimination for subjects with the RS risk ranging from 0.034 to 0.070 based on AUC; from 0.010 to 0.078 based on $pAUC_{0.1}$; from 0.032 to 0.068 based on IDI. The IDI findings and the $pAUC$ findings agree with each other. These results suggest that CPR might be useful to improve risk prediction among patients regarded as having low to moderate risk according to the FRS.

It is worth to note that the bandwidth selection procedure is not sensitive towards the choice of the number of folds in cross-validation. Using a five-fold cross-validation, the optimal bandwidths (h_0^{opt}, h_1^{opt}) are (0.121, 0.394) for the standard AUC, (0.238, 0.614) for $pAUC_{0.1}$, and (0.016, 0.272) for IDI. The resulting point estimates and CIs are almost the same as the results with the bandwidths selected via a 10-fold cross validation procedure. In addition, for calculating the weights $\hat{\omega}_i$, the survival function $G(\cdot)$ of the censoring time C is estimated by a Kaplan-Meier estimator since in the study C is likely to be independent of both T and X, Z . In section “Risk Modeling with and Without New Markers”, we commented that if this independence assumption does not hold, we could still provide a correct estimate of $G(\cdot)$ via a semi-parametric model, for example a Cox PH model. Here, we also obtained the estimates of $G(t_0)$ via a Cox PH model, i.e., $\exp\{-\hat{\Lambda}_0(t_0) \exp(\hat{\gamma}_c W_c)\}$

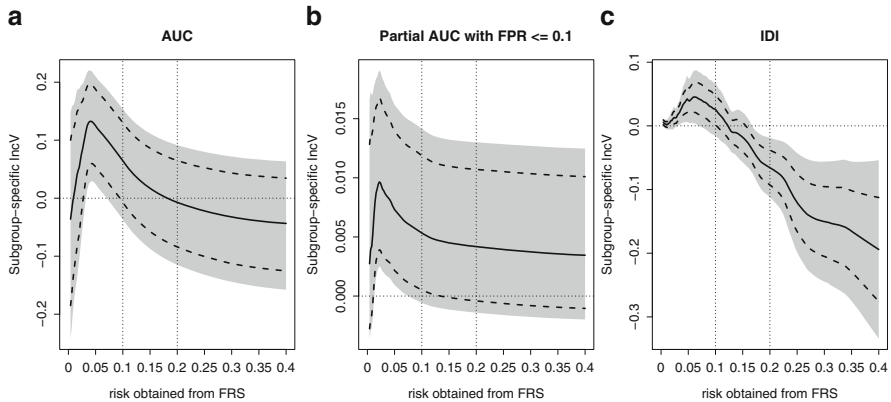


Fig. 4 The point estimates (*solid line*), and its 95% pointwise confidence intervals (*dashed lines*) and the 95% simultaneous confidence bands (*dark shaded region*) for the subgroup-specific IncV with respect to AUC and $pAUC_{0.1}$ as well as IDI. The results are based on the weights $\hat{\omega}_i$ with $G_{X,Z}(t)$ estimated via a Cox PH model. (a) AUC. (b) Partial AUC with $FPR \leq 0.1$. (c) IDI

where W_c consists of the FRS and the CRP. Based on the resulting weights $\hat{\omega}_i$, we obtained the point estimates and CIs for the subgroup-specific IncV w.r.t. AUC, $pAUC_{0.1}$ and IDI, which is presented in Fig. 4. The results are very similar to the results using Kaplan-Meier estimator of $G(\cdot)$, and therefore it implies that the independence assumption about the censoring time C is reasonable.

We are also interested in testing whether the subgroup-specific IncV of the CRP over the FRS is constant over the values $[0, 0.4]$ of the risk estimates obtained from the FRS. The p -values of testing for constant subgroup-specific IncV are 0.028 for AUC, 0.108 for $pAUC_{0.1}$ and 0.002 for IDI. These results agree with Fig. 5, which shows the point estimates and simultaneous 95% CIs for the relative subgroup-specific IncV w.r.t. AUC, $pAUC_{0.1}$ and IDI. It shows that the subgroup-specific IncVs w.r.t. AUC and IDI are not constant over the interval $[0, 0.4]$; on the other hand, the subgroup-specific IncV w.r.t. $pAUC_{0.1}$ is constant over this interval. It is worth to note that the asymptotic variance of $\hat{D}_{A[s_l, s_u]}(s)$ is larger than that of $\hat{A}(s)$, and therefore, the power of testing whether the subgroup-specific IncV is constant over a certain interval is not as strong as the power of testing whether the subgroup-specific IncV is above zero over the interval.

Concluding Remarks

In this paper, we propose a nonparametric procedure to estimate the incremental values of new markers in prediction accuracy across different subgroups defined by the conventional scoring system. We also provide the pointwise and simultaneous interval estimates via perturbation resampling. In addition, with proper adjustment

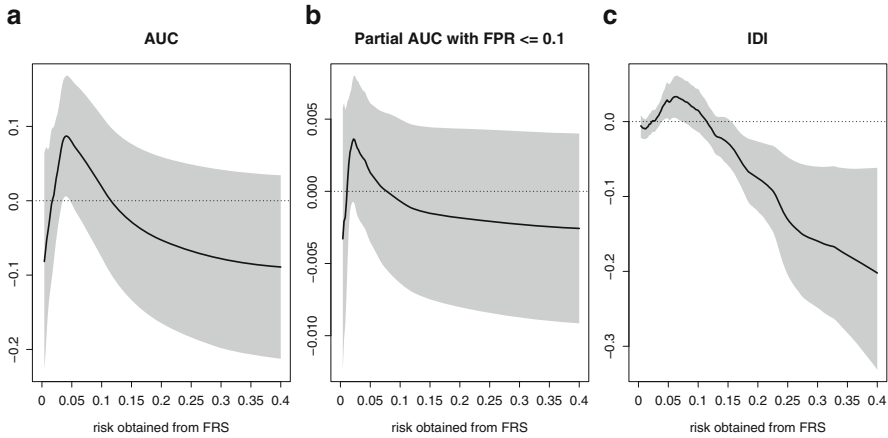


Fig. 5 The point estimates (*solid line*), and its 95% simultaneous confidence bands (*dark shaded region*) for the relative subgroup-specific IncVs, which are used to test for heterogeneous IncVs, with respect to AUC and $pAUC_{0.1}$ as well as IDI over the interval $[0, 0.4]$ of the risk estimates obtained from the FRS. (a) AUC. (b) Partial AUC with $FPR \leq 0.1$. (c) IDI

for multiple subgroups comparison, our approach is able to systematically identify the subgroups which would benefit from adding new markers. Unlike global measures which do not provide information on how the IncV may vary across subgroups, our methods enables the identification of subgroups for which the new markers may or may not be useful. Existing procedures often assess subgroup-specific IncVs empirically. We provide more rigorous and systematic analytical tools to ensure the validity of such claims and more precisely pinpoint such specific subgroups.

Appropriate choice of prediction accuracy summaries is of great importance to capture the usefulness of new markers. It is also motivated by primary research interests. Discrimination is one of the major components in assessing the accuracy of prediction models. The AUC is the most popular summary index which depicts inherent discrimination capacity. However, it is unable to capture how well the predicted risks agree with the actual observed risks [20]. In some cases, alternative summary measures should be also considered, for example, NRI, PCF and PNF. Our approach can be naturally extended to other metrics that maybe more appropriate for particular clinical applications.

The subgroup-specific TPR $\mathcal{S}_1(c; s, t) = pr \{ \bar{p}_2(X, Z) \geq c | T^\dagger \leq t, \bar{p}_1(X) = s \}$ and the subgroup-specific FPR $\mathcal{S}_0(c; s, t) = pr \{ \bar{p}_2(X, Z) \geq c | T^\dagger > t, \bar{p}_1(X) = s \}$ both depend on the time point t , which is usually pre-determined. In some applications, new biomarkers might produce relatively better long-term performance in prediction accuracy than short-term. It is straightforward to extend our procedure to different time points over an arbitrary time interval since the nonparametric estimates of the TPR and FPR, $\mathcal{S}_y(c; s, t)$, converge to a Gaussian process in time t . We could estimate the overall improvement of new markers over a certain

time interval by integrating the subgroup-specific partial AUC and the subgroup-specific IDI index w.r.t. time t . Furthermore, with properly adjusting for multiple comparison, it is possible to identify the time interval where new markers have the most incremental values for different subgroups.

Instead of focusing on the prediction of t -year survival for a fixed time point, we might be also interested in a global assessment of a fitted prediction model for the continuous event time. One example of such global measure is the C-statistic of the prediction score $\mathcal{P}_2(X, Z)$, $pr\{\bar{p}_2(X_i, Z_i) > \bar{p}_2(X_{i'}, Z_{i'}) \mid T_{i'}^\dagger > T_i^\dagger\}$ [22, 26, 32]. When the event time T^\dagger is subject to right censoring which may have finite support $[0, \tau]$, one may consider a truncated C-statistic,

$$C_\tau = pr\left\{\bar{p}_2(X_i, Z_i) > \bar{p}_2(X_{i'}, Z_{i'}) \mid T_{i'}^\dagger > T_i^\dagger, T_i^\dagger < \tau\right\},$$

as considered in Heagerty and Zheng [23] and Uno et al. [45]. It is straightforward to extend C_τ to our subgroup-specific C-statistic

$$C_\tau(s) = pr\left\{\bar{p}_2(X_i, Z_i) > \bar{p}_2(X_{i'}, Z_{i'}) \mid T_{i'}^\dagger > T_i^\dagger, T_i^\dagger < \tau, \bar{p}_1(X_i) = \bar{p}_1(X_{i'}) = s\right\}$$

and construct an IPW kernel estimator for $C_\tau(s)$ as for other accuracy measures.

Acknowledgements The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. The work is supported by grants U01-CA86368, P01- CA053996, R01-GM085047, R01-GM079330, R01-AI052817 and U54-LM008748 awarded by the National Institutes of Health.

Appendix 1

Let \mathbb{P}_n and \mathbb{P} denote expectation with respect to (w.r.t.) the empirical probability measure of $\{(T_i, \Delta_i, X_i, Z_i), i = 1, \dots, n\}$ and the probability measure of (T, Δ, X, Z) , respectively, and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$. We use $\mathcal{F}'(x)$ to denote $d\mathcal{F}(x)/dx$ for any function \mathcal{F} , \simeq to denote equivalence up to $o_p(1)$, and \lesssim to denote being bounded above up to a universal constant. Let β_0 and γ_0 denote the solution to

$$E\left[V_i\left\{Y_i^\dagger - g_1(\beta_0'V_i)\right\}\right] = 0$$

and $E\left[W_i\left\{Y_i^\dagger - g_2(\gamma_0'W_i)\right\}\right] = 0$, respectively. Let $\bar{p}_{1i} = g_1(\beta_0'V_i)$, and $\bar{p}_{2i} = g_2(\gamma_0'W_i)$. Let $\omega = \Delta I(T \leq t_0)/G_{X,Z}(T) + I(T > t_0)/G_{X,Z}(t_0)$, $\bar{M}_i(c) = I(\hat{p}_{2i} \geq c)$ and $\bar{M}_i(c) = I(\bar{p}_{2i} \geq c)$. For $y = 0, 1$, let $f_y(c; s)$ denote the conditional density of \bar{p}_{2i}

given $Y_i^\dagger = y$ and $\bar{p}_{1i} = s$ and we assumed that $f_y(c; s)$ is continuous and bounded away from zero uniformly in c and s . This assumption implies that $\text{ROC}(u; s)$ has continuous and bounded derivative $\text{ROC}'(u; s) = \partial \text{ROC}(u; s) / \partial u$. We assume that V and W are bounded, and $\tau(y; s) = \partial \text{pr}[\phi\{\bar{p}_1(X)\} \leq s, Y^\dagger = y] / \partial s$, is continuously differentiable with bounded derivatives and bounded away from zero. Throughout, the bandwidths are assumed to be of order $n^{-\nu}$ with $\nu \in (1/5, 1/2)$. For ease of presentation and without loss of generality, we assume that $h_1 = h_0$, denoted by h , and suppress h from the notations. Without loss of generality, we assume that $\sup_{t,x,z} |n^{1/2} \{\hat{G}_{X,Z}(t) - G_{X,Z}(t)\}| = O_p(1)$. When C is assumed to be independent of both T and (X, Z) , the simple Kaplan-Meier estimator satisfies this condition. When C depends on (X, Z) , $\hat{G}_{X,Z}$ obtained under the Cox model also satisfies this condition provided that W_c is bounded. The kernel function K is assumed to be symmetric, smooth with a bounded support on $[-1, 1]$ and we let $m_2 = \int K(x)^2 dx$.

Asymptotic Expansions for $\hat{\mathcal{S}}_y(c; s)$

Uniform Convergence Rate for $\hat{\mathcal{S}}_y(c; s)$ We first establish the following uniform convergence rate of $\hat{\mathcal{S}}_y(c; s) = g\{\hat{a}_y(c; s)\}$:

$$\sup_{s \in \mathcal{I}_{h,c}} |\hat{\mathcal{S}}_y(c; s) - \mathcal{S}_y(c; s)| = O_p\{(nh)^{-\frac{1}{2}} \log(n)\} = o_p(1). \tag{6}$$

To this end, we note that for any given c and s ,

$$\hat{\zeta}_y(c; s) = \begin{bmatrix} \hat{\zeta}_{a_y}(c; s) \\ \hat{\zeta}_{b_y}(c; s) \end{bmatrix} = \begin{bmatrix} \hat{a}_y(c; s) - a_y(c; s) \\ \hat{b}_y(c; s) - b_y(c; s) \end{bmatrix}$$

is the solution to the estimating equation $\hat{\Psi}_y(\zeta_y, c, s) = 0$, where $\zeta_y = (\zeta_{a_y}, \zeta_{b_y})'$ and

$$\begin{aligned} \hat{\Psi}_y(\zeta_y; c, s) &= \begin{bmatrix} \hat{\Psi}_{y1}(\zeta_y, c, s) \\ \hat{\Psi}_{y2}(\zeta_y, c, s) \end{bmatrix} \\ &= n^{-1} \sum_{i: Y_i=y} \hat{w}_i \left[\frac{1}{h^{-1} \hat{\mathcal{G}}_{i1}(s)} \right] K_h \left\{ \hat{\mathcal{E}}_{i1}(s) \right\} \left[\hat{M}_i(c) - \mathcal{G}\{\zeta_y, c, s; \phi(\hat{p}_{1i}), h\} \right], \end{aligned}$$

$a_y(c; s) = g^{-1}\{\mathcal{S}_y(c; s)\}$, $b_y(c; s) = \partial g^{-1}\{\mathcal{S}_y(c; s)\} / \partial s$ and

$$\mathcal{G}(\zeta_y, c, s; e, h) = g[a_y(c; s) + b_y(c; s)\{e - \phi(s)\} + \zeta_{a_y} + \zeta_{b_y} h^{-1}\{e - \phi(s)\}].$$

We next establish the convergence rate for $\sup_{\zeta_y, c, s} |\hat{\Psi}_y(\zeta_y; c, s) - \Psi_y(\zeta_y; c, s)|$, where

$$\Psi_y(\zeta_y; c, s) = \begin{bmatrix} \Psi_{y1}(\zeta_y, c, s) \\ \Psi_{y2}(\zeta_y, c, s) \end{bmatrix} = \tau(y; s) \begin{bmatrix} \mathcal{L}_y(c; s) - \int K(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}t\}dt \\ - \int tK(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}t\}dt \end{bmatrix}.$$

We first show that

$$\sup_{s \in \mathcal{S}_{h,c}} \left| n^{-1} \sum_{i:Y_i=y} \hat{\omega}_i K_h\{\hat{\mathcal{E}}_{i1}(s)\} \hat{M}_i(c) - \tau(y; s) \mathcal{L}_y(c; s) \right|$$

and

$$\sup_{\zeta_y, s \in \mathcal{S}_{h,c}} \left| n^{-1} \sum_{i:Y_i=y} \hat{\omega}_i K_h\{\hat{\mathcal{E}}_{i1}(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} - \tau(y; s) \int K(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}t\}dt \right|$$

are both $O_p\{(nh)^{-\frac{1}{2}} \log(n)\}$ where $\mathcal{S}_h = [\phi^{-1}(\rho_l + h), \phi^{-1}(\rho_u - h)]$ and $[\rho_l, \rho_u]$ is a subset of the support of $\phi\{g_1(\beta_0^T V)\}$. To this end, we note that since $\sup_u |\hat{G}_{X,Z}(u) - G_{X,Z}(u)| = O_p(n^{-\frac{1}{2}})$ and $|\hat{\beta} - \beta_0| = O_p(n^{-\frac{1}{2}})$,

$$\begin{aligned} & \left| n^{-1} \sum_{i:Y_i=y} (\hat{\omega}_i - \omega_i) K_h\{\hat{\mathcal{E}}_{i1}(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} \right| \\ & \leq n^{-1} \sum_{i:Y_i=y} |\hat{\omega}_i - \omega_i| K_h\{\hat{\mathcal{E}}_{i1}(s)\} = O_p(n^{-\frac{1}{2}}). \end{aligned}$$

This implies that

$$\begin{aligned} & \left| n^{-1} \sum_{i:Y_i=y} \hat{\omega}_i K_h\{\hat{\mathcal{E}}_{i1}(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} - \tau(y; s) \int K(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}t\}dt \right| \\ & \leq \left| n^{-\frac{1}{2}} \int K_h\{e - \phi(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} d\mathbb{G}_n[\omega I\{\phi(\hat{\rho}_{1i}) \leq e\}] - \omega I\{\phi(\bar{\rho}_{1i}) \leq e\} \right| \\ & + \left| \int K_h\{e - \phi(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} d\mathbb{P}[\omega I\{\phi(\bar{\rho}_{1i}) \leq e\}] - \tau(y; s) \int K(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}t\}dt \right| \\ & + \left| n^{-\frac{1}{2}} \int K_h\{e - \phi(s)\} d\mathbb{P}[\omega \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} I\{\phi(\bar{\rho}_{1i}) \leq e\}] \right| + O_p(n^{-\frac{1}{2}}) \\ & \lesssim n^{-\frac{1}{2}} h^{-1} \|\mathbb{G}_n\|_{\mathcal{X}_{\mathcal{E}_0}} + \left| n^{-\frac{1}{2}} \int K_h\{e - \phi(s)\} d\mathbb{P}[\omega \mathcal{G}\{\zeta_y, c, s; \phi(\hat{\rho}_{1i}), h\} I\{\phi(\bar{\rho}_{1i}) \leq e\}] \right| \\ & \quad + O_p(n^{-\frac{1}{2}} + h^2), \end{aligned}$$

where $\mathcal{H}_\delta = \{\omega I[\phi\{g_1(\beta'v)\} \leq e] - \omega I[\phi\{g_1(\beta'_0v)\} \leq e] : |\beta - \beta_0| \leq \delta, e\}$ is a class of functions indexed by β and e . By the maximum inequality of Van der vaart and Wellner [47], we have

$$E\|\mathbb{G}_n\|_{\mathcal{H}_\delta} \lesssim \delta^{\frac{1}{2}} \{|\log(\delta)| + |\log(h)|\} \left[1 + \frac{\delta^{\frac{1}{2}} \{|\log(\delta)| + |\log(h)|\}}{\delta n^{\frac{1}{2}}} \right]$$

Together with the fact that $|\hat{\beta} - \beta_0| = O_p(n^{-\frac{1}{2}})$ from Uno et al. [44], it implies that $n^{-\frac{1}{2}}h^{-1}\|\mathbb{G}_n\|_{\mathcal{H}_\delta} = O_p\{(nh)^{-\frac{1}{2}}(nh^2)^{-\frac{1}{4}}\log(n)\}$. In addition, with the standard arguments used in Bickel and Rosenblatt [2], it can be shown that

$$\begin{aligned} & \left| n^{-\frac{1}{2}} \int K_h\{e - \phi(s)\}d\mathbb{P}[\omega\mathcal{G}\{\zeta_y, c, s; \phi(\hat{p}_{1i}), h\}I\{\phi(\bar{p}_{1i}) \leq e\}] \right| \\ &= O_p\{(nh)^{-\frac{1}{2}}\log(n)\}. \end{aligned}$$

Therefore, for $h = n^{-\nu}$, $1/5 < \nu < 1/2$,

$$\begin{aligned} \sup_{\zeta_y, s \in \mathcal{S}_{h,c}} \left| n^{-1} \sum_{i:Y_i=y} \hat{\omega}_i K_h\{\hat{\mathcal{E}}_{i1}(s)\} \mathcal{G}\{\zeta_y, c, s; \phi(\hat{p}_{1i}), h\} \right. \\ \left. - \tau(y; s) \int K(t)g\{a_y(c; s) + \zeta_{a_y} + \zeta_{b_y}, t\}dt \right| \end{aligned}$$

is $O_p\{(nh)^{-\frac{1}{2}}\log(n)\}$. Following with similar arguments as given above, coupled with the fact that $|\hat{\gamma} - \gamma_0| = O_p(n^{-\frac{1}{2}})$, we have

$$\sup_{s \in \mathcal{S}_{h,c} \in [0,1]} \left| n^{-1} \sum_{i:Y_i=y} \hat{\omega}_i K_h\{\hat{\mathcal{E}}_{i1}(s)\} \hat{M}_i(c) - \tau(y; s) \mathcal{S}_y(c; s) \right| = O_p\{(nh)^{-\frac{1}{2}}\log(n)\}.$$

Thus, $\sup_{\zeta_y, c, s} |\hat{\Psi}_{y1}(\zeta_y; c, s) - \Psi_{y1}(\zeta_y; c, s)| = O_p\{(nh)^{-\frac{1}{2}}\log(n)\} = o_p(1)$. It follows from the same arguments as given above that

$$\sup_{\zeta_y, c, s} |\hat{\Psi}_{y2}(\zeta_y; c, s) - \Psi_{y2}(\zeta_y; c, s)| = O_p\{(nh)^{-\frac{1}{2}}\log(n) + h\} = o_p(1).$$

Therefore, $\sup_{\zeta_y, c, s} |\hat{\Psi}_y(\zeta_y; c, s) - \Psi_y(\zeta_y; c, s)| = o_p(1)$. In addition, we note that $\mathbf{0}$ is the unique solution to the equation $\Psi_y(\zeta_y; c, s) = 0$ w.r.t. ζ_y . It suggests that $\sup_{s,c} |\hat{\zeta}_{a_y}(c; s)| = O_p\{(nh)^{-\frac{1}{2}}\log(n)\} = o_p(1)$, which implies the consistency of $\mathcal{S}_y(c; s)$,

$$\sup_{s \in \mathcal{S}_{h,c} \in [0,1]} |\hat{\mathcal{S}}_y(c; s) - \mathcal{S}_y(c; s)| = O_p\{(nh)^{-\frac{1}{2}}\log(n)\} = o_p(1).$$

Asymptotic Expansion for $\hat{\mathcal{S}}_y(c; s)$ Let $\hat{d}_y(c; s) = \sqrt{nh}\{\hat{a}_y(c; s) - a_y(c; s)\}$. It follows from a Taylor series expansion and the convergence rate of $\zeta_y(c; s)$ that

$$\hat{d}_y(c; s) = \frac{\sqrt{nh}\mathbb{P}_n\left(\hat{\omega}I(Y=y)K_h\{\hat{\mathcal{E}}_1(s)\}[\hat{M}(c) - \mathcal{G}_y^0\{c, s; \phi(\hat{p}_1)\}]\right)}{\tau\{y; \phi(s)\}\dot{g}\{a_y(c; s)\}} + o_p(1), \tag{7}$$

where $\mathcal{G}_y^0(c, s; e) = g[a_y(c; s) + b_y(c; s)\{e - \phi(s)\}]$. Furthermore,

$$\hat{d}_y(c; s) = \frac{\sqrt{nh}\mathbb{P}_n\left(\omega I(Y=y)K_h\{\hat{\mathcal{E}}_1(s)\}[\hat{M}(c) - \mathcal{G}_y^0\{c, s; \phi(\hat{p}_1)\}]\right)}{\tau\{y; \phi(s)\}\dot{g}\{a_y(c; s)\}} + o_p(1),$$

since $\sup_{t \leq t_0} |\hat{G}_{X,Z}(t) - G_{X,Z}(t)| = O_p(n^{-1/2})$. We next show that $\hat{d}_y(c; s)$ is asymptotically equivalent to

$$\tilde{d}_y(c; s) = \frac{\sqrt{nh}\mathbb{P}_n\left(\omega I(Y=y)K_h\{\bar{\mathcal{E}}_1(s)\}[\bar{M}(c) - \mathcal{G}_y^0\{c, s; \phi(\bar{p}_1)\}]\right)}{\tau\{y; \phi(s)\}\dot{g}\{a_y(c; s)\}}, \tag{8}$$

where $\bar{\mathcal{E}}_1(s) = \phi(\bar{p}_1) - \phi(s)$. From (7), (8), and the fact that $\tau\{y; \phi(s)\}$ is bounded away from 0 uniformly in s , we have

$$\begin{aligned} & |\hat{d}_y(s) - \tilde{d}_y(s)| \\ & \leq h^{\frac{1}{2}} \left| \int K_h\{e - \phi(s)\} d\mathbb{G}_n(I(Y=y)\omega[\hat{M}(c)I\{\phi(\hat{p}_1) \leq e\} \right. \\ & \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \left. - \bar{M}(c)I\{\phi(\bar{p}_1) \leq e\}]\right| \\ & + h^{\frac{1}{2}} \left| \int K_h\{e - \phi(s)\} \mathcal{G}_y(c, s; e) d\mathbb{G}_n(I(Y=y)[\omega I\{\phi(\hat{p}_1) \leq e\} \right. \\ & \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \left. - \omega I\{\phi(\bar{p}_1) \leq e\}]\right| \\ & + \left| \sqrt{nh} \int K_h\{e - \phi(s)\} d\mathbb{P}(I(Y=y)[\omega \hat{M}(c)I\{\phi(\hat{p}_1) \leq e\} \right. \\ & \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \left. - \omega \bar{M}(c)I\{\phi(\bar{p}_1) \leq e\}]\right| \\ & + \left| \sqrt{nh} \int K_h\{e - \phi(s)\} d\mathbb{P}(I(Y=y)[\omega \mathcal{G}_y\{c, s; \phi(\hat{p}_1)\}I\{\phi(\hat{p}_1) \leq e\} \right. \\ & \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \left. - \omega \mathcal{G}_y\{c, s; \phi(\bar{p}_1)\}I\{\phi(\bar{p}_1) \leq e\}]\right| \\ & \leq h^{\frac{1}{2}} \|\mathbb{G}_n\|_{\mathcal{F}_\delta} + h^{\frac{1}{2}} \|\mathbb{G}_n\|_{\mathcal{H}_\delta} + O_p\{(nh)^{1/2}|\hat{\beta} - \beta_0| + |\hat{\gamma} - \gamma_0| + h^2\}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{F}_\delta = \{ & \omega I\{g_2(\gamma'w) \geq c\}I[\phi\{g_1(\beta'v)\} \leq e] \\ & - \omega I\{g_2(\gamma'_0w) \geq c\}I[\phi\{g_1(\beta'_0v)\} \leq e] : |\gamma - \gamma_0| + |\beta - \beta_0| \leq \delta, e \} \end{aligned}$$

is the class of functions indexed by γ , β and e . By the maximum inequality of Van der vaart and Wellner [47] and the fact that $|\hat{\beta} - \beta_0| + |\hat{\gamma} - \gamma_0| = O_p(n^{-\frac{1}{2}})$ from Uno et al. [44], we have $h^{\frac{1}{2}} \|\mathbb{G}_n\|_{\mathcal{F}_\delta} = O_p\{h^{-\frac{1}{2}}n^{-\frac{1}{4}}\log(n)\}$ and $h^{\frac{1}{2}} \|\mathbb{G}_n\|_{\mathcal{H}_\delta} = O_p\{h^{-\frac{1}{2}}n^{-\frac{1}{4}}\log(n)\}$. It follows that $\sup_s |\hat{d}_y(s) - \tilde{d}_y(s)| = o_p(1)$. Then, by a delta method,

$$\hat{\mathcal{W}}_{\mathcal{S}_y}(c; s) = \sqrt{nh}\{\hat{\mathcal{S}}_y(c; s) - \mathcal{S}_y(c; s)\} \simeq \sqrt{nh} \mathbb{P}_n [K_h\{\hat{\mathcal{E}}_1(s)\} \mathcal{D}_{\mathcal{S}_y}(c; s)] \tag{9}$$

where $\mathcal{D}_{\mathcal{S}_y}(c; s) = \tau\{y; \phi(s)\}^{-1} \omega I(Y = y) \{\bar{M}(c) - \mathcal{S}_y(c; s)\}$ (10)

Using the same arguments as for establishing the uniform convergence rate of conditional Kaplan-Meier estimators [12, 16], we obtain (6). Furthermore, following similar arguments as given in Dabrowska [11, 13], we have $\hat{\mathcal{W}}_{\mathcal{S}_y}(c; s)$ converges weakly to a Gaussian process in c for all s . Note that as for all kernel estimators, $\hat{\mathcal{W}}_{\mathcal{S}_y}(c; s)$ does not converge as a process in s .

Uniform Consistency of $\widehat{\text{pAUC}}_f(s)$

Next we establish the uniform convergence rate for $\widehat{\text{ROC}}(u; s)$. To this end, we write

$$\widehat{\text{ROC}}(u; s) - \text{ROC}(u; s) = \hat{\epsilon}_1(u; s) + \hat{\epsilon}_0(u; s),$$

where $\hat{\epsilon}_1(u; s) = \hat{\mathcal{S}}_1\{\hat{\mathcal{S}}_0^{-1}(u; s); s\} - \mathcal{S}_1\{\hat{\mathcal{S}}_0^{-1}(u; s); s\}$ and $\hat{\epsilon}_0(u; s) = \mathcal{S}_1\{\hat{\mathcal{S}}_0^{-1}(u; s); s\} - \mathcal{S}_1\{\mathcal{S}_0^{-1}(u; s); s\}$. It follows from (6) that $\sup_{u,s} |\hat{\epsilon}_1(u; s)| \leq \sup_{c,s} |\hat{\mathcal{S}}_1(c; s) - \mathcal{S}_1(c; s)|$. Let $\hat{\mathcal{S}}(u; s) = \mathcal{S}_0\{\hat{\mathcal{S}}_0^{-1}(u; s); s\}$. Then $\hat{\epsilon}_0(u; s) = \text{ROC}\{\hat{\mathcal{S}}(u; s); s\} - \text{ROC}(u; s)$. Noting that $\sup_u |\hat{\mathcal{S}}(u; s) - u| = \sup_u |\hat{\mathcal{S}}(u; s) - \mathcal{S}_0\{\hat{\mathcal{S}}_0^{-1}(u; s); s\}| + n^{-1} \leq \sup_c |\mathcal{S}_0(c; s) - \hat{\mathcal{S}}_0(c; s)| + n^{-1} = O_p\{(nh)^{-1/2} \log n\}$, we have $\hat{\epsilon}_0(u; s) = O_p\{(nh)^{-1/2} \log n\}$ by the continuity and boundedness of $\widehat{\text{ROC}}(u; s)$. Therefore,

$$\sup_{u,s} |\widehat{\text{ROC}}(u; s) - \text{ROC}(u; s)| = O_p\{(nh)^{-1/2} \log n\}$$

which implies

$$\begin{aligned} & \sup_{s \in \mathcal{I}_h} \left| \widehat{\text{pAUC}}_f(s) - \text{pAUC}_f(s) \right| \\ & \lesssim \sup_{s \in \mathcal{I}_h} \int_0^f \left| \widehat{\text{ROC}}(u; s) - \text{ROC}(u; s) \right| du = O_p\{(nh)^{-\frac{1}{2}} \log n\}. \end{aligned}$$

and hence the uniform consistency of $\widehat{\text{pAUC}}_f(s)$.

Asymptotic Distribution of $\hat{\mathcal{W}}_{\text{pAUC}_f}(s)$

To derive the asymptotic distribution for $\hat{\mathcal{W}}_{\text{pAUC}_f}(s)$, we first derive asymptotic expansions for $\hat{\mathcal{W}}_{\text{ROC}}(u; s) = \sqrt{nh}\{\widehat{\text{ROC}}(u; s) - \text{ROC}(u; s)\} = \sqrt{nh} \hat{\epsilon}_1(u; s) + \sqrt{nh} \hat{\epsilon}_0(u; s)$. From the weak convergence of $\hat{\mathcal{W}}_{S_y}(c; s)$ in c , the approximation in (9), and the consistency of $\hat{\mathcal{S}}_0^{-1}(c; s)$ given in the section “Uniform Consistency of $\widehat{\text{pAUC}}_f(s)$ ” in Appendix 1, we have

$$\begin{aligned} \sqrt{nh} \hat{\epsilon}_1(u; s) &\simeq \sqrt{nh} \left[\hat{\mathcal{S}}_1 \{ \mathcal{S}_0^{-1}(u; s); s \} - \text{ROC}(u; s) \right] \\ &\simeq \sqrt{nh} \mathbb{P}_n \left[K_h \{ \bar{\mathcal{E}}_1(s) \} \mathcal{D}_{\mathcal{S}_1} \{ \mathcal{S}_0^{-1}(u; s); s \} \right] \end{aligned}$$

On the other hand, from the uniform convergence of $\hat{\mathcal{S}}_0(u; s) \rightarrow u$ and the weak convergence of $\hat{\mathcal{D}}_0(c; s)$ in c , we have

$$\begin{aligned} \sqrt{nh} \{ u - \hat{\mathcal{S}}(u; s) \} &\simeq \sqrt{nh} \left[\hat{\mathcal{S}}^{-1} \{ \hat{\mathcal{S}}(u; s); s \} - \hat{\mathcal{S}}(u; s) \right] \simeq \sqrt{nh} \{ \hat{\mathcal{S}}^{-1}(u; s) - u \} \\ &\simeq \sqrt{nh} \left[\hat{\mathcal{S}}_0 \{ \mathcal{S}_0^{-1}(u; s); s \} - u \right] \end{aligned}$$

This, together with a Taylor series expansion and the expansion given (9), implies that

$$\sqrt{nh} \hat{\epsilon}_0(u; s) \simeq -\text{ROC}(u; s) \mathbb{P}_n \left[K_h \{ \bar{\mathcal{E}}_1(s) \} \mathcal{D}_{\mathcal{S}_0} \{ \mathcal{S}_0^{-1}(u; s); s \} \right]$$

It follows that

$$\hat{\mathcal{W}}_{\text{pAUC}_f}(s) \simeq \sqrt{nh} \mathbb{P}_n \left[K_h \{ \bar{\mathcal{E}}_1(s) \} \mathcal{D}_{\text{pAUC}_f}(s) \right] \tag{11}$$

where $\mathcal{D}_{\text{pAUC}_f}(s) = \int_0^f \left[\mathcal{D}_{\mathcal{S}_1} \{ \mathcal{S}_0^{-1}(u; s); s \} - \text{ROC}(u; s) \mathcal{D}_{\mathcal{S}_0} \{ \mathcal{S}_0^{-1}(u; s); s \} \right] du$. (12)

It then follows from a central limit theorem that for any fixed s , $\hat{\mathcal{W}}_{\text{pAUC}_f}(s)$ converges to a normal with mean 0 and variance

$$\sigma_{\text{pAUC}_f}^2(s) = m_2 \left[\tau \{ 1; \phi(s) \} \dot{F}_{\phi(\bar{p}_1)}(s) \right]^{-1} \sigma_1^2(s) + m_2 \left[\tau \{ 0; \phi(s) \} \dot{F}_{\phi(\bar{p}_1)}(s) \right]^{-1} \sigma_0^2(s),$$

where $\dot{F}_{\phi(\bar{p}_1)}(s)$ is the density function of $\phi(\bar{p}_1)$,

$$\sigma_1^2(s) = \text{E} \left(G(T^\dagger)^{-1} \left[\int_0^f \bar{M} \{ \mathcal{S}_0^{-1}(u; s) \} du - \text{pAUC}_f(s) \right]^2 \Big| \bar{p}_1 = s, Y^\dagger = 1 \right), \quad \text{and}$$

$$\sigma_0^2(s) = E \left(G(t_0)^{-1} \left[\int_0^f \bar{M}\{\mathcal{S}_0^{-1}(u;s)\} d\text{ROC}(u;s) - \int_0^f u d\text{ROC}(u;s) \right]^2 \mid \bar{p}_1 = s, Y^\dagger = 0 \right).$$

Justification for the Resampling Methods

To justify the resampling method, we first note that

$$|\beta^* - \hat{\beta}| + |\gamma^* - \hat{\gamma}| + \sup_{t \leq t_0} |G_{X,Z}^*(t) - \hat{G}_{X,Z}(t)| = O_p(n^{-\frac{1}{2}}).$$

It follows from similar arguments given in the Appendix 1 and Appendix 1 of [7] that $\mathcal{W}_{S_y}^*(c;s) = \sqrt{nh}\{\mathcal{S}_y^*(c;s) - \hat{\mathcal{S}}_y(c;s)\} \simeq n^{\frac{1}{2}}h^{-1/2} \sum_{i=1}^n \hat{\mathcal{D}}_{\mathcal{S}_y,i}(c;s)\xi_i$, where $\hat{\mathcal{D}}_{\mathcal{S}_y,i}(c;s)$ is obtained by replacing all theoretical quantities in $\mathcal{D}_{\mathcal{S}_y}(c;s)$ given in (10) with the estimated counterparts for the i th subject. This, together with similar arguments as given above for the expansion of $\mathcal{W}_{\text{ROC}}(u;s)$, implies that

$$\begin{aligned} \mathcal{W}_{\text{pAUC}_f}^*(s) &= \int_0^f \sqrt{nh}\{\text{ROC}^*(u;s) - \widehat{\text{ROC}}(u;s)\} du \\ &\simeq n^{-\frac{1}{2}}h^{-1/2} \sum_{i=1}^n K_h\{\hat{\mathcal{E}}_1(s)\} \hat{\mathcal{D}}_{\text{pAUC}_f}(s)\xi_i, \end{aligned}$$

where $\hat{\mathcal{D}}_{\text{pAUC}_f}(s) = \int_0^f [\hat{\mathcal{D}}_{\mathcal{S}_1,i}\{\hat{\mathcal{S}}_0^{-1}(u;s);s\} - \text{ROC}(u;s)\hat{\mathcal{D}}_{\mathcal{S}_0,i}\{\hat{\mathcal{S}}_0^{-1}(u;s);s\}] du$. Conditional on the data, $\mathcal{W}_{\text{pAUC}_f}^*(s)$ is approximately normally distributed with mean 0 and variance

$$\hat{\sigma}_{\text{pAUC}_f}^2(s) = h^{-1} \sum_{i=1}^n K_h\{\hat{\mathcal{E}}_1(s)\}^2 \hat{\mathcal{D}}_{\text{pAUC}_f}(s)^2.$$

Using the consistency of the proposed estimators along with similar arguments as given above, it is not difficult to show that the above variance converges to $\sigma_{\text{pAUC}_f}^2(s)$ as $n \rightarrow \infty$. Therefore, the empirical distribution obtained from the perturbed sample can be used to approximate the distribution of $\mathcal{W}_{\text{pAUC}_f}(s)$.

We now show that after proper standardization, the supremum type statistics Γ converges weakly. To this end, we first note that, similar arguments as given in the Appendix 1 can be used to show that $\sup_{s \in \mathcal{S}_h} |\hat{\sigma}_{\text{pAUC}_f}^2(s) - \sigma_{\text{pAUC}_f}^2(s)| = o_p(n^{-\delta})$ and

$$\Gamma = \sup_{s \in \mathcal{S}_h} \left| \frac{\sqrt{nh} \mathbb{P}_n \left[K_h\{\bar{\mathcal{E}}_1(s)\} \mathcal{D}_{\text{pAUC}_f}(s) \right]}{\sigma_{\text{pAUC}_f}(s)} \right| + o_p(n^{-\delta}),$$

for some small positive constant δ . Using similar arguments in Bickel and Rosenblatt [2], we have

$$pr\{a_n(\Gamma - d_n) < x\} \rightarrow e^{-2e^{-x}},$$

where $a_n = [2 \log \{(\rho_u - \rho_l)/h\}]^{\frac{1}{2}}$ and $d_n = a_n + a_n^{-1} \log \{ \int \dot{K}(t)^2 dt / (4m_2\pi) \}$. Now to justify the resampling procedure for constructing the confidence interval, we note that

$$\mathcal{W}_{pAUC_f}^*(s) = n^{-\frac{1}{2}} h^{\frac{1}{2}} \sum_{i=1}^n K_h \{ \hat{\mathcal{E}}_{1i}(s) \} \hat{\mathcal{D}}_{pAUC_f}(s) (\xi_i - 1) + \mathcal{E}^*(s).$$

where $pr\{\sup_{s \in \Omega(h)} |n^\delta \mathcal{E}^*(s)| \geq e \mid \text{data}\} \rightarrow 0$ in probability. Therefore,

$$\Gamma^* = \sup_{s \in \mathcal{I}_h} \left| \frac{n^{-\frac{1}{2}} h^{\frac{1}{2}} \sum_{i=1}^n K_h \{ \hat{\mathcal{E}}_{1i}(s) \} \hat{\mathcal{D}}_{pAUC_f}(s) (\xi_i - 1)}{\sigma_{pAUC_f}(s)} \right| + |\mathcal{E}_{sup}^*|.$$

where $pr\{|n^\delta \mathcal{E}_{sup}^*| \geq e \mid \text{data}\} \rightarrow 0$. It follows from similar arguments as given in Tian et al. [40] and Zhao et al. [53] that

$$\sup \left| pr\{a_n(\Gamma^* - d_n) < x \mid \text{data}\} - e^{-2e^{-x}} \right| \rightarrow 0,$$

in probability as $n \rightarrow \infty$. Thus, the conditional distribution of $a_n(\Gamma^* - d_n)$ can be used to approximate the unconditional distribution of $a_n(\Gamma - d_n)$. When $h_0 = h_1$, in general, the standardized Γ does not converge to the extreme value distribution. However, when $h_0 = h_1 = k \in (0, \infty)$, the distribution of the suitable standardized version of Γ still can be approximated by that of the corresponding standardized Γ^* conditional on the data [21].

Appendix 2

Bandwidth Selection for $pAUC_f(s)$

The choice of the bandwidths h_0 and h_1 is important for making inference about $\mathcal{S}_y(c; s)$ and consequently $pAUC_f(s)$. Here we propose a two-stage K-fold cross-validation procedure to obtain the optimal bandwidth for $\hat{\mathcal{S}}_{0, h_0}^{-1}(u; s)$ and $\hat{\mathcal{S}}_{1, h_1}(c; s)$ sequentially. Specifically, we randomly split the data into K disjoint subsets of about equal sizes denoted by $\{\mathcal{I}_k, k = 1, \dots, K\}$. The two-stage procedure is described as follows:

- (I) Motivated by the fact that $\mathcal{S}_0^{-1}(u; s)$ is essentially the $(1 - u)$ -th quantile of the conditional distribution of $\bar{p}_2(X, Z)$ given $Y^\dagger = 0$ and $\bar{p}_1(X) = s$, for each k , we use all the observations not in \mathcal{J}_k to estimate $q_{0,1-u}(s) = \mathcal{S}_0^{-1}(u; s)$ by obtaining $\{\hat{\alpha}_0(s; h), \hat{\alpha}_1(s; h)\}$, the minimizer of

$$\sum_{j \in \mathcal{J}_1, j \neq k} I(Y_j = 0) \hat{w}_j K_h\{\hat{\mathcal{E}}_{1j}(s)\} \rho_{1-u} \left[\hat{p}_{2j} - g\{\alpha_0 + \alpha_1 \hat{\mathcal{E}}_{1j}(s)\} \right]$$

w.r.t. (α_0, α_1) , where $\rho_\tau(e)$ is a check function defined as $\rho_\tau(e) = \tau e$, if $e \geq 0$; $= (\tau - 1)e$, otherwise. Let $\hat{q}_{0,1-u}^{(-k)}(s; h) = g\{\hat{\alpha}_0(s; h)\}$ denote the resulting estimator of $q_{0,1-u}(s)$. With observations in \mathcal{J}_k , we obtain

$$Err_k^{(q_0)}(h) = \sum_{i \in \mathcal{J}_k} (1 - Y_i) \hat{w}_i \int_0^f \rho_{1-u} \left[\hat{p}_{2i} - \hat{q}_{0,1-u}^{(-k)}(\hat{p}_{1i}; h) \right] du.$$

Then, we let $h_0^{\text{opt}} = \arg \min_h \sum_{k=1}^K Err_k^{(q_0)}(h)$.

- (II) Next, to find an optimal h_1 for $\hat{\mathcal{S}}_{1,h_1}(\cdot; s)$, we choose an error function that directly relates to $\text{pAUC}_f(s) = - \int_{\mathcal{S}_0^{-1}(f; s)}^{\infty} \mathcal{S}_1(c; s) d\mathcal{S}_0(c; s)$. Specifically, noting the fact that

$$E \left(\int_{\mathcal{S}_0^{-1}(f; s)}^{\infty} [I\{g_2(\gamma W_i) \geq c\} - \mathcal{S}_1(c; s)] d\mathcal{S}_0(c; s) \Big| Y_i^\dagger = 1, g_1(\beta' X_i) = s \right) = 0,$$

we use the corresponding mean integrated squared error for $I\{g_2(\gamma W_i) \geq c\} - \mathcal{S}_1(c; s)$ as the error function. For each k , we use all the observations which are not in \mathcal{J}_k to obtain the estimate of $\mathcal{S}_1(c; s)$, denoted by $\hat{\mathcal{S}}_{1,h}^{(-k)}(c; s)$ via (4). Then, with the observations in \mathcal{J}_k , we calculate the prediction error

$$Err_k^{(\mathcal{S}_1)}(h) = - \sum_{i \in \mathcal{J}_k, Y_i = 1} \hat{w}_i \int_{\mathcal{S}_{0,h_0}^{-1}(f; \hat{p}_{1i})}^{\infty} \left\{ I(\hat{p}_{2i} \geq c) - \hat{\mathcal{S}}_{1,h}^{(-k)}(c; \hat{p}_{1i}) \right\}^2 d\mathcal{S}_{0,h_0}(c; \hat{p}_{1i}).$$

We let $h_1^{\text{opt}} = \arg \min_h \sum_{k=1}^K Err_k^{(\mathcal{S}_1)}(h)$.

Since the order of h_y^{opt} is expected to be $n^{-1/5}$ [19], the bandwidth we use for estimation is $h_y = h_y^{\text{opt}} \times n^{-d_0}$ with $0 < d_0 < 3/10$ such that $h_y = n^{-\nu}$ with $1/5 < \nu < 1/2$. This ensures that the resulting functional estimator $\mathcal{S}_{y,h_y}(c; s)$ with the data-dependent smooth parameter has the above desirable large sample properties.

Bandwidth Selection for IDI(*s*)

Same as bandwidth selection for pAUC, we also propose a K-fold cross validation procedure to choose the optimal bandwidth h_1 for $IS(s) = \int_0^1 \mathcal{S}_1(c; s)dc$ and h_0 for $IP(s) = \int_0^1 \mathcal{S}_0(c; s)dc$ separately. The procedure is described as follows: we randomly split the data into K disjoint subsets of about equal sizes denoted by $\{\mathcal{J}_k, k = 1, \dots, K\}$. Motivated by the fact (3), for each k , we use all the observations not in \mathcal{J}_k to estimate $\int_0^1 \mathcal{S}_y(c, s)dc$ by obtaining $\{\hat{\varphi}_0^{(y)}(s; h), \hat{\varphi}_1^{(y)}(s; h)\}$ for $y = 0, 1$, which is the solution to the estimating equation

$$\sum_{j \in \mathcal{J}_1, l \neq k} I(Y_j = y) \hat{\omega}_j K_h \{\hat{\mathcal{E}}_{1j}(s)\} \left[\hat{p}_{2j} - g\{\varphi_0^{(y)} + \varphi_1^{(y)} \hat{\mathcal{E}}_{1j}(s)\} \right] = 0,$$

w.r.t. $(\varphi_0^{(y)}, \varphi_1^{(y)})$. Let $\widehat{IS}^{(-k)}(s; h) = g\{\hat{\varphi}_0^{(1)}(s; h)\}$ and $\widehat{IP}^{(-k)}(s; h) = g\{\hat{\varphi}_0^{(0)}(s; h)\}$. With observations in \mathcal{J}_k , we obtain

$$Err_k^{(IS)}(h) = \sum_{i \in \mathcal{J}_k} Y_i \hat{\omega}_i \left\{ \hat{p}_{2i} - \widehat{IS}^{(-k)}(\hat{p}_{1i}; h) \right\}^2,$$

or

$$Err_k^{(IP)}(h) = \sum_{i \in \mathcal{J}_k} (1 - Y_i) \hat{\omega}_i \left\{ \hat{p}_{2i} - \widehat{IP}^{(-k)}(\hat{p}_{1i}; h) \right\}^2.$$

Then, we let $h_1^{opt} = \arg \min_h \sum_{k=1}^K Err_k^{(IS)}(h)$ and $h_0^{opt} = \arg \min_h \sum_{k=1}^K Err_k^{(IP)}(h)$.

Appendix 3

R codes for application will be available from the corresponding author upon request.

References

1. Baker, S., Pinsky, P.: A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *J. Am. Stat. Assoc.* **96**, 421–428 (2001)
2. Bickel, P., Rosenblatt, M.: On some global measures of the deviations of density function estimates. *Ann. Stat.* **1**, 1071–1095 (1973)
3. Blumenthal, R., Michos, E., Nasir, K.: Further improvements in CHD risk prediction for women. *J. Am. Med. Assoc.* **297**, 641–643 (2007)

4. Cai, T., Cheng, S.: Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9**, 216–233 (2008)
5. Cai, T., Dodd, L.E.: Regression analysis for the partial area under the ROC curve. *Stat. Sin.* **18**, 817–836 (2008)
6. Cai, T., Tian, L., Wei, L.: Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika* **92**(3), 619–632 (2005)
7. Cai, T., Tian, L., Uno, H., Solomon, S., Wei, L.: Calibrating parametric subject-specific risk estimation. *Biometrika* **97**(2), 389–404 (2010)
8. Cook, N., Ridker, P.: The use and magnitude of reclassification measures for individual predictors of global cardiovascular risk. *Ann. Intern. Med.* **150**(11), 795–802 (2009)
9. Cook, N., Buring, J., Ridker, P.: The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann. Intern. Med.* **145**, 21–29 (2006)
10. Cox, D.: Regression models and life-tables. *J. R. Stat. Soc. B (Stat. Methodol.)* **34**(2), 187–220 (1972)
11. Dabrowska, D.: Non-parametric regression with censored survival time data. *Scand. J. Stat.* **14**(3), 181–197 (1987)
12. Dabrowska, D.: Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann. Stat.* **17**(3), 1157–1167 (1989)
13. Dabrowska, D.: Smoothed Cox regression. *Ann. Stat.* **25**(4), 1510–1540 (1997)
14. D’Agostino, R.: Risk prediction and finding new independent prognostic factors. *J. Hypertens.* **24**(4), 643–645 (2006)
15. Dodd, L., Pepe, M.: Partial AUC estimation and regression. *Biometrics* **59**, 614–623 (2003)
16. Du, Y., Akritas, M.: Iid representations of the conditional Kaplan-Meier process for arbitrary distributions. *Math. Method. Stat.* **11**, 152–182 (2002)
17. Dwyer, A.J.: In pursuit of a piece of the ROC. *Radiology* **201**, 621–625 (1996)
18. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults: Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *J. Am. Med. Assoc.* **285**(19), 2486–2497 (2001)
19. Fan, J., Gijbels, I.: Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation. *J. R. Stat. Soc. B (Stat. Methodol.)* **57**, 371–394 (1995)
20. Gail, M., Pfeiffer, R.: On criteria for evaluating models of absolute risk. *Biostatistics* **6**(2), 227–239 (2005)
21. Gilbert, P., Wei, L., Kosorok, M., Clemens, J.: Simultaneous inferences on the contrast of two hazard functions with censored observations. *Biometrics* **58**(4), 773–780 (2002)
22. Harrell, F. Jr., Lee, K., Mark, D.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**(4), 361–387 (1996)
23. Heagerty, P., Zheng, Y.: Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105 (2005)
24. Jiang, Y., Metz, C., Nishikawa, R.: A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **201**, 745–750 (1996)
25. Jin, Z., Ying, Z., Wei, L.: A simple resampling method by perturbing the minimand. *Biometrika* **88**(2), 381–390 (2001)
26. Korn, E., Simon, R.: Measures of explained variation for survival data. *Stat. Med.* **9**(5), 487–503 (1990)
27. Li, G., Doss, H.: An approach to nonparametric regression for life history data using local linear fitting. *Ann. Stat.* **23**, 787–823 (1995)
28. McIntosh, M., Pepe, M.: Combining several screening tests: optimality of the risk score. *Biometrics* **58**(3), 657–664 (2002)
29. Park, Y., Wei, L.: Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **9**, 717–723 (2003)

30. Park, B., Kim, W., Ruppert, D., Jones, M., Signorini, D., Kohn, R.: Simple transformation techniques for improved non-parametric regression. *Scand. J. Stat.* **24**(2), 145–163 (1997)
31. Paynter, N., Chasman, D., Pare, G., Buring, J., Cook, N., Miletich, J., Ridker, P.: Association between a literature-based genetic risk score and cardiovascular events in women. *J. Am. Med. Assoc.* **303**(7), 631–637 (2010)
32. Pencina, M., D’Agostino, R.: Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat. Med.* **23**(13), 2109–2123 (2004)
33. Pencina, M., D’Agostino, R.S., D’Agostino, R.J., Vasan, R.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond (with commentaries & rejoinder). *Stat. Med.* **27**, 157–212 (2008)
34. Pfeiffer, R., Gail, M.: Two criteria for evaluating risk prediction models. *Biometrics* **67**(3), 1057–1065 (2010)
35. Pfeffer, M., Jarcho, J.: The charisma of subgroups and the subgroups of CHARISMA. *N. Engl. J. Med.* **354**(16), 1744–1746 (2006)
36. Ridker, P.: C-Reactive protein and the prediction of cardiovascular events among those at intermediate risk: moving an inflammatory hypothesis toward consensus. *J. Am. Coll. Cardiol.* **49**(21), 2129–2138 (2007)
37. Ridker, P., Rifai, N., Rose, L., Buring, J., Cook, N.: Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N. Engl. J. Med.* **347**, 1557–1565 (2007)
38. Robins, J., Ya’Acov, R.: Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.* **16**(3), 285–319 (1997)
39. Rothwell, P.: Treating individuals 1 external validity of randomised controlled trials: to whom do the results of this trial apply? *Lancet* **365**, 82–93 (2005)
40. Tian, L., Zucker, D., Wei, L.: On the cox model with time-varying regression coefficients. *J. Am. Stat. Assoc.* **100**(469), 172–183 (2005)
41. Tian, L., Cai, T., Wei, L.J.: Identifying subjects who benefit from additional information for better prediction of the outcome variables. *Biometrics* **65**, 894–902 (2009)
42. Tibshirani, R., Hastie, T.: Local likelihood estimation. *J. Am. Stat. Assoc.* **82**(398), 559–567 (1987)
43. Tice, J., Cummings, S., Ziv, E., Kerlikowske, K.: Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res. Treat.* **94**(2), 115–122 (2005)
44. Uno, H., Cai, T., Tian, L., Wei, L.: Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**, 527–537 (2007)
45. Uno, H., Cai, T., Pencina, M., D’Agostino, R., Wei, L.: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**(10), 1105–1117 (2011)
46. Uno, H., Cai, T., Tian, L., Wei, L.J.: Graphical procedures for evaluating overall and subject-specific incremental values from new predictors with censored event time data. *Biometrics* **67**, 1389–1396 (2011)
47. Van der vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes. Springer, New York (1996)
48. Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H., Diver, W., Thun, M., Cox, D., Hankinson, S., Kraft, P., et al.: Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**(11), 986–993 (2010)
49. Wand, M., Marron, J., Ruppert, D.: Transformation in density estimation (with comments). *J. Am. Stat. Assoc.* **86**, 343–361 (1991)
50. Wang, T., Gona, P., Larson, M., Tofler, G., Levy, D., Newton-Cheh, C., Jacques, P., Rifai, N., Selhub, J., Robins, S.: Multiple biomarkers for the prediction of first major cardiovascular events and death. *N. Engl. J. Med.* **355**, 2631–2639 (2006)

51. Wang, R., Lagakos, S., Ware, J., Hunter, D., Drazen, J.: Statistics in medicine-reporting of subgroup analyses in clinical trials. *N. Engl. J. Med.* **357**(21), 2189–2194 (2007)
52. Wilson, P.W., D’Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B.: Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998)
53. Zhao, L., Cai, T., Tian, L., Uno, H., Solomon, S., Wei, L., Minnier, J., Kohane, I., Pencina, M., D’Agostino, R., et al.: Stratifying subjects for treatment selection with censored event time data from a comparative study. Harvard University Biostatistics working paper series 2010: working paper 122 (2010)

Part III

Applications

Assessing the Effects of Imprinting and Maternal Genotypes on Complex Genetic Traits

Shili Lin

Abstract A susceptibility variant may affect a trait not only through sequence variation, but also through parental origin, and even through combination with the maternal genotype. Although associations have been established for more than one thousand five hundred Single Nucleotide Polymorphisms (SNPs) and over two hundred diseases through genome-wide association studies, imprinting and maternal genotype effects (collectively referred to as parent-of-origin effects) have largely not been taken into account. The ignorance of parent-of-origin effects may have adversely contributed to “missing heritability”; thus, attempts have been made to incorporate these two epigenetic factors when assessing the effect of a genetic variant on a complex trait. In this review, we will discuss the difference between retrospective and prospective studies in genetic analysis and indicate how this difference may influence the choice of methods for assessing parent-of-origin effects on the risk of complex genetic traits. We will provide expositions on several specific study designs and their associated analysis methods, including the case-parent triad design and designs that include control samples, such as the case-parent triads/control-parent triads design. Most available methods are for retrospective studies, but a handful of methods applicable to extended pedigrees from prospective studies also exist. Although log-linear or logistic models are frequently used to factor in parent-of-origin effects, we review non-parametric approaches as well for detecting imprinting effects. We further discuss implications of various assumptions made in the modeling to avoid overparameterization. In summary, a model factoring in epigenetically modulated gene variant effects is expected to be of greater value in risk assessment and prediction if such epigenetic factors indeed play a role in the etiology of the disease.

S. Lin (✉)

Department of Statistics, The Ohio State University, Columbus, OH, USA
e-mail: shili@stat.osu.edu

Introduction

In the past decade Genome-wide association studies (GWAS) have led to the identification of many common SNPs that are believed to be associated with complex diseases. A total of 1617 published GWAS discoveries ($P\text{-value} \leq 5 \times 10^{-8}$) for 249 traits had been reported by the third quarter of 2011 [23]. However, the associated genes tend to have small effects on diseases, odds ratios of about 1.1–1.5. Moreover, the variations explain only 5–10 % of the disease burden in the population [32, 34]. These revelations set off a vigorous debate and raised the question of where to find the “missing heritability” [15, 24, 31, 34]. This has led to a watershed moment. Researchers now contend that the search for the genetic burden of complex diseases should focus not only on common, but also on rare variants. Further, since DNA sequence polymorphism is not the only factor that contributes to phenotypic variation, incorporating other mechanisms such as epigenetic modification and transcriptional/translational regulation would provide a more systematic view of genomic effects on complex traits [20, 24].

Genomic imprinting and maternal genotype effects are two epigenetic factors that modulate the genetic variants’ effects and have been explored increasingly for their role in the etiology of complex traits [30]. Such investigations should be viewed as an integral part of association studies, not supplementary to them, as a maternal effect may be disguised as an association effect in typical case-control studies [4]. Genomic imprinting, maternal or paternal, is an effect of the epigenetic process involving methylation and histone modifications in order to silence the expression of a gene inherited from a particular parent (mother or father) without altering the genetic sequence. This process leads to unequal expression of a heterozygous genotype depending on whether the imprinted variant is inherited from the mother (maternal imprinting) or from the father (paternal imprinting), which plays a key role in normal mammalian growth and development [13].

A maternal genotype effect, on the other hand, is a situation wherein the phenotype of an individual is influenced by the genotype of the mother. Maternal effects usually occur due to the additional mRNAs or proteins passed from the mother to the fetus during pregnancy. This may result in an individual showing the phenotype due to the genotype of the mother regardless of one’s own genotype, a main effect in statistical modeling. There is a more specific kind of maternal genotype effect, which is usually known as maternal-fetal genotype incompatibility [7, 42]. This kind of incompatibility arises due to interactions between the gene products of the mother and the fetus [36] and is typically modeled as interaction effects as opposed to main effects.

The first imprinted gene in humans was found 20 years ago [14]. Since then, a variety of traits (e.g. brain development [16]) and diseases (e.g. Angelman Syndrome and Prader-Willi Syndrome [5, 12, 46]) have been found to be associated with imprinting. Although it has been estimated that about 1 % of all mammalian genes are imprinted [35], only a limited number have been identified thus far. With the availability of next generation sequencing (NGS) technology, scientists are now

able to carry out direct studies of imprinting genomewide in the mouse efficiently [16, 49]. Nevertheless, the controlled mating setup that was successful in mouse studies is not feasible in humans. Thus, robust and powerful statistical methods for detecting and assessing imprinting effects on complex genetic traits are still indispensable.

Biological research increasingly reveals the presence and importance of maternal genotype effects in birth defects and in many diseases such as childhood cancer [18, 19, 29]. There are also well-known examples of different maternal-fetal genotype incompatibility mechanisms. One such example is the RhD-induced hemolytic disease [44] due to genotype incompatibility, or “mismatch”, which occurs when the immune system of a mother with two null alleles mounts an immune response when it detects “foreign” proteins produced by a fetus carrying a copy of the antigen coding allele. Another example represents a mechanism known as NIMA (non-inherited maternal antigen) as it occurs in rheumatoid arthritis [21], in which a mother’s genotype with a copy of the allele coding for an antigen at the HLA-DRB1 locus will increase the risk of her offspring when the child does not carry such an allele. Although imprinting and maternal effects arise from two different biological processes, they could give rise to “maternal lineage” of the same trait [17, 54]. Thus, it is important that these two confounding factors be studied jointly to evaluate their risk effects.

In this review, we will discuss statistical methods for assessing the effects of maternal genotype and/or imprinting on complex diseases considering several aspects of the study design (retrospective vs. prospective, nuclear family vs. extended family, affected family only vs. case-control families), data availability (complete data vs. missing father/missing a large number of individuals in a pedigree), and population genetic assumptions (no assumption vs. Hardy-Weinberg equilibrium (HWE)/mating symmetry). The focus will be on methods for binary disease traits, although methods for quantitative traits have also been proposed for assessing parent-of-origin effects.

The completion of the Human Genome Project and advances in biotechnology, including the microarray and NGS, coupled with successful identifications of genetic risk variants, have made it a reality to use the identified genetic markers for risk prediction [28, 31]. Several commercial companies even offer “risk tests” for specific diseases by marketing directly to consumers [9, 33]. However, genetically based risk assessment, with currently known variants may not be of much predictive value clinically due to their limited contributions to trait variability. Rare variants may be a key to increasing the value of genetic risk prediction [31, 33, 34]. Epigenetic factors, such as imprinting and maternal genotype effects, may also contribute significantly to genetic-based risk assessment and prediction [20, 24]. For instance, if a genetic trait is influenced by a maternally imprinted gene, then risk prediction will not only depend on whether an individual carries a copy of the risk allele but, more importantly, on whether the risk allele is inherited from the mother or from the father. Inheriting the risk allele from the father will lead to a higher risk of getting the disease, whereas inheriting the risk allele from the mother will not increase, or

will at least moderate, the risk. Most of the methods discussed in this review are model based; they can be utilized to predict an individual's risk for a disease that is influenced by genetic variants whose effects are modulated epigenetically.

Study Design

Population data do not contain information on parental transmission and therefore studies of imprinting and maternal genotype effects are family based. Nevertheless, control families or even unrelated controls may be used to help inferring population parameters [1, 11, 48, 52]. Both prospective and retrospective study designs can be utilized, although the predominant design currently is a retrospective study based on nuclear families. Retrospective family-based association studies recruit families with children affected by a certain disease. Thus the familial genotypes should be viewed as data conditional on the affected children, the probands. On the other hand, prospective studies do not specifically recruit families with a certain disease. Instead, families in prospective studies are recruited and followed over time, and therefore such studies tend to include extended families, and family members can be affected or unaffected with any disease at a given time. Data from such a design can then be used to uncover genetic associations with various complex disorders rather than just one disease as in a retrospective study [57].

Retrospective Studies

There are several frequently used study designs for retrospective studies. The earliest was the case-parent triads design, in which the genotype of an affected child (proband), together with the genotypes of both parents, are obtained and analyzed [50, 51, 53]. There are 15 possible combinations of triad genotypes (Table 1, top segment). Log-linear and logistic models have been proposed to analyze data of this kind [1]. Such models are usually parameterized in terms of the effects of the child's genotype (up to 3 parameters including the phenocopy rate), imprinting effects (up to 2 parameters), maternal genotype effects (up to 2 parameters), and interaction effects (up to 6 parameters including those signifying maternal-fetal incompatibility). In addition to these risk assessment parameters, there are also mating type probabilities (up to 9 parameters), the nuisance parameters. Table 1 (Columns 5 and 6) shows the joint probability of a child's affection status and the triad genotypes based on a log-linear risk model. Apparently, the full model is not practicable; not all parameters are identifiable since the number exceeds the number of data categories. Supplementary data are then often used to expand the capability of such models. This type of expansions leads to several other designs, which include

Table 1 Joint probabilities of disease status and triad genotypes^a

Type	<i>M</i>	<i>F</i>	<i>C</i>	$P(D = 1, M, F, C)$	$P(D = 0, M, F, C)$
1	0	0	0	$\mu_{00} \cdot 1 \cdot \delta$	$\mu_{00} \cdot 1 \cdot [1 - \delta]$
2	0	1	0	$\mu_{01} \cdot \frac{1}{2} \cdot \delta$	$\mu_{01} \cdot \frac{1}{2} \cdot [1 - \delta]$
3	0	1	1	$\mu_{01} \cdot \frac{1}{2} \cdot \delta R_1 \gamma_{01}$	$\mu_{01} \cdot \frac{1}{2} \cdot [1 - \delta R_1 \gamma_{01}]$
4	0	2	1	$\mu_{02} \cdot 1 \cdot \delta R_1 \gamma_{01}$	$\mu_{02} \cdot 1 \cdot [1 - \delta R_1 \gamma_{01}]$
5	1	0	0	$\mu_{10} \cdot \frac{1}{2} \cdot \delta S_1 \gamma_{10}$	$\mu_{10} \cdot \frac{1}{2} \cdot [1 - \delta S_1 \gamma_{10}]$
6	1	0	1	$\mu_{10} \cdot \frac{1}{2} \cdot \delta S_1 R_1 R_{im} \gamma_{11}$	$\mu_{10} \cdot \frac{1}{2} \cdot [1 - \delta S_1 R_1 R_{im} \gamma_{11}]$
7	1	1	0	$\mu_{11} \cdot \frac{1}{4} \cdot \delta S_1 \gamma_{10}$	$\mu_{11} \cdot \frac{1}{4} \cdot [1 - \delta S_1 \gamma_{10}]$
8	1	1	1	$\mu_{11} \cdot \frac{1}{4} \cdot \delta S_1 R_1 (1 + R_{im}) \gamma_{11}$	$\mu_{11} \cdot \frac{1}{4} \cdot [2 - \delta S_1 R_1 (1 + R_{im}) \gamma_{11}]$
9	1	1	2	$\mu_{11} \cdot \frac{1}{4} \cdot \delta S_1 R_2 \gamma_{12}$	$\mu_{11} \cdot \frac{1}{4} \cdot [1 - \delta S_1 R_2 \gamma_{12}]$
10	1	2	1	$\mu_{12} \cdot \frac{1}{2} \cdot \delta S_1 R_1 \gamma_{11}$	$\mu_{12} \cdot \frac{1}{2} \cdot [1 - \delta S_1 R_1 \gamma_{11}]$
11	1	2	2	$\mu_{12} \cdot \frac{1}{2} \cdot \delta S_1 R_2 \gamma_{12}$	$\mu_{12} \cdot \frac{1}{2} \cdot [1 - \delta S_1 R_2 \gamma_{12}]$
12	2	0	1	$\mu_{20} \cdot 1 \cdot \delta S_2 R_1 R_{im} \gamma_{21}$	$\mu_{20} \cdot 1 \cdot [1 - \delta S_2 R_1 R_{im} \gamma_{21}]$
13	2	1	1	$\mu_{21} \cdot \frac{1}{2} \cdot \delta S_2 R_1 R_{im} \gamma_{21}$	$\mu_{21} \cdot \frac{1}{2} \cdot [1 - \delta S_2 R_1 R_{im} \gamma_{21}]$
14	2	1	2	$\mu_{21} \cdot \frac{1}{2} \cdot \delta S_2 R_2 \gamma_{22}$	$\mu_{21} \cdot \frac{1}{2} \cdot [1 - \delta S_2 R_2 \gamma_{22}]$
15	2	2	2	$\mu_{22} \cdot 1 \cdot \delta S_2 R_2 \gamma_{22}$	$\mu_{22} \cdot 1 \cdot [1 - \delta S_2 R_2 \gamma_{22}]$
				$P(D = 1, M, C)$	$P(D = 0, M, C)$
1, 2	0	—	0	$(\mu_{00} + \frac{1}{2} \mu_{01}) \cdot \delta$	$(\mu_{00} + \frac{1}{2} \mu_{01}) \cdot [1 - \delta]$
3, 4	0	—	1	$(\frac{1}{2} \mu_{01} + \mu_{02}) \cdot \delta R_1 \gamma_{01}$	$(\frac{1}{2} \mu_{01} + \mu_{02}) \cdot [1 - \delta R_1 \gamma_{01}]$
5, 7	1	—	0	$(\frac{1}{2} \mu_{10} + \frac{1}{4} \mu_{11}) \cdot \delta S_1 \gamma_{10}$	$(\frac{1}{2} \mu_{10} + \frac{1}{4} \mu_{11}) \cdot [1 - \delta S_1 \gamma_{10}]$
6, 8, 10	1	—	1	$\frac{1}{2} \mu_{10} \cdot \delta S_1 R_1 R_{im} \gamma_{11}$ $+ \frac{1}{4} \mu_{11} \cdot \delta S_1 R_1 (1 + R_{im}) \gamma_{11}$ $+ \frac{1}{2} \mu_{12} \cdot \delta S_1 R_1 \gamma_{11}$	$\frac{1}{2} \mu_{10} \cdot [1 - \delta S_1 R_1 R_{im} \gamma_{11}]$ $+ \frac{1}{4} \mu_{11} \cdot [2 - \delta S_1 R_1 (1 + R_{im}) \gamma_{11}]$ $+ \frac{1}{2} \mu_{12} \cdot [1 - \delta S_1 R_1 \gamma_{11}]$
9, 11	1	—	2	$(\frac{1}{4} \mu_{11} + \frac{1}{2} \mu_{12}) \cdot \delta S_1 R_2 \gamma_{12}$	$(\frac{1}{4} \mu_{11} + \frac{1}{2} \mu_{12}) \cdot [1 - \delta S_1 R_2 \gamma_{12}]$
12, 13	2	—	1	$(\mu_{20} + \frac{1}{2} \mu_{21}) \cdot \delta S_2 R_1 R_{im} \gamma_{21}$	$(\mu_{20} + \frac{1}{2} \mu_{21}) \cdot [1 - \delta S_2 R_1 R_{im} \gamma_{21}]$
14, 15	2	—	2	$(\frac{1}{2} \mu_{21} + \mu_{22}) \cdot \delta S_2 R_2 \gamma_{22}$	$(\frac{1}{2} \mu_{21} + \mu_{22}) \cdot [1 - \delta S_2 R_2 \gamma_{22}]$

^a*M*, *F*, and *C* are the number of variant allele(s) carried by the mother, the father and the child in a family, which are equal to 0, 1 or 2; *F* = — indicates that paternal genotype is missing in case-mother and control-mother pairs; μ_{mf} denotes the mating type probability of (*M*, *F*) = (*m*, *f*), that is, probability of parental pairs in which the mothers carry *m* copies and the fathers *f* copies of the variant allele; δ is the phenocopy rate of the disease in the population; *R*₁ and *R*₂ are relative risks due to 1 and 2 copies of the variant allele carried by the offspring, respectively; *R*_{im} is the relative risk due to the single copy of the variant allele being inherited from the mother (another imprinting parameter may also be introduced, see [1]); *S*₁ and *S*₂ are the maternal effect of 1 and 2 copies of the variant allele carried by the mother, respectively; the $\gamma_{m,c}$'s are the interaction effects between the mother's genotype *m* and the child's genotype *c*, where γ_{01} is measuring the "mismatch" incompatibility and γ_{10} is measuring the NIMA effect

data that can be useful for estimating mating type probabilities¹ [1]. Such designs include case-parent triads + controls [11], case-parent triads + control parents [52], case-parent triads + control-parent triads [1, 55], case-parent triads + control-mother pairs [48]. The last design reflects the concern that fathers of controls are typically much more difficult to recruit than mothers of controls. The same concern also arises in recruiting case families, and thus designs that do away with fathers or allow for missing fathers have also been considered: case-mother pairs + control-mother pairs [41], case-parent/control-parent triads + case-mother/control-mother pairs [1, 55, 58]. With the additional data, more parameters are identifiable and can be estimated. Regardless, the full model (even without the parameters for interactions) is still overparameterized with a full likelihood approach, and, as such, various assumptions are made to reduce the parameter space. A partial likelihood solution, on the other hand, is able to overcome the problem by circumventing the need to estimate the nuisance parameters [58]. Other extensions to the retrospective design are mainly for the purpose of more fully utilizing the available data, including nuclear families with multiple affected children [42] and, further, allowing for one of the parents to be missing [60]. Extended pedigrees have also been considered [7, 62], although such data maybe more appropriately analyzed using alternative methods if they come from a prospective family design, so that unaffected individuals in the pedigree can also contribute to the parameter estimation and hypothesis testing [57].

Prospective Studies

Although retrospective designs are popular for assessing the role of imprinting and maternal genotype on disease risk, there are numerous prospective family-based association studies. Well-known family-based epidemiology projects that are prospective in nature include studies of genetic isolates such as the Hutterite [3, 45] and the Amish populations [10]. Other large studies include the deCODE project [37], the Busselton Health Study [26], and the Framingham Heart Study (FHS) [43]. There is little doubt that genotype profiles of large families from such prospective cohort studies will become more and more available [27, 39] as genotyping technique advances and cost reduces. Such data could lead to uncovering genetic association with various complex diseases, including consideration of pleiotropy, rather than just the one disease ascertained in a retrospective study. Data from prospective studies have been analyzed using methods devised for data from retrospective studies [7, 56, 62]. In such analyses, unaffected siblings are typically ignored or used only for aiding the estimation of population parameters. A more appropriate alternative is to model the joint likelihood of genotypes and disease status, and use both affected and unaffected children through a generalized linear model [56, 57]. By utilizing the data more fully and appropriately, one can obtain a higher power for detecting imprinting and maternal genotype effects [57].

¹Although such data are typically only used to help estimating mating type probabilities, they can contribute to the estimation of risk parameters under certain formulations [57].

Statistical Methods

We first discuss methods that can be used to detect imprinting at an associated locus assuming there is no maternal genotype effect. Such methods are usually very simple and powerful if there is indeed no maternal effect. However, when the assumption is violated, there can be severely inflated type I error rate or reduced power [57]. We then turn our attention to methods that are designed to detect only maternal genotype effects assuming there is absence of imprinting. Such methods may also suffer from large biases in the parameter estimates and inflated type I error rates should the assumption be violated [57]. Due to the confounding between imprinting and maternal genotype effects [17], it is important that both effects be accounted for simultaneously in a statistical test of parent-of-origin effects. If, however, there is a priori and unequivocal information that either imprinting or maternal effect is indeed absent, then a method that assumes the null effect of such factor would indeed be more powerful and appropriate.

Detection of Imprinting Assuming No Maternal Effect

Based on the assumption that there is association between the trait and the genetic variant of interest but no maternal genotype effect, the parental-asymmetry test (PAT) [50, 60, 62] is a simple but powerful method for testing imprinting effect. It basically measures whether there is an imbalance between the number of the variant allele inherited from the mother and that from the father among affected children. The original study design considered by PAT is case-parent triads in retrospective studies. Due to familial aggregation of genetic diseases, it is likely that an affected child may have affected siblings. Therefore, the original PAT has been extended to nuclear families with an arbitrary number of affected children [60]. For a family (the i th family) with n_i children, we denote the genotype scores (the number of the variant allele carried by an individual) of the mother, father and affected child j by M_i, F_i and $C_{ij}, 1 \leq j \leq n_i$, respectively, which take values in $\{0, 1, 2\}$. To measure the imbalance, we consider the statistic

$$\sum_{j=1}^{n_i} [I(F_i > M_i, C_{ij} = 1) - I(F_i < M_i, C_{ij} = 1)],$$

where $F_i > M_i, C_{ij} = 1$ ($F_i < M_i, C_{ij} = 1$) represents the event that child j 's only variant allele is inherited from the father (mother), and I is the usual indicator function that takes the value of 1 if the event inside the parentheses is true and 0 otherwise. Since the contributions from multiple siblings within a family are correlated, such correlations need to be taken into account in computing the variance of the statistic. Assuming the availability of n nuclear families with independent contributions and assuming mating symmetry in the population lead to the following test statistic that is distributed as $N(0, 1)$ asymptotically [60]:

$$\text{PAT} = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} [I(F_i > M_i, C_{ij} = 1) - I(F_i < M_i, C_{ij} = 1)]}{\sqrt{\sum_{i=1}^n \left[\sum_{j=1}^{n_i} I(F_i \neq M_i, C_{ij} = 1) + 2 \sum_{j < k} I(F_i \neq M_i, C_{ij} = 1, C_{ik} = 1) \right]}}$$

PAT's further extension, the pedigree parental-asymmetry test (PPAT) and the Monte Carlo pedigree parental-asymmetry test (MCPAT), can accommodate extended families and missing genotypes [62]. These model-free tests have been implemented in software packages and are powerful tools for detecting imprinting effects. Nevertheless, caution needs to be exercised when applying these methods. As we mentioned earlier, the assumption of no maternal genotype effect should not be taken lightly. Due to confounding, an unaccounted maternal effect may magnify paternal imprinting while canceling out maternal imprinting. For analysis that utilize family data, it would be a good idea to ascertain whether the data were collected prospectively or retrospectively since PAT type tests are designed for retrospective designs.

Detection of Maternal Genotype Effect Assuming No Imprinting

For retrospective study designs that recruit child-mother pairs only, there is little information contained in such data about imprinting, and therefore, the focus is typically on studying the maternal genotype main effect assuming the absence of imprinting [41]. Such an assumption is also due to practical reason, as the log-linear or logistic models will otherwise be overparameterized. The models and methods in such scenarios are special cases of those discussed in section "Joint Consideration of Imprinting and Maternal Effects: Retrospective Studies" and will not be discussed separately.

There are also specialty tests for detecting interaction effects between the genotypes of mother and child, the maternal-fetal genotype incompatibility test (MFG) [42]. Under the case-parent triad design, a number of scenarios are considered, corresponding to different parametrizations of the log-linear model for connecting the disease phenotype to the case-parent triad genotypes depending on different hypotheses of biological interactions. The method for case-parent design has been extended to nuclear families or child-mother pairs data [6, 8]. The log-linear approach in the MFG test may also be viewed as a special case of those discussed in section "Joint Consideration of Imprinting and Maternal Effects: Retrospective Studies".

A more recent extension, the extended-MFG test [7], considers extended families with arbitrary pedigree structures. For a pedigree with complete genotype vector \mathbf{G} and trait phenotype vector \mathbf{D} , its contribution to the likelihood is the conditional probability

$$P(\mathbf{G}|\mathbf{D}) = \frac{P(\mathbf{G}, \mathbf{D})}{P(\mathbf{D})} = \frac{P(\mathbf{D}|\mathbf{G})P(\mathbf{G})}{\sum_{\mathbf{G}} P(\mathbf{D}|\mathbf{G})P(\mathbf{G})}.$$

The penetrance contained in $P(\mathbf{D}|\mathbf{G})$ can be modeled as log-linear. Under the assumption of random mating, the probability of the pedigree genotype $P(\mathbf{G})$ can be factored into the products of founder probabilities and Mendelian transmission probabilities assuming HWE. Discussion on how to relax this assumption is given at the end of section “Joint Consideration of Imprinting and Maternal Effects: Prospective Studies”. If there is missing data, the probability will sum over all possible genotypes that are compatible with the observed ones.

Joint Consideration of Imprinting and Maternal Effects: Retrospective Studies

The most frequently used model for connecting the underlying family triad genotypes with disease risk for the child are either log-linear or logistic. These two models would provide equivalent inferences unless additional constraints are imposed in the log-linear formulation [41]. As in section “Detection of Imprinting Assuming No Maternal Effect”, we denote the genotype scores of the mother, father and child in a triad by M , F and C , respectively, which take values in $\{0, 1, 2\}$. The disease status $D = 1$ indicates that the child is affected, $D = 0$ otherwise. Let $\eta = E[D|M, F, C] = P(D = 1|M, F, C)$. Then the log-linear/logic model can be expressed as:

$$g(\eta) = \delta R_1^{I(C=1)} R_2^{I(C=2)} R_{im}^{I(C=1 \& \text{origin}=M)} S_1^{I(M=1)} S_2^{I(M=2)} \gamma_{01}^{I(M=0, C=1)} \gamma_{10}^{I(M=1, C=0)} \gamma_{11}^{I(M=1, C=1)} \gamma_{12}^{I(M=1, C=2)} \gamma_{21}^{I(M=2, C=1)} \gamma_{22}^{I(M=2, C=2)}, \quad (1)$$

where g , the link function, is $\eta/(1 - \eta)$ for the logit model or the identity function for the log-linear model. For the parameters in the model, δ is the phenocopy rate of the disease; R_1 and R_2 are the variant allele effect of 1 and 2 copies carried by the child, respectively; R_{im} is the effect when the single copy of the variant allele carried by the child is inherited from the mother; S_1 and S_2 are the maternal effect when the mother carries 1 and 2 copies of the variant allele, respectively; the γ parameters denote the interaction effects; $I(\cdot)$ is the usual indicator function that is equal to 1 or 0 depending on whether the condition within the parentheses is met or not. A similar model can be written down for child-mother pairs (Table 1, bottom segment). Based on the parametrization, $R_{im} = 1$ signifies no imprinting effect; $S_1 = S_2 = 1$ indicates no maternal genotype (main) effects, whereas $\gamma_{01} = \gamma_{10} = \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 1$ indicates no interaction effects between the mother and the child’s genotypes. We further note that a model with only interaction effect γ_{01} or γ_{10} codes for the RhD type “mismatch” or NIMA.

Taking the log-linear model as an example, one can write down the contribution of a case-parent triad to the likelihood:

$$\begin{aligned} P(M, F, C|D = 1) &= P(D = 1|M, F, C)P(M, F, C)/P(D = 1) \\ &= \eta\mu_{MF}P(C|M, F)/P(D = 1), \end{aligned}$$

where η is as defined in Eq. (1) with g taken to be the identity function; $\mu_{MF} = P(M, F)$ is the mating type probability; $P(C|M, F)$ is the Mendelian transmission probability; $P(D = 1)$ is the disease prevalence. There are 9 possible mating types for a SNP, but the total probabilities need to sum to 1, so there are 8 independent nuisance parameters. Given the large number of parameters, assumptions about genotype risks (e.g., only main effects are considered) and mating type probabilities (e.g., mating symmetry or HWE) are typically made to avoid overparameterization in a full likelihood approach [1].

A partial likelihood approach using the case-parent triads/control-parent triads design but allowing for missing fathers in both cases and controls circumvents the need to estimate the mating type probabilities [58]. The key is the recruitment of control families of the same structure as case families, thus creating “internal matches” stratified by the familial genotypes. A partial likelihood component can then be extracted from the full likelihood of the retrospective design. This partial likelihood can be thought of as the products of likelihoods from stratified prospective designs according to the triad/pair genotypes. Through conditional on the familial genotypes, the partial likelihood is free of the nuisance parameters with respect to the population mating type probabilities. Therefore, it is no longer necessary to make any assumption about mating type probabilities. This makes the partial likelihood approach more robust and more efficient by reducing the parameter space. However, there is a trade-off. The data in which both the mother and the child are heterozygous while the father’s genotype is missing cannot be used in the procedure, which may lead to reduction in power in some situations.

Joint Consideration of Imprinting and Maternal Effects: Prospective Studies

For prospective studies, the familial genotypes and the disease data need to be modeled jointly. Since such studies usually involve extended pedigrees, it is rather common that some of the genotypes are missing in a family. Let $\mathbf{G} = (G_1, \dots, G_c, \dots, G_n)$ denote the genotype scores of all n members of a family, with the c non founders preceding the founders. Since some of the genotype scores may be unavailable, we further divide \mathbf{G} into \mathbf{G}_o , the observed scores, and \mathbf{G}_m , the missing scores. We let $\mathbf{D} = (D_1, D_2, \dots, D_c)$ denote the binary disease status of all c non founders. Further, F_k and M_k denote the genotype score of the father and mother of non founder k in the pedigree. Assuming that the disease status of the nonfounders

are conditionally independent given the familial genotypes, the contribution of this family to the likelihood is the joint probability of all observed genotypes and disease status:

$$P(\mathbf{G}_o, \mathbf{D}) = \sum_{\mathbf{G}_m} P(\mathbf{G}_o, \mathbf{G}_m, \mathbf{D}) = \sum_{\mathbf{G}_m} P(\mathbf{G}_o, \mathbf{G}_m) \prod_{k=1}^c P(D_k | M_k, F_k, G_k).$$

The penetrance $P(D_k | M_k, F_k, G_k)$ can be modeled as before using Eq. (1). To compute the probability of the familial genotypes, $P(\mathbf{G}_o, \mathbf{G}_m)$, one can factor it into the products of the probabilities of founders and each nonfounder conditioning on the parental genotypes. The latter, the conditional probabilities, are simply transmission probabilities under Mendel’s law of segregation. Although random mating and HWE are usually assumed when computing founder genotype probabilities, as in section “Detection of Maternal Genotype Effect Assuming No Imprinting”, these assumptions are strong and likely to be violated in reality. A solution to avoid such strong assumption is to model each founder couple jointly and to model a married-in founder conditional on the spouse’s genotypes [57].

Missing genotypes have long been a concern in family-based association studies. A widely employed strategy for handling missing data is to recover missing information based on what have been observed. For pedigrees with a moderate number of missing genotypes, it is feasible to enumerate all possible unobserved genotypes that are compatible with the observed genotypes of other family members [57]. This practice can be very fruitful, as the power can be much higher than simply excluding individuals with missing genotypes from the analysis. However, for large pedigrees with a lot of missing genotypes, enumeration may become impracticable. Thus, computational methods such as reverse peeling [38] may be considered.

Assumptions and Their Effects

Mating Type Probabilities and Population Stratification

Hardy Weinberg equilibrium is the strongest assumption that portrays random mating. This assumption is needed for parameter identifiability with limited data, especially in the case-mother design [1]. A less stringent assumption is mating symmetry, which is almost universally assumed to avoid overparameterization [1], with only a few exceptions [48, 58]. However, if there is gender-specific assortative mating, then the assumption of mating symmetry does not hold anymore. It has been shown that some of the methods in the literature are not robust to departure from the mating symmetry assumption [42, 58], and can lead to greatly inflated type I errors. The partial likelihood approach discussed in section “Joint Consideration of Imprinting and Maternal Effects: Retrospective Studies” deviates from the rest by

circumventing the need to estimate the mating type probabilities, and thus is robust to violation of the underlying population mating type probability assumptions.

For study designs that recruit both case and control families, the effect of population stratification is of concern. The degree of influence of stratification depends on the assumptions about mating type probabilities. For a population consisting of two subpopulations each in HWE, the population as a whole is no longer in HWE. Thus, population stratification will have a profound effect on any method that assumes HWE. On the other hand, although mating symmetry, or even parental allelic exchangeability (a stronger condition than mating symmetry [41]), continues to hold in the whole population if the assumption is true in each of the subpopulation, there can still be considerable biases in parameter estimates and inflated type I error rates for methods that require estimation of the mating type probabilities [58]. This is in part due to other hidden assumptions, as discussed in the following subsection. In contrast, for a method that circumvents the need to estimate such probabilities, the effect is much smaller, even when the disease prevalence is different in the subpopulations [58].

Disease Rarity

The assumption that the probabilities of child-mother pair genotype combinations in the controls are approximately the same as in the general population is sometimes made in the literature. It is typically argued that rare disease is a sufficient condition for the assumption to hold [1,41]. Although the rationale seems plausible, analytical as well as simulation results indicate that the rare disease assumption is only a necessary, not a sufficient, condition, for the frequencies to be roughly equal. It is the interplay of allele frequency and the underlying genetic model, not the rare disease assumption alone, that determines whether the pair frequencies are roughly equal [58]. The rarity assumption has driven other assumptions about population frequency relationships [41], which is a hidden factor contributing to the biases and inflated type I errors seen in methods that make such an assumption [58].

Multiple SNPs and Haplotypes

The methods discussed in this review are all single-SNP based, in that each SNP is examined one at a time for its association with a trait, although SNPs are available genomewide in the order of hundreds of thousands or even millions. A common approach to utilize multiple SNPs is through haplotype analysis, which can be more powerful for detecting association in complex diseases, especially when (1) the causative variant is not investigated directly, or (2) when there are multiple disease-causing alleles, or (3) when there are several mutations within the same gene in cis formation [25,59]. Nevertheless, availability of haplotype-based methods

is limited compared to SNPs-based ones, partially due to computational intensity. Among methods for haplotype association studies, there are only a handful that account for imprinting and/or maternal effects [2, 8, 61]. A rather interesting use of haplotype is to aid deduction of parental origin for imprinting analysis at a test locus using data from distant relatives [30]. This approach effectively borrows information from neighboring markers to enhance one's chance of identifying the parental origin of an individual's genes when information on both parents are missing and the individual is otherwise not informative for imprinting analysis without bringing in extra information.

Environmental Covariates and Quantitative Traits

Environmental factors, especially gene-environment interactions, are believed to be another type of major contributors to the "missing heritability". Although the log-linear or logistic regression models discussed in this review can, in theory, easily accommodate such factors, there are practical limitations. Even without parameters representing environmental covariate effects, the models are already overparameterized for studying binary traits. It would likely be more promising to consider the main and interacting effects of environmental factors for traits that are measured quantitatively. Several methods have been proposed for assessing imprinting effects for quantitative traits [22, 40, 47], but investigation on environmental covariates is limited. In [22], the environmental covariate effect was first regressed out before carrying out an analysis to study imprinting effects. It was shown that there was a gain in power after taking into account the covariate, although gene-environment interactions cannot be investigated in this two-step approach.

Concluding Remarks

Since epigenetic factors such as imprinting and maternal genotype effects may contribute to the explanation of "missing heritability", there is an increasing interest in factoring in such effects in studies that assess the effects of genetic variants, with the goal of achieving a better understanding of the underlying genetic mechanism. Most of the methods in the literature are for qualitative traits with a retrospective study design. However, there are large epidemiologic studies that are prospective in nature with extended pedigrees, and thus methods that can fully utilize such data are clearly needed. Overparametrization is a major concern; assumptions needed to avoid such a problems are difficult to check and likely to be unrealistic. The partial likelihood approach is a step in the direction of reducing the parameter space without commonly made assumptions, leading to a more robust and efficient procedure. Environmental factors, especially their interacting effects with genetic variants, are hypothesized to contribute significantly to missing heritability as well.

Thus, more research in this direction, especially for quantitative traits, is warranted. Genetic variants whose effects are modulated epigenetically through imprinting or maternal genotypes can be of greater value in disease risk assessment and prediction by including such epigenetic factors.

Acknowledgements This work was supported in part by the National Science Foundation grant DMS-1208968. The author would like to thank Dr. Lynn Friedman for her valuable comments on an earlier version of the manuscript.

References

1. Ainsworth, H.F., Unwin, J., Jamison, D.L., Cordell, H.J.: Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet. Epidemiol.* **35**, 19–45 (2011)
2. Becker, T., Baur, M.P., Knapp, M.: Detection of parent-of-origin effects in nuclear families using haplotype analysis. *Hum. Hered.* **62**, 64–76 (2006)
3. Boycott, K.M., Parboosingh, J.S., Chodirker, B.N., Lowry, R.B., McLeod, D.R., Morris, J., Greenberg, C.R., Chudley, A.E., Bernier, F.P., Midgley, J., Moller, L.B., Innes, A.M.: Clinical genetics and the Hutterite population: a review of Mendelian disorders. *Am. J. Med. Genet. Part A* **146A**, 1088–1098 (2008)
4. Buyske, S.: Maternal genotype effects can alias case genotype effects in case-control studies. *Eur. J. Hum. Genet.* **16**, 784–785 (2008)
5. Cassidy, S.B., Driscoll, D.J.: Prader-Willi syndrome. *Eur. J. Hum. Genet.* **17**, 3–13 (2009)
6. Chen, J., Zheng, H., Wilson, M.L.: Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications. *Genet. Epidemiol.* **33**, 526–538 (2009)
7. Childs, E.J., Palmer, C.G., Lange, K., Sinsheimer, J.S.: Modeling maternal-offspring gene-gene interactions: the extended-MFG test. *Genet. Epidemiol.* **34**, 512–521 (2010)
8. Cordell, H.J., Barratt, B.J., Clayton, D.G.: Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet. Epidemiol.* **26**, 167–185 (2004)
9. Couzin, J.: Genetics: DNA test for breast cancer risk draws criticism. *Science* **322**, 357 (2008)
10. Edwards, D.R.V., Gilbert, J.R., Jiang, L., Gallins, P.J., Caywood, L., Creason, M., Fuzzell, D., Knebusch, C., Jackson, C.E., Pericak-Vance, M.A., Haines, J.L., Scott, W.K.: Successful aging shows linkage to chromosomes 6, 7, and 14 in the Amish. *Ann. Hum. Genet.* **75**, 516–528 (2011)
11. Epstein, M., Veal, C., Trembath, R., Barker, J., Li, C., Satten, G.: Genetic association analysis using data from triads and unrelated subjects. *Am. J. Hum. Genet.* **76**, 592–608 (2005)
12. Falls, J.G., Pulford, D.J., Wylie, A.A., Jirtle, R.L.: Genomic imprinting: implications for human disease. *Am. J. Pathol.* **154**, 635–647 (1999)
13. Ferguson-Smith, A.C.: Genome imprinting: the emergence of an epigenetic paradigm. *Nat. Rev.* **12**, 565–575 (2011)
14. Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C.G., Polychronakos, C.: Parental genomic imprinting of the human IGF2 gene. *Nat. Genet.* **4**, 98–101 (1993)
15. Goldstein, D.B.: Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009)
16. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D., Dulac, C.: High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643–648 (2010)

17. Hager, R., Cheverud, J.M., Wolf, J.B.: Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics* **178**, 1755–1762 (2008)
18. Haig, D.: Genetic conflicts in human pregnancy. *Q. Rev. Biol.* **68**, 495–532 (1993)
19. Haig, D.: Evolutionary conflicts in pregnancy and calcium metabolism – a review. *Placenta* **25**, S10–S15 (2004)
20. Hardy, J., Singleton, A.: Genomewide association studies and human disease. *N. Engl. J. Med.* **360**, 1759–1768 (2009)
21. Harney, S., Newton, J., Milicic, A., Brown, M.A., Wordsworth, B.P.: Non-inherited maternal HLA alleles are associated with rheumatoid arthritis. *Rheumatology* **42**, 171–174 (2003)
22. He, F., Zhou, J.Y., Hu, Y.Q., Sun, F., Yang, J., Lin, S., Fung, W.K.: Detection of parent-of-origin effects for quantitative traits in complete and incomplete nuclear families with multiple children. *Am. J. Epidemiol.* **174**, 226–233 (2011)
23. Hindorf, L.A., Junkins, H.A., Hall, P.N., Mehta, J.P., Manolio, T.A.: A catalog of published genome-wide association studies. Available from www.genome.gov/gwastudies (2010). Accessed 15 Oct 2012
24. Hirschhorn, J.N.: Genomewide association studies – illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009)
25. Huang, B.E., Amos, C.I., Lin, D.Y.: Detecting haplotype effects in genomewide association studies. *Genet. Epidemiol.* **31**, 803–812 (2007)
26. Jamrozik, E.F., Knuiiman, M.W., James, A., Divitini, M., Musk, A.W.: Risk factors for adult-onset asthma: a 14-year longitudinal study. *Respirology* **14**(6), 814–821 (2009)
27. Jamrozik, E.F., Warrington, N., Mcclenaghan, J., Hui, J., Musk, A.W., James, A., Beilby, J.P., Hansen, J., De Klerk, N.H., Palmer, L.J.: Functional haplotypes in the PTGDR gene fail to associate with asthma in two Australian populations. *Respirology* **16**, 359–366 (2010)
28. Janssens, A.C.J.W., Ioannidis, J.P.A., van Dujin, C.M., Little, J., Khoury, M.J.: Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. *Genome Med.* **3**, 16 (2011)
29. Jensen, L.E., Etheredge, A.J., Brown, K.S., Mitchell, L.E., Whitehead, A.S.: Maternal genotype for the monocyte chemoattractant protein 1 A(-2518)G promoter polymorphism is associated with the risk of spina bifida in offspring. *Am. J. Med. Genet.* **140A**, 1114–1118 (2006)
30. Kong, A., Steinthorsdottir, V., Masson, G., et al.: Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009)
31. Kraft, P., Hunter, D.J.: Genetic risk prediction – are we there yet? *N. Engl. J. Med.* **360**, 1701–1703 (2009)
32. Maher, B.: Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008)
33. Manolio, T.A.: Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010)
34. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorf, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al.: Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009)
35. Morison, I.M., Paton, C.J., Cleverley, S.D.: The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29**, 275–276 (2001)
36. Ober, C.: HLA and pregnancy: the paradox of the fetal allograft. *Am. J. Hum. Genet.* **62**, 1–5 (1998)
37. Peltonen, L., Palotie, A., Lange, K.: Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* **1**, 182–190 (2000)
38. Ploughman, L., Boehnke, M.: Estimating the power of a proposed linkage study for a complex genetic trait. *Am. J. Hum. Genet.* **44**, 543–551 (1989)
39. Scuteri, A., Sanna, S., Chen, W., Uda, M., Albai, G., et al.: Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**(7), e1151 (2007)
40. Shete, S., Amos, C.I.: Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *Am. J. Hum. Genet.* **70**, 751–757 (2002)

41. Shi, M., Umbach, D.M., Vermeulen, S.H., Weinberg, C.R.: Making the most of case-mother/control-mother studies. *Am. J. Epidemiol.* **168**, 541–547 (2008)
42. Sinsheimer, J.S., Palmer, C.G.S., Woodward, J.A.: Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incomparability test. *Genet. Epidemiol.* **24**, 1–13 (2003)
43. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., et al.: The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.* **165**, 1328–1335 (2007)
44. Strachan, T., Read, A.P. (eds.): *Human Molecular Genetics*, 2nd edn. Wiley, New York (1999)
45. Thompson, E.E., Sun, Y., Nicolae, D., Ober, C.: Shades of gray: a comparison of linkage disequilibrium between Hutterites and Europeans. *Genet. Epidemiol.* **34**, 133–139 (2010)
46. Van Buggenhout, G., Fryns, J.P.: Angelman syndrome (AS, MIM 105830). *Eur. J. Hum. Genet.* **17**, 1367–1373 (2009)
47. van den Oord, E.J.: The use of mixture models to perform quantitative tests for linkage disequilibrium, maternal effects, and parent-of-origin effects with incomplete subject-parent triads. *Behav. Genet.* **30**, 335–343 (2000)
48. Vermeulen, S.H., Shi, M., Weinberg, C.R., Umbach, D.M.: A hybrid design: case-parent triads supplemented by control-mother dyads. *Genet. Epidemiol.* **33**, 136–144 (2009)
49. Wang, X., Sun, Q., McGrath, S.D., Mardis, E.R., Soloway, P.D., Clark, A.G.: Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* **3**, e3839 (2008)
50. Weinberg, C.R.: Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* **65**, 229–235 (1999)
51. Weinberg, C.R., Wilcox, A.J., Lie, R.T.: A log-linear approach to case-parent triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subjected to parental imprinting. *Am. J. Hum. Genet.* **62**, 969–978 (1998)
52. Weinberg, C.R., Umbach, D.M.: A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am. J. Hum. Genet.* **77**, 627–636 (2005)
53. Wilcox, A.J., Weinberg, C.R., Lie, R.T.: Distinguishing the effects of maternal and offspring genes through studies of case-parent triads. *Am. J. Epidemiol.* **148**, 893–901 (1998)
54. Wittkopp, P.J., Haerum, B.K., Clark, A.G.: Parent-of-origin effects on mRNA expression in *Drosophila melanogaster* not caused by genomic imprinting. *Genetics* **173**, 1817–1821 (2006)
55. Yang, J.: Likelihood approaches for detecting imprinting and maternal effects in family-based association studies. Ph.D. dissertation, The Ohio State University (2010)
56. Yang, J., Lin, S.: Detection of imprinting and heterogeneous maternal effects on high blood pressure using Framingham Heart Study data. *BMC Proc.* **3**, S125 (2009)
57. Yang, J., Lin, S.: Likelihood approach for detecting imprinting and maternal effect using general pedigrees from prospective family-based association studies. *Biometrics* **68**, 477–485 (2012)
58. Yang, J., Lin, S.: Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *Ann. Appl. Stat.* **7**, 249–268 (2013)
59. Yu, Z., Schaid, D.J.: Sequential haplotype scan methods for association analysis. *Genet. Epidemiol.* **31**, 553–564 (2007)
60. Zhou, J., Hu, Y., Lin, S., Fung, W.K.: Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children. *Hum. Hered.* **67**, 1–12 (2009)
61. Zhou, J., Lin, S., Fung, W.K., Hu, Y.-Q.: Detection of parent-of-origin effects in complete and incomplete nuclear families with multiple affected children using multiple tightly linked markers. *Hum. Hered.* **67**, 116–127 (2009)
62. Zhou, J., Ding, J., Fung, W.K., Lin, S.: Detection of parent-of-origin effects using general pedigree data. *Genet. Epidemiol.* **34**, 151–158 (2010)

Competing Risks Models and Breast Cancer: A Brief Review

Sharareh Taghipour, Dragan Banjevic, Anthony Miller, and Bart Harvey

Abstract In this paper, we present a brief overview of the methods which are commonly used for statistical analysis of data in competing risks settings. Moreover, we review 37 recent published clinical papers on breast cancer which consider an event of interest, such as breast cancer incident, while they also take into account the competing events, such as death due to other causes. The papers are selected based on the number of citations and publication year.

Introduction

In breast cancer follow-up studies, the risk of developing breast cancer or breast cancer mortality may be affected by deaths due to other causes. Ignoring competing events can change the probability of the event of interest and overestimate it. Appropriate statistical methods such as cause-specific hazard model or hazard of subdistribution should be used for proper regression modeling of an event in the presence of competing risks. In this paper, we first present a brief overview of the

S. Taghipour (✉)

Reliability, Risk and Maintenance Research Laboratory (RRMR),
Department of Mechanical and Industrial Engineering,
Ryerson University, Toronto, ON M5B 2K3, Canada
e-mail: sharareh@ryerson.ca

D. Banjevic

Centre for Maintenance Optimization and Reliability Engineering (C-MORE), Department of
Mechanical & Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, Canada
e-mail: banjev@mie.utoronto.ca

A. Miller • B. Harvey

Dalla Lana School of Public Health, Health Science Building, 155 College Street,
Toronto, ON M5T 3M7, Canada
e-mail: ab.miller@sympatico.ca; bart.harvey@utoronto.ca

statistical methods which are commonly used to analyze the data in competing risks settings. We then review the literature for recent clinical papers on breast cancer which have been interested in a particular event, such as breast cancer incident or death from breast cancer following a breast cancer diagnosis, while they have also considered the competing events.

Competing Risks Models

Cause-Specific Hazard and Cumulative Incidence Function

When there are K competing causes of death, the cause-specific hazard of survival time T , which is defined as the instantaneous probability of failing from cause k in presence of covariates Z is

$$\lambda_k(t|Z) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t < T \leq t + \Delta t, C = k | T \geq t, Z). \tag{1}$$

The overall hazard is then $\lambda(t|Z) = \sum_{k=1}^K \lambda_k(t|Z)$. In this model, a subject who fails due to other causes than k will no longer be at risk of failing from cause k .

Let us define $S_k^*(t|Z) = e^{-\Lambda_k(t|Z)}$, where $\Lambda_k(t|Z) = \int_0^t \lambda_k(s|Z) ds$. The

overall survival function is then equal to $S(t|Z) = e^{-\sum_{k=1}^K \Lambda_k(t|Z)} = \prod_{k=1}^K S_k^*(t|Z)$.

It should be noted that in this context, $S_k^*(t|Z)$ cannot be interpreted as the marginal probability that a person with covariates Z and subject only to risk from cause k will survive to time t . In the context of independent latent failures [1], $S_k^*(t|Z)$ is interpreted as the marginal survival function of cause k , and $1 - S_k^*(t|Z)$ is the cumulative “pure” risk. The cumulative incidence function of cause k is the probability of failing from cause k by time t and is equal to

$$I_k(t|Z) = P(T \leq t, C = k) = \int_0^t \lambda_k(s|Z) S(s|Z) ds. \tag{2}$$

This quantity is sometimes called the “absolute risk” or “crude” risk.

Methods to Estimate Cause-Specific Hazard

A Kaplan-Meier estimate for cause-specific hazard is given by Kalbfleisch and Prentice [2]. Cause-specific hazard can be also estimated using semi-parametric models such as Cox proportional hazards model (PHM) which models the effect of covariates. The cause-specific hazard of cause k for a subject with covariate Z is

$$\lambda_k(t|Z) = \lambda_{k,0}(t) \exp(\beta_k^T Z), \tag{3}$$

where $\lambda_{k,0}(t)$ is the baseline cause-specific hazard for cause k , and vector β_k is the covariate effect on cause k [3]. The cumulative incidence function of cause k is then obtained from Eq. 2.

A “naïve” (biased) formula for the cumulative incidence function of cause k is obtained if we replace $S(s|Z)$ in Eq. 2 with $S_k^*(s|Z)$:

$$\tilde{I}_k(t|Z) = \int_0^t \lambda_k(s|Z) S_k^*(s|Z) ds. \tag{4}$$

Hazard of Subdistribution and Cumulative Incidence Function

Fine and Gray [4] propose to use Cox PHM as a model for so-called hazard of subdistribution, to perform regression directly on cumulative incidence function. This model keeps in the risk sets the subjects who fail from the competing causes, but uses an estimate of the survivor function of the censoring distribution to reweight their contributions to the risk sets. The hazard of subdistribution is defined as

$$\bar{\lambda}_k(t|Z) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t < T \leq t + \Delta t, C = k | T \geq t \text{ or } C \neq k, Z). \tag{5}$$

The hazard of subdistribution can be modeled using Cox proportional hazards model as

$$\bar{\lambda}_k(t|Z) = \bar{\lambda}_{k,0}(t) \exp(\beta_k^T Z), \tag{6}$$

where $\bar{\lambda}_{k,0}(t)$ is the baseline subdistribution hazard of cause k .

The cumulative incidence function of cause k is then calculated by

$$I_k(t|Z) = 1 - e^{-\int_0^t \bar{\lambda}_k(s|Z) ds} = 1 - \exp\left(-\exp(\beta_k^T Z) \int_0^t \bar{\lambda}_{k,0}(s) ds\right). \tag{7}$$

Equation 7 unlike Eq. 2 uses only the hazard of subdistribution to obtain the cumulative incidence function of cause k , so when using Eq. 6 it is easier to interpret the effect of the vector of covariates Z on the cumulative incidence function.

Conditional Probability Function

Pepe and Mori [5] describe the conditional probability function as the probability of failure due to cause k by time t given that the subject has not failed from any other causes by this time. The conditional probability function is the proportion of the events of interest after removing observations that have experienced a competing event up to time t .

$$CP_k(t) = \frac{P(\text{failure from cause } k \text{ by time } t)}{1 - P(\text{non-failure from cause } k \text{ by time } t)} = \frac{I_k(t|Z)}{1 - I_{other}(t|Z)}, \quad (8)$$

where $I_{other}(t|Z)$ is the cumulative incidence function for all causes other than k . Pepe and Mori [5] warn against using Eq. 4, as it requires the optimistic assumption that incidence of an event completely eliminates occurrence of other competing events.

Type of Data Required for Competing Risks Models

Cause-specific hazard model or conditional probability function can be used for competing risks analysis of data from different types of clinical studies, such as full cohort, case-control, nested case-control, and case-cohort. However, estimation of cumulative incidence based on subdistribution hazard model requires full cohort data.

Competing Events in Breast Cancer Publications

In many breast cancer clinical studies, follow-up of a woman is terminated as soon as she experiences an event, either the event of interest or a competing event. We reviewed the literature for more recent clinical journal publications on breast cancer, which have considered the competing risk settings of their studies and employed one or more of the statistical methods described in section “[Competing Risks Models](#)” to address it. The papers are selected based on two criteria: number of citations and

publication year. 37 papers published after 2002 were reviewed. We classified the papers based on the main event of interest which they considered, or for which they developed a prediction model. The events include “breast cancer incidence”, “breast cancer mortality”, “breast cancer recurrence”, and “occurrence of a second cancer, other diseases, or metastasis following breast cancer”, shown in Tables 1, 2, 3, and 4. In each table, we present the main event of interest, the competing events, and the competing risks method which was used for analysis. Some of the papers have employed more than one method. In the following subsections, a summary of the reviewed papers in each category are presented.

Articles with Breast Cancer Incidence as the Outcome

This subsection provides a summary of the articles listed in Table 1, which consider breast cancer incidence as the outcome.

Travis et al. [6] use population breast cancer incidence rates and competing causes of death to estimate cumulative absolute breast cancer risk for women treated for Hodgkin lymphoma at age 30 years or younger. Degnim et al. [7] stratified the risk of breast cancer in women with atypia. Hwang et al. [8] estimate ductal carcinoma in situ prevalence and incidence in known BRCA-positive and BRCA-negative women who were undergoing genetic testing for a BRCA mutation. Kerlikowske et al. [9] identify characteristics of women who have been diagnosed with ductal carcinoma in situ (DCIS) and have a high or low risk of subsequent invasive cancer. These women were older than 40 years at diagnosis and treated by lumpectomy alone. Kurian et al. [10] estimate subtype-specific lifetime breast cancer risks.

Yi et al. [11] identify the factors that may affect a decision to undergo contralateral prophylactic mastectomy for patients with unilateral breast cancer who underwent breast-conserving surgery and/or mastectomy. Biggar et al. [12] examine the association of using digoxin with risk of breast cancer. Luo et al. [13] investigate the effect of smoking on the risk of developing invasive breast cancer among postmenopausal women aged 50–79 years.

Petracci et al. [14] estimate the effects of changes in modifiable risk factors on the absolute risk of breast cancer for women aged 23–74 years with breast cancer. Warner et al. [15] compare the incidence of advanced-stage breast cancers in women undergoing MRI screening with those undergoing conventional screening for women with a BRCA1 or BRCA2 mutation. Taghipour et al. [16] consider 39 risk factors collected at the time of enrolment in the Canadian National Breast Screening Study (CNBSS), and predict age-specific cumulative risk of invasive breast cancer using data from the CNBSS.

Table 1 Articles with breast cancer incidence as the outcome

Author ^a	Main event of interest	Competing events	Hazard model
Travis [6]	Breast cancer incidence	Death from other causes	CS
Degnim [7]	Breast cancer incidence	Death from other causes	CS
Hwang [8]	Incidence of ductal carcinoma in situ and invasive breast cancer	Mastectomy; salpingo-oophorectomy; tamoxifen use; death	CS
Kerlikowske [9]	Subsequent invasive cancer	Ductal carcinoma in situ and death from causes other than breast cancer	S
Kurian [10]	Age-specific incidence of breast cancer subtypes defined by estrogen receptor, progesterone receptor, and HER2/neu status	All non-breast cancer causes of death	CS
Yi [11]	Contralateral breast cancer incidence	Recurrence from the primary breast cancer	CS
Biggar [12]	Incidence of ER (estrogen receptor)-positive and ER-negative breast cancers	ER-negative and ER-unknown were competing events for ER-positive	CS
Luo [13]	Breast cancer incidence	Non-breast cancer mortality	CS
Petracci [14]	Breast cancer incidence	Deaths from other causes	CS
Warner [15]	Incidence of noninvasive, invasive, early-stage, and late-stage breast cancers	Mutually exclusive events	CS, CP
Taghipour [16]	Invasive breast cancer incidence	Non-breast cancer mortality	S

CS cause-specific hazard model, S subdistribution hazard model, CP conditional probability function

^aAll papers have more than two authors and we represent a paper by its first author

Table 2 Articles with breast cancer mortality as the outcome

Author ^a	Main event of interest	Competing events	Hazard model
Schairer [17]	Breast cancer mortality	Death from other causes	CP
Dalton [18]	Breast cancer mortality	Death from other causes	CS
Hanrahan [19]	Breast cancer mortality	Death from other causes	CS, CP
Chapman [20]	Breast cancer mortality	Deaths from other malignancies and/or other causes	CS
Newcomb [21]	Breast cancer mortality	Death from other causes	CP
Kalinsky [22]	Breast cancer mortality	Death from other causes	CS, S
Komenaka [23]	Breast cancer mortality	Death from other causes	S
Vinh-Hung [24]	Breast cancer mortality	Deaths from other causes	CS

CS cause-specific hazard model, S subdistribution hazard model, CP conditional probability function

^aAll papers have more than two authors and we represent a paper by its first author

Articles with Breast Cancer Mortality as the Outcome

In this subsection, a summary of the articles listed in Table 2 are given, which consider breast cancer mortality as the main event of interest.

Schairer et al. [17] estimate probabilities of death from breast cancer and other causes for white and black women with breast cancer. Dalton et al. [18] study the importance of a range of socioeconomic factors and comorbid disorders on survival after breast cancer surgery for women with a primary invasive breast cancer who were less than 70 years of age at the time of diagnosis. Hanrahan et al. [19] investigate the impact of prognostic factors on breast cancer-specific and non-breast cancer related mortality for T1a,bN0M0 breast cancer cases.

Chapman et al. [20] examine factors associated with cause-specific death in disease free breast cancer patients after adjuvant tamoxifen treatment. Newcomb et al. [21] evaluate the influence of prediagnostic use of hormone therapy on breast cancer mortality for women over 50 with invasive breast cancer. Kalinsky et al. [22] investigate the association of PIK3CA mutation with breast cancer for women who underwent surgery for primary breast cancer.

Komenaka et al. [23] compare the breast cancer outcomes of underinsured African American and non-Hispanic white women. Vinh-Hung et al. [24] examine the relationship between age and lymph node ratio and determine their effects on breast cancer and overall mortality for women over 50 with a unilateral histologically confirmed T1-T2 node positive surgically treated primary breast carcinoma.

Articles with Breast Cancer Recurrence as the Outcome

This subsection describes the articles which consider breast cancer recurrence as the outcome (listed in Table 3).

Table 3 Articles with breast cancer recurrence as the outcome

Author ^a	Main event of interest	Competing events	Hazard model
Fisher [25]	Ipsilateral breast tumor recurrence	Other recurrences; contralateral breast cancer; other second primary cancer; death	CS
Dignam [26]	Tumor recurrence; second primary cancers; contralateral breast tumors; cause-specific mortality	Mutually exclusive events	CS
Nottage [27]	Ipsilateral breast tumour recurrence	Deaths from or with breast cancer	S
Nguyen [28]	Local recurrence	Isolated regional nodal recurrence; distant metastasis; contralateral breast cancer; second malignancy; death without recurrence; loss to follow-up	S
Schaapveld [29]	Incidence of metachronous contralateral breast cancer	Death or synchronous contralateral breast cancer	CS
Yerushalmi [30]	Incidence of contralateral tumors	Death before developing contralateral breast cancer	S
Mell [31]	Incidence of locoregional recurrence, distant recurrence, and competing mortality	Mutually exclusive events	S
Galimberti [32]	Incidence of axillary recurrences	Local events; locoregional recurrences; distant metastases; contralateral breast cancers; other primary cancers; deaths	CS, S
Nsouli-Maktabi [33]	Incidence of second primary breast cancer	Death and second primary endometria or ovarian cancers	CS, CP

CS cause-specific hazard model, S subdistribution hazard model, CP conditional probability function

^aAll papers have more than two authors and we represent a paper by its first author

Table 4 Articles with a second cancer, other diseases, or metastasis following breast cancer as the outcome

Author ^a	Main event of interest	Competing events	Hazard model
Crump [34]	Secondary acute leukemia	Recurrence and death from any causes	CP
Ryberg [35]	Central nervous system metastasis	Death from progressive breast cancer	CS
Pestalozzi [36]	Central nervous system metastases as the first site of recurrence	Other sites of first recurrence; contralateral breast cancer; non-breast cancer second primary tumors; death without recurrence	CS
Brown [37]	Incidence of second cancers	Breast cancer and non-breast cancer deaths	CS
Howard [38]	Incidence of leukemia after diagnosis of breast cancer	Death and occurrence of a second cancer other than leukemia	CS
Marees [39]	Second malignancies by type of retinoblastoma and treatment	Death from other causes	CS
Ryberg [40]	Cardiotoxicity	Death from all causes including breast cancer	CS
Schaapveld [41]	Secondary nonbreast cancers	Death and second breast cancers	CS
Kennecke [42]	Relapses to specific sites	Death from other causes	S

CS cause-specific hazard model, S subdistribution hazard model, CP conditional probability function

^aAll papers have more than two authors and we represent a paper by its first author

Fisher et al. [25] investigate the need for breast irradiation after lumpectomy for node-negative women with invasive breast cancers of less than one centimeter. The effects of obesity and race on prognosis in lymph node-negative, estrogen receptor-negative breast cancer women are investigated by Dignam et al. [26]. Nottage et al. [27] establish the incidence of ipsilateral breast tumour recurrence in a community treatment setting for women with node negative breast cancer who diagnosed between ages 18–75 years.

Nguyen et al. [28] determine whether breast cancer subtype is associated with outcome after breast-conserving therapy for women with invasive breast cancer. Schaapveld et al. [29] also investigate the impact of age and adjuvant therapy on contralateral breast cancer for surgically treated stage I–IIIA patients. Yerushalmi et al. [30] compare the incidence of contralateral breast cancer between the multifocal/multicentric and unifocal groups for women diagnosed with stage I–III breast cancer.

Mell et al. [31] identify predictors of competing mortality in women with stage I to II invasive breast cancer. Galimberti et al. [32] analyze the outcomes in single micrometastatic sentinel node patients who did not receive axillary dissection. Nsouli-Maktabi et al. [33] describe the cumulative incidence function of second primary breast in first primary breast cancer black and white female survivors.

Articles with a Second Cancer, Other Diseases, or Metastasis Following Breast Cancer as the Outcome

In this subsection, a summary of the articles listed in Table 4 are given, whose main event of interest is a second cancer, other diseases, or metastasis following breast cancer.

Crump et al. [34] estimate the risk of secondary acute leukemia following epirubicin-containing chemotherapy regimens for women who received adjuvant or neoadjuvant chemotherapy. Ryberg et al. [35] consider patients treated with epirubicin-based chemotherapy and identify predictive factors for central nervous system metastasis. Pestalozzi et al. [36] determine whether a high-risk group could be defined among patients with operable breast cancer in whom a search of occult central nervous system metastases was justified. They consider women with early breast cancer as their study population.

Brown et al. [37] consider women diagnosed with breast cancer as a first primary cancer who survived at least one year, and examine the absolute risk of second cancer risk thirty or more years after diagnosis. Howard et al. [38] quantify long-term temporal trends in the excess absolute risk of secondary leukemia among women diagnosed with a first primary breast cancer who survived one or more years.

Marees et al. [39] estimate the risk of second malignancies in the survivors of retinoblastoma. Ryberg et al. [40] identify the risk factors for cardiotoxicity and overall mortality for anthracycline-naive patients treated for metastatic breast cancer

with epirubicin. The risk of secondary non-breast cancers is assessed by Schaapveld et al. [41] for patients diagnosed with invasive breast cancer. Kennecke et al. [42] investigate metastatic behavior of breast cancer subtypes for patients with early-stage of breast cancer.

Discussion

In this paper we briefly reviewed several recent clinical publications on breast cancer which considered two or more competing events in their studies. We were particularly interested to identify in each paper the main event of interest and the other competing events and investigate the statistical method which was used. The main event of interest in the majority of papers is either breast cancer incidence or mortality, and death due to causes other than breast cancer is the competing event. Some studies were interested in more than one event, such as incidence of both ductal carcinoma in situ and invasive breast cancer. In some studies, more than one event was competing with the main event, for example death and second breast cancers were the competing events with secondary non-breast cancer as the main event of interest. Most papers used cause-specific hazard model as an approach for analyzing their competing risks data. More recent papers have opted for the hazard of subdistribution as the statistical method for addressing competing events, probably due to availability of more recently developed statistical tools, such as package *cmprsk* in R [43].

There are other review articles on competing risk models and their uses for breast cancer. Gail [44, 45] defines absolute and pure risks, and describes some applications of absolute risk in breast cancer counseling and prevention.

Competing risks are relevant for medical research and their ignorance has significant clinical consequences [46]. A review of clinical studies performed by Koller et al. [46] reveals competing risks issues in 70 % of articles. Statistical methods developed for competing risks data should be used to properly model and estimate an event in the presence of other competing events.

Acknowledgements We acknowledge the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research for their financial support.

References

1. Pintilie, M.: *Competing Risks: A Practical Perspective*. Wiley, New York (2006)
2. Kalbfleisch, J.D., Prentice, R.L.: *The Statistical Analysis of Failure Time Data*, 1st edn. Wiley, New York (1980)
3. Putter, H., Ficco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* **26**(11), 2389–2430 (2007)
4. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**(446), 496–509 (1999)

5. Pepe, M.S., Mori, M.: Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat. Med.* **12**(8), 737–751 (1993)
6. Travis, L.B., Hill, D., Dores, G.M., et al.: Cumulative absolute breast cancer risk for young women treated for Hodgkin lymphoma. *J. Natl. Cancer Inst.* **97**, 1428–1437 (2005)
7. Degnim, A.C., Visscher, D.W., Berman, H.K., et al.: Stratification of breast cancer risk in women with atypia: a Mayo cohort study. *J. Clin. Oncol.* **25**, 2671–2677 (2007)
8. Hwang, E.S., McLennan, J.L., Moore, D.H., et al.: Ductal carcinoma in situ in BRCA mutation carriers. *J. Clin. Oncol.* **25**(6), 642–647 (2007)
9. Kerlikowske, K., Molinaro, A.M., Gauthier, M.L., et al.: Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis. *J. Natl. Cancer Inst.* **102**, 627–637 (2010)
10. Kurian, A.W., Fish, K., Shema, S., et al.: Lifetime risks of specific breast cancer subtypes among women in four racial/ethnic groups. *Breast Cancer Res.* **12**(6), R99 (2010)
11. Yi, M., Hunt, K.K., Arun, B.K., et al.: Factors affecting the decision of breast cancer patients to undergo contralateral prophylactic mastectomy. *Cancer Prev. Res.* **3**(8), 1026–1034 (2010)
12. Biggar, R.J., Wohlfahrt, J., Oudin, A., et al.: Digoxin use and the risk of breast cancer in women. *J. Clin. Oncol.* **29**(16), 2165–2170 (2011)
13. Luo, J., Horn, K., Ockene, J.K., et al.: Interaction between smoking and obesity and the risk of developing breast cancer among postmenopausal women: the Women’s Health Initiative Observational Study. *Am. J. Epidemiol.* **174**(8), 919–928 (2011)
14. Petracci, E., Decarli, A., Schairer, C., et al.: Risk factor modification and projections of absolute breast cancer risk. *J. Natl. Cancer Inst.* **103**(13), 1037–1048 (2011)
15. Warner, E., Hill, K., Causer, P., et al.: Prospective study of breast cancer incidence in women with a BRCA1 or BRCA2 mutation under surveillance with and without magnetic resonance imaging. *J. Clin. Oncol.* **29**, 1664–1669 (2011)
16. Taghipour, S., Banjevic, D., Fernandes, J., et al.: Incidence of invasive breast cancer in the presence of competing mortality: The Canadian National Breast Screening Study. *Breast Cancer Res. Treat.* **134**(2), 839–851 (2012)
17. Schairer, C., Mink, P.J., Carroll, L., et al.: Probabilities of death from breast cancer and other causes among female breast cancer patients. *J. Natl. Cancer Inst.* **96**(17), 1311–1321 (2004)
18. Dalton, S.O., Ross, L., Doring, M., et al.: Influence of socioeconomic factors on survival after breast cancer – a nationwide cohort study of women diagnosed with breast cancer in Denmark 1983–1999. *Int. J. Cancer* **121**, 2524–2531 (2007)
19. Hanrahan, E.O., Gonzalez-Angulo, A.M., Giordano, S.H., et al.: Overall survival and cause-specific mortality of patients with stage T1a, bN0M0 breast carcinoma. *J. Clin. Oncol.* **25**, 4952–4960 (2007)
20. Chapman, J.A., Meng, D., Shepherd, L., et al.: Competing causes of death from a randomized trial of extended adjuvant endocrine therapy for breast cancer. *J. Natl. Cancer Inst.* **100**(4), 252–260 (2008)
21. Newcomb, P.A., Egan, K.M., Trentham-Dietz, A., et al.: Prediagnostic use of hormone therapy and mortality after breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **7**(4), 864–871 (2008)
22. Kalinsky, K., Jacks, L.M., Heguy, A., et al.: PIK3CA mutation associates with improved outcome in breast cancer. *Clin. Cancer Res.* **15**(16), 5049–5059 (2009)
23. Komenaka, I.K., Martinez, M.E., Pennington Jr., R.E., et al.: Race and ethnicity and breast cancer outcomes in an underinsured population. *J. Natl. Cancer Inst.* **102**(15), 1178–1187 (2010)
24. Vinh-Hung, V., Joseph, S.A., Coutty, N., et al.: Age and axillary lymph node ratio in postmenopausal women with T1-T2 node positive breast cancer. *Oncologist* **15**, 1050–1062 (2010)
25. Fisher, B., Bryant, J., Dignam, J.J., et al.: Tamoxifen, radiation therapy, or both for prevention of ipsilateral breast tumor recurrence after lumpectomy in women with invasive breast cancers of one centimeter or less. *J. Clin. Oncol.* **20**(20), 4141–4149 (2002)
26. Dignam, J.J., Wieand, K., Johnson, K.A., et al.: Effects of obesity and race on prognosis in lymph node-negative, estrogen receptor-negative breast cancer. *Breast Cancer Res. Treat.* **97**, 245–254 (2006)

27. Nottage, M.K., Kopciuk, K.A., Tzontcheva, A., et al.: Analysis of incidence and prognostic factors for ipsilateral breast tumour recurrence and its impact on disease-specific survival of women with nodenegative breast cancer: a prospective cohort study. *Breast Cancer Res.* **8**, R44 (2006)
28. Nguyen, P.L., Taghian, A.G., Katz, M.S., et al.: Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. *J. Clin. Oncol.* **26**(14), 2373–2378 (2008)
29. Schaapveld, M., Visser, O., Louwman, W.J., et al.: The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: a population based study in the Netherlands. *Breast Cancer Res. Treat.* **110**, 189–197 (2008)
30. Yerushalmi, R., Kennecke, H., Woods, R., Olivotto, I.A., Speers, C., Gelmon, K.A.: Does multicentric/multifocal breast cancer differ from unifocal breast cancer? An analysis of survival and contralateral breast cancer incidence. *Breast Cancer Res. Treat.* **117**(2), 365–370 (2009)
31. Mell, L.K., Jeong, J., Nichols, M.A., et al.: Predictors of competing mortality in early breast cancer. *Cancer* **116**(23), 5365–5373 (2010)
32. Galimberti, V., Botteri, E., Chifu, C., et al.: Can we avoid axillary dissection in the micrometastatic sentinel node in breast cancer? *Breast Cancer Res. Treat.* **131**, 819–825 (2012)
33. Nsouli-Maktabi, H., Henson, D., Younes, N., Young, H., Cleary, S.: Second primary breast, endometrial, and ovarian cancers in Black and White breast cancer survivors over a 35-year time span: effect of age. *Breast Cancer Res. Treat.* **129**(3), 963–969 (2011)
34. Crump, M., Tu, D., Sheperd, L., et al.: Risk of acute leukemia following epirubicin-based adjuvant chemotherapy: a report from the National Cancer Institute of Canada Clinical Trials Group. *J. Clin. Oncol.* **21**(16), 3066–3071 (2003)
35. Ryberg, M., Nielsen, D., Osterlind, K., et al.: Predictors of central nervous system metastasis in patients with metastatic breast cancer. A competing risk analysis of 579 patients treated with epirubicin-based chemotherapy. *Breast Cancer Res. Treat.* **91**, 217–225 (2005)
36. Pestalozzi, B.C., Zahrieh, D., Price, K.N., et al.: Identifying breast cancer patients at risk for central nervous system (CNS) metastases in trials of the International Breast Cancer Study Group (IBCSG). *Ann. Oncol.* **17**, 935–944 (2006)
37. Brown, L.M., Chen, B.E., Pfeiffer, R.M., et al.: Risk of second non-hematological malignancies among 376,825 breast cancer survivors. *Breast Cancer Res. Treat.* **106**, 439–451 (2007)
38. Howard, R.A., Gilbert, E.S., Chen, B.E., et al.: Leukemia following breast cancer: an international population-based study of 376,825 women. *Breast Cancer Res. Treat.* **105**, 359–368 (2007)
39. Marees, T., Moll, A.C., Imhof, S.M., et al.: Risk of second malignancies in survivors of retinoblastoma: more than 40 years of follow-up. *J. Natl. Cancer Inst.* **100**(24), 1771–1779 (2008)
40. Ryberg, M., Nielsen, D., Cortese, G., et al.: New insight into epirubicin cardiac toxicity: competing risks analysis of 1097 breast cancer patients. *J. Natl. Cancer Inst.* **100**(15), 1058–1067 (2008)
41. Schaapveld, M., Visser, O., Louwman, M.J., et al.: Risk of new primary nonbreast cancers after breast cancer treatment: a Dutch population-based study. *J. Clin. Oncol.* **26**(8), 1239–1246 (2008)
42. Kennecke, H., Yerushalmi, R., Woods, R., et al.: Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol.* **28**(20), 3271–3277 (2010)
43. Scrucca, L., Santucci, A., Aversa, F.: Regression modeling of competing risk using R: an in depth guide for clinicians. *Bone Marrow Transplant.* **45**, 1388–1395 (2010)
44. Gail, M.H.: Estimation and interpretation of models of absolute risk from epidemiologic data, including family-based studies. *Lifetime Data Anal.* **14**, 18–36 (2008)
45. Gail, M.H.: Personalized estimates of breast cancer risk in clinical practice and public health. *Stat. Med.* **30**(10), 1090–1104 (2011)
46. Koller, M.T., Raatz, H., Steyerberg, E.W., et al.: Competing risks in the clinical community: irrelevance or ignorance? *Stat. Med.* **31**, 1089–1097 (2012)

Quantifying Relative Potency in Dose-Response Studies

Gregg E. Dinse and David M. Umbach

Abstract Relative potency is an important concept in the comparative evaluation of chemicals via dose-response studies. For example, toxicologists use relative potency estimates to rank chemicals with respect to a given response endpoint, to convert doses of one chemical to equivalent doses of another chemical, and to combine information across studies and endpoints when calculating toxic equivalency factors. The conventional definition of relative potency, arising historically from dilution assays, is a ratio of equi-effective doses, that is, those doses that produce the same mean response. Specifically, the ratio is the dose of a reference chemical divided by the dose of a test chemical. In an analytical dilution assay, relative potency is constant regardless of the mean response used to select equi-effective doses. Nevertheless, researchers often observed data that were inconsistent with constant relative potency and desired ways to characterize non-constant relative potency. This article reviews various approaches for quantifying relative potency when it cannot be regarded as constant, including modifications to the usual definition. In particular, we focus on recent proposals that describe the relative potency of two chemicals as functions of dose or of response.

Introduction

Relative potency plays a critical role in toxicology. For example, toxicologists estimate relative potency to rank chemicals with respect to a toxicity endpoint of interest (e.g., [1]), to convert a dose of one chemical to an equivalent dose of another chemical (e.g., [2]), and to combine information across studies and endpoints when calculating a chemical's toxic equivalency factor (e.g., [3]). Relative potency is

G.E. Dinse (✉) • D.M. Umbach
Biostatistics Branch, National Institute of Environmental Health Sciences, MD A3-03,
P.O. Box 12233, Research Triangle Park, NC 27709, USA
e-mail: dinse@niehs.nih.gov; umbach@niehs.nih.gov

typically derived from the parameters in a mathematical (dose-response) model that expresses a toxicity response as a function of a chemical dose.

Consider a dose-response function that relates the mean response for a particular endpoint to the dose of a given chemical. Let $f(d; \theta)$ be a model that specifies mean response in terms of dose d and parameter vector θ . We focus on models for which f is a monotone increasing function of d , though the same methods can be modified easily to handle monotone decreasing dose-response functions. Early methods for comparative bioassays often assumed a linear model for f , possibly after transforming dose, response, or both. Often, linearity is reasonable over some restricted dose-response region only. A linear dose-response model specifies $f(d; \theta) = \alpha + \beta d^*$, where $\theta = (\alpha, \beta)$, α is an intercept, β is a slope, and d^* is a dose metric (typically either dose itself or log dose). Other assays, especially those for binary endpoints, frequently employed a sigmoid model with lower and upper response asymptotes and expressed generally as:

$$f(d; \theta) = L + (U - L)g(d; \phi), \quad (1)$$

where L is the lower response limit, U is the upper response limit, and the dose-quantile function g is a monotone increasing function of d that ranges from 0 (at $d=0$) to 1 (at $d=\infty$) and depends on a parameter vector ϕ , with $\theta = (L, U, \phi)$. If mean response decreases as dose increases, we associate U with $d=0$ and L with $d=\infty$ and require g to be monotone decreasing in dose. In either case, the elements of ϕ typically govern the location and shape of the dose-response curve.

Now consider multiple chemicals. Without loss of generality, we focus on two chemicals: a reference chemical, C_0 , and a test chemical, C_1 . Rooted in ideas from dilution assays, relative potency, denoted ρ , is classically defined as the ratio of equi-effective doses (reference divided by test), i.e., doses of the two chemicals that elicit the same response. Ideally, in dilution assays, this ratio does not change with the response level chosen. Faced with examples where the ratio did vary with response level, investigators had to grapple with ways to characterize non-constant relative potency.

This article reviews approaches that have been proposed for assessing non-constant relative potency. Some of these retain the classical definition of relative potency as a ratio of equi-effective doses but abandon the notion that a single numerical constant suffices to compare potency of two chemicals. Others retain the simplicity of a single constant to compare potency between chemicals but abandon or modify the classical definition. The most recent developments describe non-constant relative potency using the notation of mathematical functions.

Constant Relative Potency in Bioassay

The classical concept of relative potency arises from analytical dilution assays, where each test preparation is constructed as a dilution of a reference preparation [4]. In this context, relative potency as the ratio of equi-effective doses is a

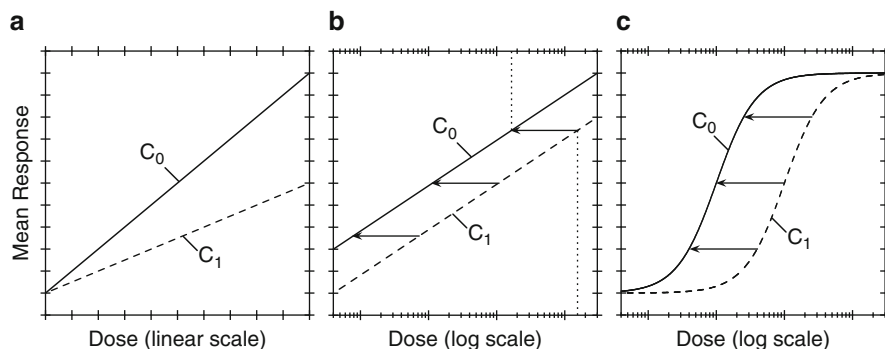


Fig. 1 Dose-response curves producing constant relative potency. Panels: **(a)** diverging lines with equal intercepts, where mean response is linear in dose; **(b)** parallel lines with equal slopes, where mean response is linear in log dose; and **(c)** similar sigmoid curves, generated by Hill functions with equal response limits and shapes. In all three panels, the ratio of any equi-effective doses for reference chemical C_0 and test chemical C_1 is constant and equals the relative potency. In panels **(b)** and **(c)**, the length of each horizontal arrow from C_1 to C_0 is constant and equals the log relative potency. In panel **(b)**, the vertical dotted lines illustrate that a given arrow (or relative potency) can be indexed by the dose of either chemical, as well as by mean response

constant, ρ , regardless of the response level considered. When relative potency is constant, ranking chemicals is straightforward: simply rank them by the relative potencies. Dose conversion is also simple: the dose of chemical C_0 that is equivalent to dose d_1 of chemical C_1 is $d_0 = d_1\rho$, and the dose of C_1 that is equivalent to dose d_0 of C_0 is $d_1 = d_0/\rho$. Furthermore, because the ratio of equi-effective doses is constant, the difference between the logs of those doses is also constant. Thus, as often noted, relative potency is constant if and only if the dose-response functions are identical except for a horizontal shift when plotted against log dose (though this graphical definition can be inconvenient when zero doses are involved). When the relative potency of two chemicals is constant, their dose-response curves are referred to as similar.

Slope Ratio Assays

A slope ratio assay is based on dose-response curves that are linear functions of dose with a common intercept (usually the origin) but possibly distinct slopes [4]. Thus, the dose-response function for C_i is $f(d;\theta_i) = \alpha + \beta_i d$, where $\beta_i > 0$ and $\theta_i = (\alpha, \beta_i)$ for $i = 0, 1$ (Fig. 1a). Denoting the dose of C_i that produces mean response μ by $d_i(\mu)$, the corresponding inverse function for C_i is $d_i(\mu) = f^{-1}(\mu; \theta_i) = (\mu - \alpha)/\beta_i$ and relative potency is a constant ratio of the slopes: $d_0(\mu)/d_1(\mu) = \beta_1/\beta_0$ for all values of μ .

Parallel Line Assays

A parallel line assay is based on dose-response curves that are linear functions of log dose with a common slope but possibly distinct intercepts [4]. Thus, the dose-response function for C_i is $f(d; \theta_i) = \alpha_i + \beta \log(d)$, where $\beta > 0$ and $\theta_i = (\alpha_i, \beta)$ for $i=0,1$ (Fig. 1b). The corresponding inverse function is $d_i(\mu) = f^{-1}(\mu; \theta_i) = \exp[(\mu - \alpha_i)/\beta]$ and relative potency is again constant: $d_0(\mu)/d_1(\mu) = \exp[(\alpha_1 - \alpha_0)/\beta]$ for all values of μ .

Assays Involving Similar Sigmoid Curves

Consider chemicals that have sigmoid dose-response functions of the form given in Eq. 1. Suppose L and U are the same for both chemicals and that the vector ϕ is the same for both chemicals up to a location parameter for log dose. Then, the dose-response curves are similar, and the chemicals have constant relative potency. The Hill [5] model is an example. It is obtained by setting $g(d; \phi) = d^S / (d^S + M^S)$, where S is a shape parameter and M is the median effective dose (ED_{50}), which is the dose producing a mean response halfway between L and U . Similar Hill curves have identical response limits and shapes; only their ED_{50} s differ (Fig. 1c). The corresponding inverse function for C_i is $d_i(\mu) = f^{-1}(\mu; \theta_i) = M_i [(\mu - L)/(U - \mu)]^{1/S}$ and relative potency is a constant equal to the ED_{50} ratio: $d_0(\mu)/d_1(\mu) = M_0/M_1$ for all values of μ between L and U . The Hill model can be rewritten in its log logistic form by setting $g(d; \phi) = 1/[1 + \exp(-X)]$ with $X = S[\log(d) - \log(M)]$. Here, $g(d; \phi)$ is a logistic distribution function for $\log(d)$ with location parameter $\log(M)$ and scale parameter $1/S$ [6]. Analogously, the probit model takes the dose-quantile function $g(d; \phi)$ as the standard normal distribution function evaluated at X [6]. Other distribution functions, such as the Weibull [7], can be used for $g(d; \phi)$, and ϕ can contain more than two parameters [8]. In any of these cases, similar sigmoid curves (and thus constant relative potencies) are obtained by constraining the dose-response models for C_0 and C_1 to be identical except for the location parameter.

Non-constant Relative Potency

In many situations, the notion of constant relative potency is inconsistent with observed data, and investigators face a dilemma. One strategy is to retain the simplicity of a single constant as a descriptor of relative potency, even though treating relative potency as fixed when it is not can generate misleading conclusions [9]. This strategy can involve modifying or abandoning the classical definition of relative potency based on a ratio of equi-effective doses. An alternate strategy is to adopt a descriptor of relative potency that involves more than a single constant,

but this alternative has the undesirable side effect of making dose conversion or chemical ranking problematic. Despite an awareness that many pairs of chemicals have non-constant relative potency, few general approaches for handling non-constant relative potencies were developed until recently.

Defining Relative Potency as the Ratio of ED_{50} s

Because similar sigmoid dose-response functions have constant relative potency given by the ratio of their ED_{50} s, some authors have simply employed that ratio as a measure of relative potency even for data where dose-response curves in log dose may differ by more than a constant horizontal shift (e.g., [10]). Others have pointed out that this approach is simple and convenient but less than ideal theoretically [11]. The convenience arises because an estimate of the ED_{50} is usually output by software for fitting dose-response models. On the other hand, because this approach treats relative potency as constant despite evidence to the contrary, it can lead to flawed conclusions when ranking chemicals [9] and would certainly distort dose conversions.

A more subtle issue also arises. When two sigmoid curves have the same upper and lower response limits, the ED_{50} values for each curve correspond to the same value of mean response for both curves. In that case, the ratio of ED_{50} s meets the classical definition of relative potency, at least at the single chosen response level. On the other hand, when the two curves differ in their upper and/or lower response limits, the ED_{50} values for each curve typically correspond to distinct values of mean response for each curve and the classical definition of relative potency is lost. The doses are no longer equi-effective in the sense of having the same mean response; the doses instead mark the same proportional change in mean response between the respective lower and upper limits for each chemical.

Deforming the Log-Dose and Response Axes to Achieve Similarity via Splines

Guardabasso et al. [12] proposed to fit the reference chemical's dose-response curve using a cubic spline function of log dose and then obtain the test chemical's dose-response curve by horizontally shifting and stretching the reference chemical's spline by constant amounts along the log dose axis – essentially deforming the log dose axis with a two-parameter transformation. They assumed that both chemicals had the same response limits and equated log relative potency with the constant shift parameter, even if the stretch (i.e., scale) parameter differed from 1. Thus, even though they reported a constant value that they called 'relative potency', they invoked an unconventional definition by allowing the dose-response curves to differ

by more than a constant horizontal shift along the log dose axis. Later, Guardabasso et al. [13] extended this approach to accommodate chemicals with different response limits by also allowing vertical shifting and stretching of the reference spline along the response axis. Their methods retained the simplicity of characterizing relative potency by a single parameter at the expense of redefining relative potency in a way that no longer matched the classical definition. Although the construction is a clever one, the utility of this approach for the traditional uses of relative potency, such as chemical ranking or dose conversion, seems questionable.

Evaluating Relative Potency at Multiple $ED_{100\pi}$ Values

We have already mentioned the common approach of using the ratio of ED_{50} s to assess relative potency even if the dose-response curves are not similar. Of course, with non-constant relative potency, the ED_{50} ratio can differ greatly from the ED_{10} ratio, the ED_{75} ratio, or any other ratio of $ED_{100\pi}$ values (for any $0 < \pi < 1$). One slight improvement on estimating non-constant relative potency by a single $ED_{100\pi}$ ratio would be to report several ratios [14] or the range between two effective doses, such as the ED_{20} and the ED_{80} [15]. Insofar as these proposals rely on $ED_{100\pi}$ values, as mentioned earlier, they entail a modification of the classical definition of relative potency when the two chemicals differ in their lower and/or upper response limits.

Relative Potency Functions

From evaluating relative potency at a finite list of equi-effective dose levels, it is a short step to evaluating relative potency at every relevant dose level, that is, to defining a relative potency function.

Parallel Line Assays Where Similarity Fails

Cornfield [16] derived a relative potency function under separate linear log-dose-response models. Assume that the mean response to dose d_i of C_i is $f(d_i; \theta_i) = \alpha_i + \beta_i \log(d_i)$ and $\theta_i = (\alpha_i, \beta_i)$ for $i=0,1$ (for similarity, the slopes would be equal). The corresponding inverse function is $d_i(\mu) = f^{-1}(\mu; \theta_i) = \exp[(\mu - \alpha_i)/\beta_i]$, which allowed Cornfield to express log relative potency as a linear function of mean response μ :

$$\lambda_\mu(\mu) = \log[\rho_\mu(\mu)] = \log\left(\frac{d_0(\mu)}{d_1(\mu)}\right) = \left(\frac{\alpha_1}{\beta_1} - \frac{\alpha_0}{\beta_0}\right) + \left(\frac{1}{\beta_0} - \frac{1}{\beta_1}\right)\mu. \quad (2)$$

Here, the notation $\rho_{\mu}(\mu)$ denotes a relative potency function that maps μ to the relative potency at response level μ . Cornfield noted that relative potency also can be indexed by the dose of either chemical (Fig. 1b) and derived formulae for $\rho_{d1}(d_1)$ and $\rho_{d0}(d_0)$ that express relative potency as functions of the doses of the test and reference chemicals, respectively. All three relative potency functions reduce to the constant obtained under the parallel line model if $\beta_0 = \beta_1 = \beta$. Cornfield's approach, which assumes a separate linear model in log dose for each dose-response curve, produces relative potency functions that are log-linear either in mean response or in log dose. His approach would be effective whenever a suitable transformation of the response yields a pair of dose-response models that are linear in log dose.

Specifying a Relative Potency Function a Priori

DeVito et al. [17] addressed the problem of estimating relative potency when data on the reference chemical are adequate to fit a non-linear (i.e., Hill) dose-response model, but data on the test chemical are not. For example, when fitting a sigmoid, if responses at the highest tested doses do not level out, estimation of the upper response limit (and thus the ED_{50}) becomes problematic. DeVito et al. [17] proposed the following ad hoc solution: (i) fit a Hill model to the reference chemical data; (ii) invert this Hill model to express dose as a function of mean response; (iii) for each (dose-specific) sample mean response in the test group, apply the inverse model to predict an equivalent dose of the reference chemical (say \hat{d}_0); and (iv) fit a linear model for equivalent reference dose in terms of actual test dose (say d_1) to give: $\hat{d}_0 = \alpha + \beta d_1$. If the dose-response curves are similar, α is zero and the relative potency equals the constant β . However, if α is nonzero, relative potency is linear in the reciprocal of test dose, namely: $\rho_{d1}(d_1) = \hat{d}_0/d_1 = \beta + \alpha/d_1$. Later, facing data where the simple linear regression of \hat{d}_0 on d_1 seemed inadequate, DeVito et al. [18] extended their procedure to give a relative potency function that was constant up to a threshold and then linear in the reciprocal of test dose.

This approach differs in a fundamental way from Cornfield's approach. Cornfield specified two dose-response models and deduced the appropriate relative potency function. DeVito et al. specified a dose-response model for the reference chemical but not for the test chemical. Instead, by assuming a simple linear regression of \hat{d}_0 on d_1 , their procedure in effect specifies a relative potency function and uses that function together with the dose-response model for the reference chemical to implicitly induce a dose-response model for the test chemical. With such a procedure, the induced dose-response model for the test chemical may not have the same functional form as the dose-response model for the reference chemical.

Sigmoid Dose-Response Models

Ritz et al. [19] derived a general formula for relative potency as a function of mean response for dose-response model (1). If $f(d; \theta)$ is monotone, one can invert $\mu = f(d; \theta)$ to express dose as a function of mean response: $d = f^{-1}(\mu; \theta)$. Suppose $d_0(\mu)$ and $d_1(\mu)$ are doses of C_0 and C_1 that both produce the same mean response μ . Dividing $d_0(\mu)$ by $d_1(\mu)$ expresses relative potency as a function of mean response μ :

$$\rho_{\mu}(\mu) = f^{-1}(\mu; \theta_0) / f^{-1}(\mu; \theta_1), \quad (3)$$

where θ_0 and θ_1 are the parameter vectors in the dose-response models for chemicals C_0 and C_1 . If L_i and U_i are the lower and upper response limits for C_i ($i=0, 1$), $\rho_{\mu}(\mu)$ is positive and finite for any μ in the intersection of the response ranges: $\max(L_0, L_1) < \mu < \min(U_0, U_1)$. Conversely, $\rho_{\mu}(\mu)$ is undefined for any $\mu < \min(L_0, L_1)$ or $\mu > \max(U_0, U_1)$; and if μ lies between two distinct lower (or upper) response limits, $\rho_{\mu}(\mu)$ is either 0 or ∞ .

Dinse and Umbach [9] extended these ideas by expressing relative potency as functions of reference dose, of test dose, and of response quantile. Recall that similar sigmoid curves are identical up to a constant shift along the log dose axis (Fig. 1c). In fact, if we draw a horizontal arrow from the dose-response curve for C_1 to the dose-response curve for C_0 , the length and direction of the arrow correspond to the magnitude and sign of the log relative potency (with left being negative). For similar dose-response curves, any horizontal arrow will have the same length and direction (Fig. 1c). For non-similar curves, each length can be distinct and the direction may change. Nevertheless, each arrow, and thus each log relative potency (or relative potency), can be indexed by mean response, reference dose, and test dose (Fig. 2). Indexing by response quantile is somewhat different, and we will return to it later.

Consider expressing relative potency as a function of dose. Substituting $f(d_0; \theta_0)$ for μ in Eq. 3 and noting that $f^{-1}(f(d_0; \theta_0); \theta_0) = d_0$, one may express relative potency as a function of reference dose d_0 :

$$\rho_{d_0}(d_0) = d_0 / f^{-1}(f(d_0; \theta_0); \theta_1). \quad (4)$$

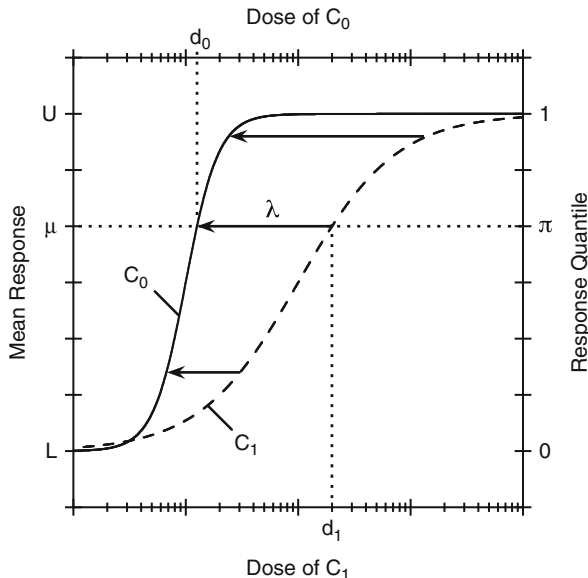
Substituting $f(d_1; \theta_1)$ instead, one may express relative potency as a function of test dose d_1 :

$$\rho_{d_1}(d_1) = f^{-1}(f(d_1; \theta_1); \theta_0) / d_1. \quad (5)$$

These relative potency functions are defined or undefined according to where the corresponding mean responses, $f(d_0; \theta_0)$ and $f(d_1; \theta_1)$, fall with respect to the bounds for $\rho_{\mu}(\mu)$.

Relative potency also can be indexed by response quantile (denoted by π), which is the fraction of the distance between the lower and upper response limits (i.e., mean response standardized to the unit interval). As mean response μ varies from L to U , the corresponding quantile $\pi = (\mu - L)/(U - L)$ varies from 0 to 1. Let $ED_{100\pi}$

Fig. 2 Classical definition of relative potency indexed by dose, response, and response quantile. Reference chemical C_0 and test chemical C_1 have the same lower (L) and upper (U) response limits. The length of each horizontal arrow, drawn from C_1 to C_0 , varies and represents a changing log relative potency. For illustration, the arrow labeled λ (and its corresponding relative potency) can be indexed by mean response (μ), reference chemical dose (d_0), test chemical dose (d_1), or response quantile (π), as indicated by the dotted lines



be the dose producing a mean response $100\pi\%$ of the way from L to U (e.g., $\pi = 0.5$ gives the ED_{50}). If C_0 and C_1 have the same upper and same lower response limits, each value of π corresponds to the same value of μ for both chemicals (Fig. 2). On the other hand, if the chemicals differ in one or both response limits, each value of π will correspond to a distinct value of μ for each chemical (Fig. 3).

Consider the ratio of $ED_{100\pi}$ values for C_0 and C_1 as an alternative definition of relative potency [9]. If C_0 and C_1 have the same response limits, the log of the $ED_{100\pi}$ ratio is the horizontal distance between their dose-response curves on a log dose axis (Fig. 2). Thus, when chemicals have equal response limits, a definition based on the $ED_{100\pi}$ ratio corresponds exactly to the classical concept of relative potency. If the limits differ, however, the $ED_{100\pi}$ ratio is no longer the ratio of doses producing the same mean response. Instead, the log $ED_{100\pi}$ ratio is the horizontal component of the non-horizontal line segment connecting the dose-response curves at responses $100\pi\%$ of the way from L_i to U_i ($i = 0, 1$) (Fig. 3). Thus, when C_0 and C_1 have unequal limits, a definition based on the $ED_{100\pi}$ ratio embodies a modified concept of relative potency. For a given quantile π , the mean response to C_i is $\mu_i = L_i + (U_i - L_i)\pi$. Dividing dose $f^{-1}(\mu_0; \theta_0)$ by dose $f^{-1}(\mu_1; \theta_1)$, Dinse and Umbach [9] obtained:

$$\rho_{\pi}^*(\pi) = f^{-1}(L_0 + (U_0 - L_0)\pi; \theta_0) / f^{-1}(L_1 + (U_1 - L_1)\pi; \theta_1);$$

and, under the sigmoid model in Eq. 1, they showed that $\rho_{\pi}^*(\pi)$ reduces to:

$$\rho_{\pi}^*(\pi) = g^{-1}(\pi; \phi_0) / g^{-1}(\pi; \phi_1). \tag{6}$$

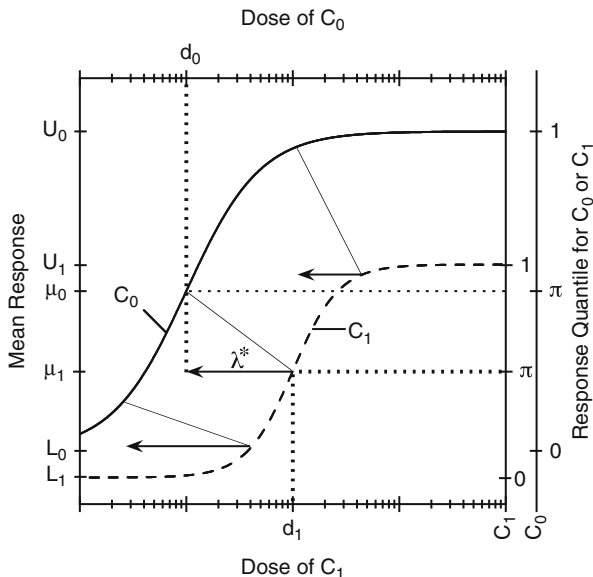


Fig. 3 Modified definition of relative potency (based on the ratio of $ED_{100\pi}$ values) indexed by response quantile and dose. Reference chemical C_0 and test chemical C_1 have different lower ($L_0 \neq L_1$) and upper ($U_0 \neq U_1$) response limits and, hence, distinct scales for each chemical on the response quantile axis. For a selected response quantile (same π but distinct μ_i for each chemical), an oblique line segment connects the point $(ED_{100\pi,1}, \mu_1)$ for C_1 to the point $(ED_{100\pi,0}, \mu_0)$ for C_0 . The horizontal component (depicted by an arrow) of each oblique segment represents a modified concept of log relative potency, whose value varies with π . For illustration, the arrow labeled λ^* (and its corresponding relative potency) can be indexed by response quantile (π), reference chemical dose (d_0), or test chemical dose (d_1), as indicated by the dotted lines

These equations express relative potency as a function of response quantile π for any $0 < \pi < 1$. We use the modified notation ρ^* to emphasize that this particular relative potency function does not, in general, embody the classical definition of relative potency.

Also, because one can index the log $ED_{100\pi}$ ratio by either the dose of the reference or test chemicals (Fig. 3), the modified definition of relative potency admits two other relative potency functions. Substituting the dose-quantile function $g(d_0; \phi_0)$ for π in Eq. 6 and noting that $g^{-1}(g(d_0; \phi_0); \phi_0) = d_0$, one may express the modified definition of relative potency as a function of reference dose d_0 :

$$\rho_{d_0}^*(d_0) = d_0 / g^{-1}(g(d_0; \phi_0); \phi_1). \tag{7}$$

Substituting $g(d_1; \phi_1)$ instead, the modified relative potency becomes a function of test dose d_1 :

$$\rho_{d_1}^*(d_1) = g^{-1}(g(d_1; \phi_1); \phi_0) / d_1. \tag{8}$$

Equations 6, 7 and 8 express the modified relative potency as functions of response quantile π , reference dose d_0 , and test dose d_1 , respectively, for all $\pi \in (0,1)$, $d_0 > 0$, and $d_1 > 0$.

Consideration of the modified definition of relative potency embodied in the ρ^* functions arose for two reasons. First, as mentioned in sections “[Defining Relative Potency as the Ratio of ED₅₀s](#)” and “[Evaluating Relative Potency at Multiple ED₁₀₀ \$\pi\$ Values](#)”, earlier authors have suggested using the ratio of ED₅₀s or of ED₁₀₀ π s to measure relative potency. The ρ^* functions are a natural extension of those earlier approaches so examining the implications of this modified definition seemed worthwhile. Second, before fitting dose-response models to compare chemicals, toxicologists sometimes re-express measured responses as a percent of a control mean for each chemical (e.g., perhaps a zero dose is expected to give a maximal response) or rescale them to a range set by mean responses to both positive and negative control treatments (e.g., normalized percent of activation) [20]. These transformations seem designed to remove extraneous variability from the data under a belief that rescaling makes sense when comparing chemicals (a point we return to later). Thus, consideration of the ρ^* functions also represented an effort to reflect common toxicologic practice, though without transforming measured responses.

Solving Eq. 5 for $f(d_1; \theta_1)$ yields $f(d_1; \theta_1) = f(d_1 \rho_{d_1}(d_1); \theta_0)$; that is, the dose-response function for C_1 can be expressed as the dose-response function for C_0 evaluated at dose $d_1 \rho_{d_1}(d_1)$. Similarly, Eq. 8 implies $g(d_1; \phi_1) = g(d_1 \rho_{d_1}^*(d_1); \phi_0)$. Consequently, specifying a dose-response (or dose-quantile) model and a relative potency model together is equivalent to specifying a pair of dose-response (or dose-quantile) models, a fact implicitly used by DeVito et al. [17, 18]. Recently, Dinse and Umbach [21] described conditions where modeling $\rho_{d_1}(d_1)$ (or $\rho_{d_1}^*(d_1)$) as a power function, $e^\eta d_1^\psi$, guaranteed that, for a wide range of popular dose-response models, the dose-response (or, respectively, dose-quantile) models for both chemicals would have the same functional form. They also pointed out that directly modeling ρ or ρ^* can sometimes facilitate inferences about relative potency functions.

Selecting Among Various Relative Potency Functions

The primary question is whether to use $\{\rho_\mu(\mu), \rho_{d_0}(d_0), \rho_{d_1}(d_1)\}$, the functions that embody the classical concept of relative potency, or $\{\rho_\pi(\pi), \rho_{d_0}^*(d_0), \rho_{d_1}^*(d_1)\}$, the functions that embody the modified concept. If the dose-response curves have identical response limits, both sets of functions are direct generalizations of the usual definition of relative potency as a ratio of equi-effective doses. Graphically these six relative potency functions convey essentially the same information because they all plot the same dose ratio as the ordinate, though each against a distinct abscissa, so the curves are differentially stretched horizontally (Fig. 4a–d).

If the response limits are not equal, however, $\{\rho_\pi(\pi), \rho_{d_0}^*(d_0), \rho_{d_1}^*(d_1)\}$, in using a modified definition of relative potency, can give a different impression than the other three relative potency functions based on the classical definition

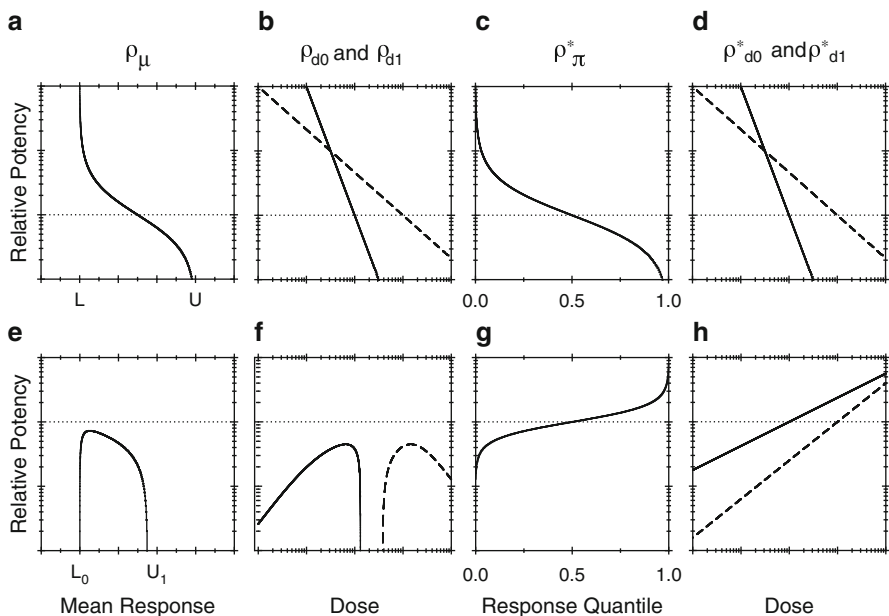


Fig. 4 Relative potency functions corresponding to the pairs of dose-response functions in Fig. 2 (panels a–d) and in Fig. 3 (panels e–h). The relative potency functions are: $\rho_\mu(\mu)$ (panels a, e); $\rho_{d0}(d_0)$ (solid) and $\rho_{d1}(d_1)$ (dashed) (panels b, f); $\rho_\pi^*(\pi)$ (panels c, g); and $\rho_{d0}^*(d_0)$ (solid) and $\rho_{d1}^*(d_1)$ (dashed) (panels d, h). The relative potency and dose axes are logarithmic; the mean response and response quantile axes are linear. Relative potency functions fall into two equivalence classes, $\{\rho_\mu(\mu), \rho_{d0}(d_0), \rho_{d1}(d_1)\}$ and $\{\rho_\pi^*(\pi), \rho_{d0}^*(d_0), \rho_{d1}^*(d_1)\}$, corresponding to the classical definition and to a modified definition of relative potency, respectively. The dotted horizontal line in each panel represents the ratio of ED_{50} s (Some panels are reproduced in part from Dinse and Umbach [9])

(Fig. 4e–h). When the response limits differ, the choice between these definitions depends on whether those differences are intrinsic or extrinsic to the chemicals [9]. For example, suppose two pesticides are compared with respect to the percentage of pests killed and a subset of the population is immune to one pesticide; thus, the upper response limit would be 100% for one pesticide and less than 100% for the other. These differences are intrinsic to the chemicals and should be taken into account by using $\{\rho_\mu(\mu), \rho_{d0}(d_0), \rho_{d1}(d_1)\}$. The convenient choice is to use $\rho_\mu(\mu)$ for ranking chemicals and $\rho_{d0}(d_0)$ or $\rho_{d1}(d_1)$ for dose conversion. On the other hand, suppose each chemical's dose-response study is performed in a different laboratory. Differences in response limits would be considered extrinsic if they were idiosyncratic to the specific laboratories rather than a property of the chemicals themselves. If response-limit differences are extrinsic, $\rho_\pi^*(\pi)$ should be used for ranking chemicals because it rescales the dose-response curves to the same response range. Likewise, $\rho_{d0}^*(d_0)$ and $\rho_{d1}^*(d_1)$ would be used to calculate equivalent doses of one chemical in terms of the other on a standardized response scale. Use of

$\{\rho_{\pi}^*(\pi), \rho_{d0}^*(d_0), \rho_{d1}^*(d_1)\}$ is in accord with the toxicologic practice of rescaling responses as a percent of control mean response and is preferable to rescaling the data, which can introduce correlations that are not accounted for by most standard analyses.

Example

We analyzed data from U.S. National Toxicology Program (NTP) bioassays evaluating 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) and 2,3,4,7,8-pentachlorodibenzofuran (PeCDF) [22, 23]. We focused on cytochrome *P*450 1A1-associated 7-ethoxyresorufin-*O*-deethylase (EROD) activity measured in liver tissue of female Harlan Sprague–Dawley rats treated by oral gavage for 14 weeks. Both studies involved 10 rats in each of 6 dose groups (control plus 5 exposure levels). Our estimates of relative potency functions were derived from parameters estimated by fitting dose-response models using the relationships described earlier.

We analyzed log-transformed enzyme activity via least squares. We used Proc GLM in SAS (version 9.3, SAS Institute Inc., Cary, NC, USA) to fit a saturated analysis-of-variance model that estimates a mean response for each dose level of each chemical and Proc NLIN to fit nonlinear regression models. All analyses assumed a common residual variance across dose levels and chemicals. Dose-response models were Hill models based on Eq. 1 with $g(d; \phi) = d^S / (d^S + M^S)$. We compared the fit of nested models with *F* tests [24] based on residual sums of squares and constructed simultaneous confidence bands for relative potency functions using Scheffe's method [24].

An 8-parameter model based on two separate Hill models (Table 1) showed no lack of fit (Fig. 5a, b) compared to a saturated analysis-of-variance model with 12 parameters ($F_{4,108} = 0.06$, $p = 0.99$). However, a 6-parameter model with common response limits for TCDD and PeCDF did not fit as well as the 8-parameter model ($F_{2,112} = 14.44$, $p < 0.0001$). We conclude that the chemicals have different response limits. Consider ρ_{d1} as an example. If one regarded these response-limit differences as intrinsic to the chemicals, estimation of ρ_{d1} as a function of PeCDF dose should use Eq. 5. The differences in response limits guarantee that ρ_{d1} is non-constant. The estimated ρ_{d1} is below one for most of the dose range but exceeds one at either edge of that range (Fig. 5c), suggesting that PeCDF is generally less toxic than TCDD. On the other hand, if one regarded the response-limit differences as extrinsic to the chemicals, estimation of ρ_{d1}^* as a function of PeCDF dose should use Eq. 8. Relative potency modeled as a power function of PeCDF dose, $\rho_{d1}^*(d_1) = e^{\eta} d_1^{\psi}$, a straight line in log-log plots (Fig. 5d), fit no better than $\rho_{d1}^*(d_1) = e^{\eta}$ for these data ($F_{1,112} = 0.24$, $p = 0.63$) (Table 1). This conclusion is consistent with the horizontal line at 0.06 ($= e^{-2.76}$), the estimate of modified relative potency as constant, remaining within the 95% confidence band for the power-function estimate (Fig. 5d). We do not know enough about the details of

Table 1 Parameter and standard error (SE) estimates for nested Hill dose-response models fitted to liver EROD activity^a in rats after 14-week exposure to TCDD (reference chemical, subscripted 0) or PeCDF (test chemical, subscripted 1)^b

Separate 4-parameter Hill Model for Each Chemical ^c			
Parameterization 1		Parameterization 2	
Parameter	Estimate (SE)	Parameter	Estimate (SE)
L_0	30.21 (1.96)	L_0	30.21 (1.96)
U_0	2241 (217)	U_0	2241 (217)
S_0	0.94 (0.24)	S_0	0.94 (0.24)
M_0	4.79 (1.30)	M_0	4.79 (1.30)
L_1	48.66 (3.16)	L_1	48.66 (3.16)
U_1	2852 (389)	U_1	2852 (389)
S_1	1.09 (0.09)	η	-3.30 (1.05)
M_1	65.06 (18.3)	ψ	0.16 (0.31)
MSE (df) ^e	0.0422 (112)		0.0420 (113)

^aActivity measured as nmol of resorufin formed per min per mg of microsomal protein

^bDoses measured as ng per kg of body weight per day

^cThe parameterizations are related by: $\psi = S_1/S_0 - 1$ and $\eta = \log(M_0) - (S_1/S_0)\log(M_1)$. The first parameterization is used to estimate $\rho_{d1}(d_1)$ via Eq. 5.

The second parameterization is used to estimate $\rho_{d1}^*(d_1) = e^{\eta} d_1^{\psi}$

^dRelative potency (modified definition) is constant: $\rho_{d1}(d_1) = e^{\eta}$

^eMean Squared Error; an estimate of the residual variance, and associated degrees of freedom

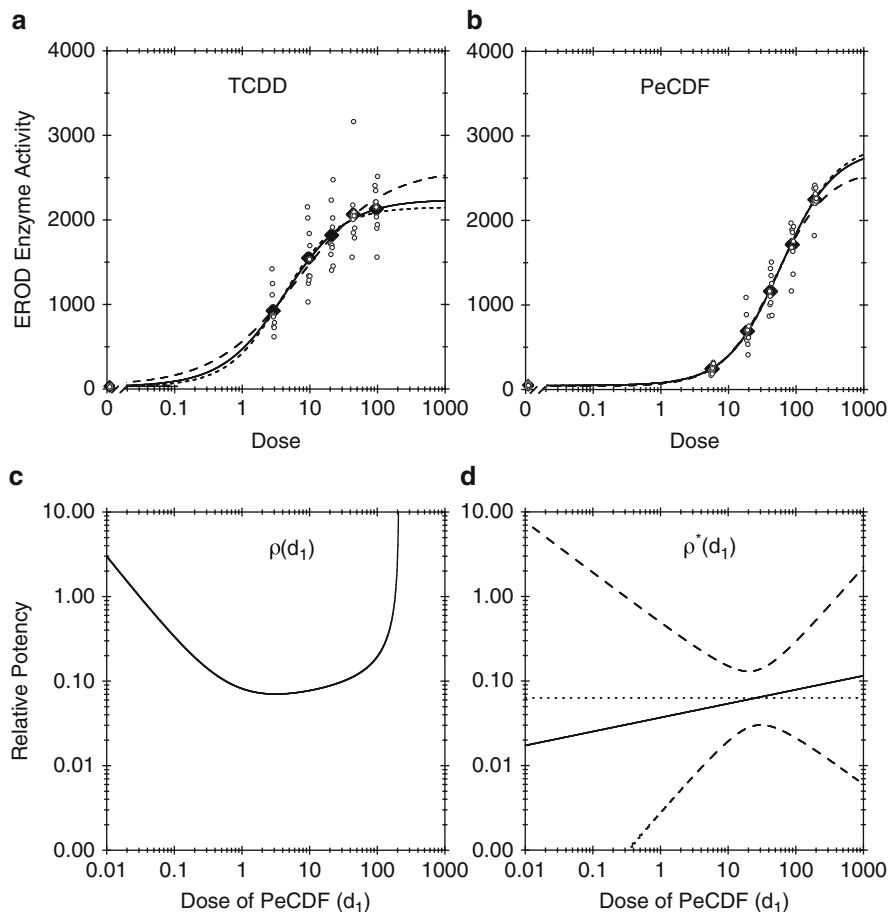


Fig. 5 Dose-response and relative potency for TCDD and PeCDF for liver EROD activity (pmol of resorufin formed per min per mg of microsomal protein) in rats after 14-week exposure via oral gavage. Dose units are ng per kg of body weight per day. Panels (a) TCDD and (b) PeCDF show observed activity for each rat (\circ), dose-specific means (\blacklozenge), and estimated dose-response curves (solid, 8-parameter model with a separate Hill function for each chemical; dashed, 6-parameter model honoring a constraint that both chemicals have same response limits; dotted, 7-parameter model honoring a constraint that $\rho_{d1}^*(d_1)$ is constant). Panel (c) shows an estimate of $\rho_{d1}(d_1)$. Panel (d) shows estimates of $\rho_{d1}^*(d_1)$ (solid, as a power-function; dashed, its 95% simultaneous confidence band; dotted, as a constant)

the experiments and the biology to decide whether the response-limit differences should be regarded as intrinsic or extrinsic to these chemicals. Regardless of that judgment, however, these data support a conclusion that PeCDF is less potent than TCDD.

Summary

The idea that relative potency should be constant is rooted historically in analytical dilution assays. It simplifies chemical ranking and dose conversion. If relative potency is a constant equal to ρ , the dose of chemical C_0 that is equivalent to dose d_1 of chemical C_1 is $d_0 = d_1\rho$, and the dose of C_1 that is equivalent to dose d_0 of C_0 is $d_1 = d_0/\rho$. Chemical ranking is even easier: order each chemical by its value of ρ .

Toxicologists, however, have long been faced with data from comparative assays that indicate that relative potency is not generally constant. Over the years, various investigators have suggested ways to cope with non-constant relative potency. Extending the concept that relative potency is the ratio of equi-effective doses, Cornfield [16] showed that linear dose-response models in log dose induce relative potency functions that are log-linear in log dose or response. In fact, a wide variety of monotone dose-response models can be inverted to express relative potency as a function of reference dose, test dose, or mean response. Analogously, using a modified concept of relative potency as the ratio of $ED_{100\pi}$ s, one can express (modified) relative potency as a function of reference dose, test dose, or response quantile. If the chemicals have the same response limits, the classical and modified definitions of relative potency coincide. Relative potency functions allow chemicals to be ranked with respect to toxicity, though that ranking may change for different dose or response levels. For dose conversion, the dose of C_0 that is equivalent to dose d_1 of C_1 is $d_0 = d_1\rho_{d1}(d_1)$ and the dose of C_1 that is equivalent to dose d_0 of C_0 is $d_1 = d_0/\rho_{d0}(d_0)$. The choice between $\{\rho_{\pi}^*(\pi), \rho_{d0}^*(d_0), \rho_{d1}^*(d_1)\}$, based on the modified concept of relative potency, and $\{\rho_{\mu}(\mu), \rho_{d0}(d_0), \rho_{d1}(d_1)\}$, based on the classical definition, depends on whether response limits differ for extrinsic or intrinsic reasons. Relative potency functions appear to be a promising avenue for characterizing non-constant relative potency.

Acknowledgements We are grateful to Grace Kissling for reading an early version of the manuscript and to an anonymous reviewer for helpful comments. This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES-102685).

References

1. Hannas, B.R., Lambricht, C.S., Furr, J., Evans, N., Foster, P.M.D., Gray, E.L., Wilson, V.S.: Genomic biomarkers of phthalate-induced male reproductive developmental toxicity: a targeted RT-PCR array approach for defining relative potency. *Toxicol. Sci.* **125**, 544–557 (2012)
2. Jensen, B.H., Petersen, A., Christiansen, S., Boberg, J., Axelstad, M., Herrmann, S.S., Poulsen, M.E., Hess, U.: Probabilistic assessment of the cumulative dietary exposure of the population of Denmark to endocrine disrupting pesticides. *Food Chem. Toxicol.* **55**, 113–120 (2013)
3. Haws, L.C., Su, S.H., Harris, M., DeVito, M.J., Walker, N.J., Farland, W.H., Finley, B., Birnbaum, L.S.: Development of a refined database of mammalian relative potency estimates for dioxin-like compounds. *Toxicol. Sci.* **89**, 4–30 (2006)

4. Finney, D.J.: The meaning of bioassay. *Biometrics* **21**, 785–798 (1965)
5. Hill, A.V.: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.* **40**(Suppl), iv–vii (1910)
6. Finney, D.J.: Radioligand assay. *Biometrics* **32**, 721–740 (1976)
7. Christensen, E.R., Nyholm, N.: Ecotoxicological assays with algae: Weibull dose-response curves. *Environ. Sci. Technol.* **18**, 713–718 (1984)
8. Brain, P., Cousens, R.: An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Res.* **29**, 93–96 (1989)
9. Dinse, G.E., Umbach, D.M.: Characterizing non-constant relative potency. *Regul. Toxicol. Pharmacol.* **60**, 342–353 (2011)
10. Bagdy, G., To, C.T.: Comparison of relative potencies of i.v. and i.c.v. administered 8-OH-DPAT gives evidence of different sites of action for hypothermia, lower lip retraction and tail flicks. *Eur. J. Pharmacol.* **323**, 53–58 (1997)
11. Rodbard, D., Frazier, G.R.: Statistical analysis of radioligand assay data. *Methods Enzymol.* **37**, 3–22 (1975)
12. Guardabasso, V., Rodbard, D., Munson, P.J.: A model-free approach to estimation of relative potency in dose-response curve analysis. *Am. J. Physiol.* **252**, E357–E364 (1987)
13. Guardabasso, V., Munson, P.J., Rodbard, D.: A versatile method for simultaneous analysis of families of curves. *FASEB J.* **2**, 209–215 (1988)
14. Streibig, J.C., Walker, A., Blair, A.M., Anderson-Taylor, G., Eagle, D.J., Friedlander, H., Hacker, E., Iwanzik, W., Kudsk, P., Labhart, C., Luscombe, B.M., Madafiglio, G., Nel, P.C., Pestemer, W., Rahman, A., Retzlaff, G., Rola, J., Stefanovic, L., Straathof, H.J., Thies, E.P.: Variability of bioassays with metsulfuron-methyl in soil. *Weed Res.* **35**, 215–224 (1995)
15. Villeneuve, D.L., Blankenship, A.L., Giesy, J.P.: Derivation and application of relative potency estimates based on in vitro bioassay results. *Environ. Toxicol. Chem.* **19**, 2835–2843 (2000)
16. Cornfield, J.: Comparative assays and the role of parallelism. *J. Pharmacol. Exp. Ther.* **144**, 143–149 (1964)
17. DeVito, M.J., Diliberto, J.J., Ross, D.G., Menache, M.G., Birnbaum, L.S.: Dose-response relationships for polyhalogenated dioxins and dibenzofurans following subchronic treatment in mice: I. CYP1A1 and CYP1A2 enzyme activity in liver, lung, and skin. *Toxicol. Appl. Pharmacol.* **147**, 267–280 (1997)
18. DeVito, M.J., Menache, M.G., Diliberto, J.J., Ross, D.G., Birnbaum, L.S.: Dose-response relationships for induction of CYP1A1 and CYP1A2 enzyme activity in liver, lung, and skin in female mice following subchronic exposure to polychlorinated biphenyls. *Toxicol. Appl. Pharmacol.* **167**, 157–172 (2000)
19. Ritz, C., Cedergreen, N., Jensen, J.E., Streibig, J.C.: Relative potency in nonsimilar dose-response curves. *Weed Sci.* **54**, 407–412 (2006)
20. Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J., Nadon, R.: Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **89**, 4–30 (2006)
21. Dinse, G.E., Umbach, D.M.: Parameterizing dose-response models to estimate relative potency functions directly. *Toxicol. Sci.* **129**, 447–455 (2012)
22. National Toxicology Program: NTP technical report on the toxicology and carcinogenesis studies of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) (CAS no. 1746-01-6) in female Harlan Sprague-Dawley rats (Gavage studies). Technical report series no. 521, NIH publication no. 06-4468, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, RTP, NC (2006)
23. National Toxicology Program: NTP technical report on the toxicology and carcinogenesis studies of 2,3,4,7,8-pentachlorodibenzofuran (PeCDF) (CAS no. 57117-31-4) in female Harlan Sprague-Dawley rats (Gavage studies). Technical report series no. 525, NIH publication no. 06-4461, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, RTP, NC (2006)
24. Seber, G., Wild, C.: *Nonlinear Regression*. Wiley, New York (1989)

Development and Validation of Exposure Biomarkers to Dietary Contaminants Mycotoxins: A Case for Aflatoxin and Impaired Child Growth

Paul Craig Turner and Barbara Zappe Pasturel

Abstract Mycotoxins are toxic secondary metabolites that globally contaminate an estimated 25% of cereal crops and thus exposure is frequent in many populations. The heterogeneous distribution of mycotoxins in food restricts the usefulness of food sampling and intake estimates for epidemiological studies; instead exposure biomarkers provide better tools for informing epidemiological investigations. Aflatoxins, fumonisins and deoxynivalenol are amongst those mycotoxins of particular concern from a human health perspective. Validated exposure biomarkers for aflatoxin (urinary aflatoxin M1, aflatoxin-N7-guanine, serum aflatoxin-albumin) were important in confirming aflatoxins as ‘Group 1’ liver carcinogens. For fumonisins and deoxynivalenol these steps for exposure biomarker development and validation have significantly advanced in recent years. Such biomarkers should better inform epidemiological studies and thus improve our understanding of their potential risk to human health. In West Africa it has been suggested that growth faltering in children is not fully explained by poor nutrition and infection. This review highlights some of the recently emerging epidemiology that strongly implicates a role for aflatoxins in this growth faltering, and suggests potential mechanisms. The use of aflatoxin exposure biomarkers were essential in understanding the observational data reviewed, and will likely be critically monitors of the effectiveness of interventions to restrict aflatoxin exposure.

Introduction to Mycotoxins

Fungi are important sources of dietary nutrition (mushrooms, cheeses) and medicines (penicillin, statins), but can also produce toxic secondary metabolites known as mycotoxins. These potent dietary toxins are estimated to contaminate 25%

P.C. Turner (✉) • B.Z. Pasturel
Maryland Institute for Applied Environmental Health, School of Public Health,
University of Maryland, College Park, MD, USA
e-mail: pturner3@umd.edu; barbaraz_hap@yahoo.com

M.-L.T. Lee et al. (eds.), *Risk Assessment and Evaluation of Predictions*,
Lecture Notes in Statistics 215, DOI 10.1007/978-1-4614-8981-8_16,
© Springer Science+Business Media New York 2013

of the world's cereal crops [1] making exposure frequent among many populations. Among the hundreds of mycotoxins identified, those of major public health concern include aflatoxins (AF) produced from *Aspergillus*, and both fumonisins (FB), and trichothecenes (deoxynivalenol (DON), nivalenol, T2-toxin) from *Fusarium*. AF and FB are more frequent contaminants of crops in hot and humid climates as in Central America, tropical Asia and sub-Saharan Africa where staple foods such as maize and groundnuts (peanuts) are often contaminated. For AF both field growth and long-term storage contribute to the burden of contamination, whilst FB is predominantly a field toxin of maize [2], Trichothecenes e.g. DON are more prevalent in temperate climates, and also tend to accumulate in the field to a greater extent than during dry storage. They occur in a variety of grains, though wheat and maize are the predominantly contaminated cereal [3]. Mycotoxins' resistance to processing, and their stability during cooking, also contribute to dietary exposure [2]. Particularly vulnerable populations are those with limited dietary variation and a heavy reliance on one or two high risk dietary staples. For this reason those individuals at greatest risk are often from some of the poorest countries, with inadequate or absent regulations and limited food choices.

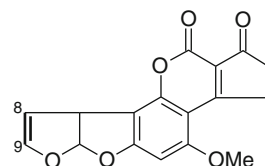
With the exception of the aflatoxins, mycotoxins as a group of contaminants remain a mostly poorly examined global health issue, despite the known frequency of exposure and the demonstrated animal toxicities [4]. Aflatoxins are potent human carcinogens, suspected human growth modulators, and in animals cause cancer and effect growth and immune function; fumonisins are suspected human carcinogens, and recently postulated growth modulators, and in animals cause diverse toxicity including cancer, neural tube defects, equine leukoencephalomalacia and porcine pulmonary edema; whilst deoxynivalenol has effects on the GI tract and immune system of animals, and is suspected to cause growth faltering [4]. The heterogeneous distribution of mycotoxins in the diet has restricted more classical epidemiological approaches partly because these struggle to clearly define exposure. However the development, validation and use of exposure biomarkers offer improved exposure assessment. This short review highlights the development and use of mycotoxin exposure biomarkers, and focuses on the emerging relationship between early life exposures to and growth faltering in West African infants and young children.

Mycotoxin Exposure Biomarkers

Biomarkers for Aflatoxin

Among the naturally occurring aflatoxins, aflatoxin B1 (AFB1) occurs most frequently and is the most toxic and carcinogenic. AFB1 is metabolized by a number of cytochrome P450 enzymes [5, 6] generating hydroxy-metabolites (e.g. AFM1, AFQ1 and AFP1) and two reactive epoxide species, aflatoxin B1 exo-8,9-epoxide and endo-8,9-epoxide. The epoxides can cause cellular and macromolecule

Fig. 1 Chemical structure of the major naturally occurring aflatoxin – aflatoxin B1



damage by binding to proteins and nucleic acids [7–10]. Aflatoxin B1 exo-epoxide preferentially binds to guanine residues in DNA, and following depurination of this adduct, aflatoxin-N7-guanine (AF-N7-Gua) is detected in urine, in addition to the hydroxyl metabolites, and the un-metabolized parent toxins [7, 11] AFB1-N7-Gua and AFM1 in urine are both strongly correlated with aflatoxin intake in chronically exposed individuals, ($r = 0.80$, $p < 0.0001$ and $r = 0.82$, $p < 0.0001$ for AFB1-N7-Gua, and AFM1) [12–15]. These quantitative relationships for urinary aflatoxins provide confidence in the use of these measures as exposure assessment tools (Fig. 1).

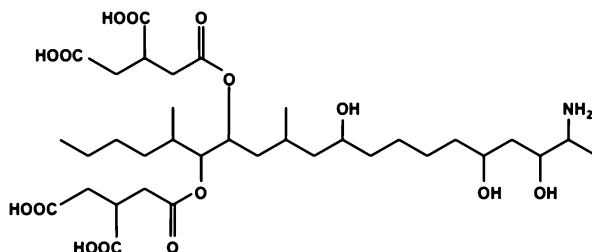
Hydrolysis of both epoxides allows protein adduction and toxicity, and the formation of aflatoxin-albumin, which can readily be observed in the sera of exposed animals and humans [16–41]. The concentration of aflatoxin-albumin in dietary exposed individuals strongly correlates with aflatoxin intake ($r = 0.69$, $p < 0.0001$); providing an additional exposure biomarker that based on the half-life of albumin represents an integrated assessment of exposure over a period of two to three months [18, 22]. Aflatoxin-albumin adduct levels were additionally demonstrated to be linear with dose in rodents across a dosing range from 0.16 ng/kg body weight (bw) to 12,300 ng/kg bw ($r^2 = 0.98$), and importantly, typical human exposures within low, moderate and high risk communities all fall within this experimental range [41]. Neither AFB1 nor other AFB1 metabolites in urine have been demonstrated to be correlated with the dose [12]. For this reason, urinary AFB1 is informative to some extent that exposure occurred, but does not provide a useful indicator of the amount of that exposure.

In high risk regions of the world, greater than 95% of those individuals tested are positive for aflatoxin-albumin over a 3 log range, from approximately 3–5 pg/mg albumin to >1,000 pg/mg [16–38], while more developed regions rarely have detectable levels of the biomarker [23, 42].

Biomarkers for Fumonisin

Fumonisin do not appear to undergo significant metabolism [43–49], thus biomarker development has not followed the metabolite profile approach used for aflatoxin. Fumonisin inhibit sphingolipid metabolism by competing with ceramide synthase [43, 50, 51]. Their capacity to alter levels of sphingoid bases, as observed in experimental animals [50–53], is plausibly linked to their suggested mechanism of toxicity [43, 52–54] including cancer and neural tube defects. Animal studies

Fig. 2 Chemical structure of the major naturally occurring fumonisin – fumonisin B1



indicate that the transfer of fumonisins to urine was around 0.4–2.0% of that ingested [44–49], though typically these percentages refer to total transfer over several days, and often at doses higher than would be observed in humans.

One human study examined tortilla consumption in Mexican women and found that urinary FB1 was detected more frequently in the group with high consumption (96%) compared to medium (80%) and low (45%) consumption [55]. The geometric mean urinary FB concentrations were also associated ($p < 0.001$) with consumption of tortillas (geometric means and 95th percentile were 147 pg/ml (88, 248 pg/ml), 63 pg/ml (37, 108 pg/ml) and 35 pg/ml (19, 65 pg/ml), respectively). In a separate study the concentration of urinary FB1 was measured in Chinese adults. Comparison between two counties with similar frequency of FB1 detection revealed mean concentrations of 13,630 pg/mg creatinine (range nd – 256,000 pg/mg; median 3,910 pg/mg) in Huaian county as compared to 720 pg/mg (range nd – 3,720 pg/mg; median 390 pg/mg) in Fusui county [56], though no significant correlation between urinary FB1 and estimated intake was found. This lack of correlation may reflect the fact that food frequency questionnaires (FFQ) measured typical intake over weeks while urinary measures more typically reflect more recent intake, though the toxicokinetics of urinary FB are not clearly defined as yet. Their data suggested that FB intakes were at least threefold higher in Huaian County and that about 1–2% of the ingested FB was transferred to urine. A study in South Africa attempted to better assess the relationship between urinary FB1 and FB1 ingestion using measures from plate ready food (a maize porridge). A moderate correlation ($r^2 = 0.31$, $p < 0.001$) was observed between estimated FB1 intake/kg bw/day and urinary FB1 adjusted for creatinine. In that study the transfer of ingested FB1 to urine was estimated to be 0.075%. Fumonisin disruption of sphingolipid metabolism and the associated levels of sphingoid bases e.g. sphinganine-1-phosphate is another area being investigated for the development of fumonisin exposure biomarkers [57, Riley et al. 2012, manuscript in preparation] (Fig. 2).

Biomarkers for DON

DON, is a type B trichothecene mycotoxin predominantly associated with crop contaminations such as *Fusarium* head blight in wheat and *Gibberella* ear rot in maize [1]. Also known as vomitoxin due to its potent gastrointestinal effects,

DON can be detoxified by gut microbiota to DOM-1, a de-epoxy metabolite, or metabolized in the liver to a glucuronide. Meko and colleagues [58] suggested the combined measure of un-metabolized DON or 'free' DON (fD) and DON-glucuronide (DG) as a putative urinary exposure biomarker [59]. In a survey of UK adults, urinary fD + DG from a single 24 hour void was detected in 98.7% of individuals (geometric mean 8.9 ng/mg; 95%CI: 8.2, 9.7; range nd to 48.2 ng/mg) [60] and a modest, but significant, positive association was observed between the urinary measure and cereal consumption ($p < 0.001$, $R^2 = 0.23$). A survey over six days confirmed this relationship between cereal intake and the urinary bio-measure ($p < 0.001$, $R^2 = 0.23$) [61]. In addition, a four-day survey assessing DON intake revealed that on a daily basis, urinary fD + DG was strongly correlated with DON intake ($p < 0.001$, $R^2 = 0.56, 0.49, 0.54, 0.64$, for each day respectively), and an integrated assessment of the four days combined revealed a highly significant correlation, which remained after adjustment for age, sex and BMI ($p < 0.001$, $R^2 = 0.83$).

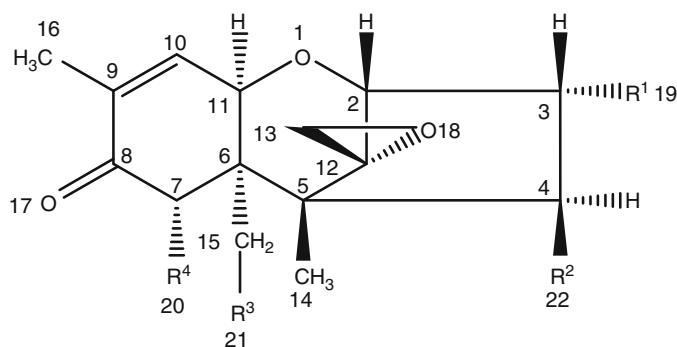
Based on the strong quantitative relationship between exposure and the bio-measure, and the stability in both ambient room temperature in short term (24 hour) and cryo-preservation in the long term (years), urinary fD + DG is now regarded as useful exposure biomarker [59–63]. DOM-1 is an important detoxification product of DON in many species, but to date it is either absent or rarely detected in human urine from DON exposed individuals, suggestive that humans may be one of the more sensitive species to DON toxicity. Approximately 73% of the ingested DON is transferred to urine as fD + DG [61]. To date DON is enriched from urine samples prior to being quantified by LC/MS. The relatively high levels of DON that are typically being observed in urine provides an opportunity to explore more rapid methods as suggested by colleagues in Austria [64] (Fig. 3).

Summary of Biomarker Approaches

Table 1 summarizes the mycotoxin biomarker approaches discussed here. There are significant differences in the urinary assays for aflatoxin, fumonisin and deoxynivalenol as analytic sensitivity depends on the sensitivity of exposure assessment as well as the transfer kinetics of each toxin. For this reason, similar levels of urinary bio-measures do not necessarily represent similar levels of exposure. Of course similar levels of exposure do not necessarily reflect similar risk of concern; for example aflatoxin is significantly more toxic than both fumonisin and DON.

Aflatoxins and Chronic Disease

Aflatoxins are proven hepatocarcinogens, classified by the International Agency for Research on Cancer (IARC) as Group 1 human carcinogens [9, 65]. The role of aflatoxin in the etiology of liver cancer is widely recognized. This role has been



	R ¹	R ²	R ³	R ⁴
Nivalenol	OH	OH	OH	OH
Deoxynivalenol	OH	H	OH	OH
3-Acetyl-Deoxynivalenol	OCOCH ₃	H	OH	OH
15-Acetyl-Deoxynivalenol	OH	H	OCOCH ₃	OH
Fusarenon X	OH	OCOCH ₃	OH	OH

Fig. 3 Chemical structure of the major naturally occurring trichothecenes including DON

systematically demonstrated through animal models and molecular epidemiology approaches involving the use of biomarkers of exposure and effect. A synergistic interaction between aflatoxin exposure and hepatitis B virus carriage is well documented [65], however the mechanism remains unclear, and worthy of further investigation [8, 30, 34]. More recently, however, AF exposure biomarkers have begun to reveal additional health concerns, namely growth faltering, immune suppression.

Aflatoxin Exposure and Impaired Growth

Chronic aflatoxin exposure in many regions of Sub-Saharan is endemic, and two decades of biomarker-driven research demonstrate that exposure occurs in utero, during early life and childhood and continues into adulthood [4]. In early childhood (<5 years) growth faltering in many sub-Saharan Africa is common, in excess of 30% are stunted (height for age Z-score (HAZ) < -2) or extremely stunted (HAZ < -3) [66], but in the Gambia at least it was revealed that the stunting was not sufficiently explained by either lack of nutrition or by infectious episodes [67–69]. In separate studies inverse relationships between growth and aflatoxin exposure are being revealed [27, 28, 31, 33]. Infants are typically weaned using family foods, which are frequently contaminated by aflatoxins. Significant transitions

Table 1 Summary of mycotoxin biomarkers

Parent mycotoxin Biomarker	<i>Aflatoxin B1</i> Aflatoxin -N7-Guanine (urine)	<i>Aflatoxin B1</i> Aflatoxin M1 (urine)	<i>Aflatoxin B1</i> (Aflatoxin-albumin) (sera)	<i>Fumonisin B1</i> Fumonisin B1 (urine)	<i>Deoxynivalenol</i> Deoxynivalenol plus deoxynivalenol-glucuronide (urine)
<i>Transfer</i> ^a	Approx. 1%	1-3%	1-2%	0.075%	73%
<i>Correlation coefficient</i> ^b	0.80*	0.82*	0.69*	0.49**	0.91*

Modified from [4]

* $p < 0.001$; ** $p < 0.01$ ^aAssessment of transfer from dietary intake to biomarker in sera or urine^bCorrelation coefficient of the regression model for intake against the urinary biomarker

in the mean aflatoxin biomarker levels are apparent as infants are first introduced to weaning foods [31] and when they transition from receiving a mixture of weaning foods and breast milk to family foods [70].

One Gambian study in which aflatoxin biomarkers were assessed in maternal (during pregnancy), cord blood, week 16 infant and week 52, aflatoxin-albumin adduct positivity was 100% (range 5–400 pg aflatoxin-lysine/mg albumin), 49% (range nd-50 pg/mg), 11% (nd – 50 pg/mg) and 92% (nd – 390 pg/mg) respectively [31, 32]. The cord blood data indicates both in utero exposure, and, the requisite metabolic capacity to activate the toxin to reactive epoxides. Maternal and week 16 aflatoxin-albumin combined were significantly negatively correlated ($p < 0.001$) with growth velocity of the infant in the first year of life [31]. These data suggested that reduction in maternal aflatoxin-albumin adduct during pregnancy from 110 pg/mg to 10 pg/mg would improve linear growth in the first year of life by 2 cm and weight by 0.8 kg.

In Benin HAZ and weight for age Z-scores (WAZ) in children aged 9 months to 5 years were inversely associated with the aflatoxin exposure biomarker, indicative of a relationship between aflatoxin and both stunting and being underweight, $p < 0.001$ for both; data supported by a subsequent longitudinal study in Beninese children aged around 2–3 years from regional villages [28]. Data from this latter study was suggestive that a 100 pg/mg difference in exposure approximates to about a 1 cm reduction in height over an 8-month period in this age group.

A cross sectional study of slightly older Gambian children (aged six to nine years) revealed a less significant relationship between aflatoxin exposure and growth [33]. This observation could be suggestive of more significant effect of younger children, which is plausible given the more rapid growth in younger children, perhaps providing a greater opportunity for a toxic insult to have an observable effect. However, it is also worth noting that despite both the Beninese with Gambian cross-sectional studies having chronic aflatoxin exposure, there were some differences. In the former 99% of the children were positive (geometric mean 32.8 pg/mg: range 5–1064 pg/mg) [27], whilst in the latter 93% of the children were positive (geometric mean 22.3 pg/mg; range 5–456 pg/mg), [33]. Perhaps most importantly in the former study the percentage of children exceeding a biomarker concentration of 100 pg/mg (16%) was more than twice that of the latter (7%). Thus it remains unclear whether age or precise exposure burden was the stronger driver for these observational data; to date no threshold has been established for aflatoxin and growth.

One important aspect of all these data is that to date no single study has followed exposure through pregnancy and into the first 3–5 years of life. This could be valuable in terms of clearly establishing critical windows of exposure, understanding mechanisms of effect, and opportunities and timing of intervention. One important public health message could simply be the reinforcement of the prolonging of breastfeeding in these populations. Whilst aflatoxins are transferred to breast milk, the levels are modest compared to that in weaning foods [9, 65] and through the weaning process aflatoxin biomarker levels follow a pattern from low to high as you move from exclusive breast fed < partial breast fed < fully weaned

[31, 70]. Other intervention strategies that either reduce aflatoxin contamination [30] or effect uptake or metabolism [8] will also be important to protect maternal and post weaning phases of exposure. The development of sustainable targeted interventions should be a priority given the clear burden of exposure.

Aflatoxin and Gastrointestinal Toxicity

The mechanism(s) by which aflatoxin may affect child growth remain unclear, but possibilities include immune suppression, altered growth factor expression or intestinal toxicity [71, 72]. Each of these could contribute to growth faltering at different stages of child development, especially during the period of dynamic changes to nutritional intake in early life when the shift from breast milk to solid food exposes a child to dietary contaminants that could affect immunity and gastrointestinal tract integrity. Since the CYP3A enzymes that bio-activate AFB1 are also expressed in human intestinal epithelial cells, the GI tract is a primary target for aflatoxin-induced damage, particularly in the tight junctions that regulate paracellular permeability. Aflatoxin appears to modulate paracellular transport in confluent Caco-2 monolayers, making the barrier more 'leaky' [73]. Since one of the key toxic effects of aflatoxin is disruption of phosphorylation patterns of structural and enzymic proteins [74], it is plausible that 'leaky' tight junctions reflect aflatoxin-induced disruption of intercellular functional protein complexes that form tight junctions. Indeed, West African studies have indicated such intestinal membrane permeability in young children and this 'leakiness' aka intestinal enteropathy is strongly associated with the degree of growth faltering [68, 75, 76]. The observed villous shortening, crypt hyperplasia and lymphocyte infiltration [68, 75–77] which lead to a decrease in intestinal surface area, elevated inflammatory markers and subsequent decreased absorption of sugars, may stem from recurrent exposure to infectious agents and to damage caused by aflatoxin. This latter hypothesis requires further supportive mechanistic data.

Aflatoxin and Immune Suppression

Aflatoxins have potent effects on the immune system and host suppression and increased susceptibility to infections are clearly demonstrated in animals, whilst the effects in humans' remains poorly examined. In one study of Gambian children the concentration of salivary sIgA, which binds to bacterial and viral surface antigens as part of the mucosal barrier, were significantly reduced in aflatoxin-exposed children [33]. Alterations in cellular immunity have also been observed in Ghanaian adults [78]. Intestinal reduction of sIgA may be a contributing factor to decreased bacterial resistance and to increased epithelial inflammation.

Aflatoxin, Zinc and Insulin-Like Growth Factor

Dietary zinc deficiency, whose human symptoms include growth retardation, skin abnormalities and mental lethargy [79], has been recognized as a health concern for about 50 years and poses a particular problem in developing countries where studies support supplementation of children aged <5 years in order to improve linear growth and reduce stunting [80]. In a study investigating the effects of aflatoxin exposure during pregnancy in swine, the offspring from dosed sows showed significant reduced growth, an effect related to a reduced capacity to properly utilize zinc, despite the diet being zinc sufficient [81]. It was notable that this study involved relatively low levels of aflatoxin. Measures of zinc- both free and carried by thymulin- would be valuable in aflatoxin-exposed children, in relation to anthropometry.

Another potential link between aflatoxin exposure and growth concerns the liver-derived insulin-like growth factor IGF1, which affects linear bone growth. Microarray data revealed down-regulation of genes responsible for, among other things, IGF1 in aflatoxin-treated chicks [82]. One study in Gabon examined nutritional status (kwashiorkor and marasmus) in children less than 30 months and found reduced IGF1 levels in malnourished subjects [83], though aflatoxins were not measured. Additional research is required to understand the potential mechanism(s) of aflatoxin induced growth faltering. It will be important to consider other mycotoxins, as co-exposure to multiple mycotoxins will be the norm rather than the exception. With the recent development of exposure assessment tools for some *Fusarium* mycotoxins as described above, this need can now to some extent be met. Intervention studies to restrict exposure may be invaluable in clearly defining the role and mechanism of aflatoxins and infant growth.

Conclusion

When one considers that, worldwide, 40% of the 11 million deaths in children aged less than 5 years old occur in sub-Saharan Africa [66] and that approximately half of the deaths linked to infectious diseases in sub-Saharan African children point to under-nutrition and slowed growth as an underlying cause, the urgent need for immediate interventions and further research into the effect of food contaminants on public health becomes self-evident. Since mycotoxin-contaminated foods constitute a large portion of daily dietary intake for many of the world's developing nations, assessments of mycotoxin exposure are essential, and the need for clarification of the biological mechanisms involved. Such understanding of the health risks may lead to targeted, affordable and sustainable methods being established to restrict such exposures among those at highest risk and to reduce the overall burden of mycotoxin-driven chronic disease.

References

1. Council for Agricultural Science and Technology (CAST): Potential economic costs of mycotoxins in the United States. In: *Mycotoxins: Risks in Plant, Animal and Human Systems*. Task force report no 139, Ames, pp. 136–142 (2003)
2. Miller, J.D.: Fungi and mycotoxins in grain: implications for stored product research. *J. Stored Prod. Res.* **31**, 1–16 (1995)
3. Pestka, J.J., Smolinski, A.T.: Deoxynivalenol: toxicology and potential effects on humans. *J. Toxicol. Environ. Health B Crit. Rev.* **8**, 39–69 (2005)
4. Turner, P.C., Flannery, B., Isitt, C., et al.: The role of biomarkers in evaluating human health concerns from fungal contaminants in food. *Nutr. Res. Rev.* **25**, 162–179 (2012)
5. Guengerich, F.P., Ueng, Y.F., Kim, B.R., et al.: Activation of toxic chemicals by cytochrome P450 enzymes: regio- and stereoselective oxidation of aflatoxin B1. *Adv. Exp. Med. Biol.* **387**, 7–15 (1996)
6. Gallagher, E.P., Kunze, K.L., Stapleton, P.L., et al.: The kinetics of aflatoxin B1 oxidation by human cDNA-expressed and human liver microsomal cytochromes P450 1A2 and 3A4. *Toxicol. Appl. Pharmacol.* **141**, 595–606 (1996)
7. Wild, C.P., Turner, P.C.: The toxicology of aflatoxins as a basis for public health decisions. *Mutagenesis* **17**(6), 471–481 (2002)
8. Kensler, T.W., Roebuck, B.D., Wogan, G.N., et al.: Aflatoxin: a 50 year odyssey of mechanistic and translational toxicology. *Toxicol. Sci.* **120**(Suppl 1), S28–S48 (2011)
9. IARC, Monograph 56.: Some naturally occurring substances: food items and constituents, heterocyclic aromatic amines and mycotoxins. International Agency for Research on Cancer, Lyon (1993)
10. Raney, V.M., Harris, T.M., Stone, M.P.: DNA conformation mediates aflatoxin B1-DNA binding and the formation of guanine N7 adducts by aflatoxin B1 8,9-exo-epoxide. *Chem. Res. Toxicol.* **6**, 64–68 (1993)
11. Wild, C.P., Turner, P.C.: Exposure biomarkers in chemoprevention studies of liver cancer. In: Millar, A.B., Bartsch, H., Boffetta, P., Dragsted, L., Vainio, H. (eds.) *Biomarkers in Cancer Chemoprevention*. IARC Scientific Pub. no. 154, pp. 215–222. (2001)
12. Groopman, J.D., Wild, C.P., Hasler, J., et al.: Molecular epidemiology of aflatoxin exposures: validation of aflatoxin-N7-guanine levels in urine as a biomarker in experimental rat models and humans. *Environ. Health Perspect.* **99**, 107–113 (1993)
13. Groopman, J.D., Hall, A., Whittle, H., et al.: Molecular dosimetry of aflatoxin-N7-guanine in human urine obtained in The Gambia, West Africa. *Cancer Epidemiol. Biomarkers Prev.* **1**, 221–228 (1992)
14. Groopman, J.D., Zhu, J.Q., Donahue, P.R., et al.: Molecular dosimetry of urinary aflatoxin-DNA adducts in people living in Guangxi Autonomous Region, People's Republic of China. *Cancer Res.* **52**, 45–52 (1992)
15. Zhu, J.Q., Zhang, L.S., Hu, X., et al.: Correlation of dietary aflatoxin B1 levels with excretion of aflatoxin M1 in human urine. *Cancer Res.* **47**, 1848–1852 (1987)
16. Sabbioni, G., Skipper, P.L., Büchi, G., et al.: Isolation and characterization of the major serum albumin adduct formed by aflatoxin B1 in vivo in rats. *Carcinogenesis* **8**, 819–824 (1987)
17. Sabbioni, G., Ambs, S., Wogan, G.N., et al.: The aflatoxin-lysine adduct quantified by high-performance liquid chromatography from human serum albumin samples. *Carcinogenesis* **11**, 2063–2066 (1990)
18. Gan, L.S., Skipper, P.L., Peng, X.C., et al.: Serum albumin adducts in the molecular epidemiology of aflatoxin carcinogenesis: correlation with aflatoxin B1 intake and urinary excretion of aflatoxin M1. *Carcinogenesis* **9**, 1323–1325 (1988)
19. Anwar, W.A., Khalil, M.M., Wild, C.P.: Micronuclei, chromosomal aberrations and aflatoxin-albumin adducts in experimental animals after exposure to aflatoxin B1. *Mutat. Res.* **322**, 61–67 (1994)

20. Wild, C.P., Hasegawa, R., Barraud, L., et al.: Aflatoxin-albumin adducts: a basis for comparative carcinogenesis between animals and humans. *Cancer Epidemiol. Biomarkers Prev.* **5**, 179–189 (1996)
21. Wild, C.P., Jiang, Y.Z., Sabbioni, G., et al.: Evaluation of methods for quantitation of aflatoxin-albumin adducts and their application to human exposure assessment. *Cancer Res.* **50**, 245–251 (1990)
22. Wild, C.P., Hudson, G.J., Sabbioni, G., et al.: Dietary intake of aflatoxins and the level of albumin-bound aflatoxin in peripheral blood in The Gambia. *West Afr. Cancer Epidemiol. Biomarkers Prev.* **1**, 229–234 (1992)
23. Wild, C.P., Jiang, Y.Z., Allen, S.J., et al.: Aflatoxin-albumin adducts in human sera from different regions of the world. *Carcinogenesis* **11**, 2271–2274 (1990)
24. Wild, C.P., Rasheed, F.N., Jawla, M.F., et al.: In utero exposure to aflatoxin in West Africa. *The Lancet* **337**, 1602 (1990)
25. Allen, S.J., Wild, C.P., Wheeler, J.G., et al.: Aflatoxin exposure, malaria and hepatitis B infection in rural Gambian children. *Trans. R. Soc. Trop. Med. Hyg.* **86**, 426–430 (1992)
26. Wild, C.P., Yin, F., Turner, P.C., et al.: Environmental and genetic determinants of aflatoxin-albumin adducts in the Gambia. *Int. J. Cancer* **86**, 1–7 (2000)
27. Gong, Y.Y., Cardwell, K., Hounsa, A., et al.: Dietary aflatoxin exposure and impaired growth in young children from Benin and Togo, West Africa: cross sectional study. *Br. Med. J.* **325**, 20–21 (2002)
28. Gong, Y.Y., Hounsa, A., Egal, S., et al.: Post-weaning exposure to aflatoxin results in impaired child growth: a longitudinal study in Benin, West Africa. *Environ. Health Perspect.* **112**, 1334–1338 (2004)
29. Shuaib, F.M., Jolly, P.E., Ehiri, J.E., et al.: Association between anemia and aflatoxin B1 biomarker levels among pregnant women in Kumasi, Ghana. *Am. J. Trop. Med. Hyg.* **83**, 1077–1083 (2010)
30. Turner, P.C., Sylla, A., Kuang, S.Y., et al.: Absence of TP53 codon 249 mutations in Guinean infants with high aflatoxin exposure. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2053–2055 (2005)
31. Turner, P.C., Collinson, A.C., Cheung, Y.B., et al.: Aflatoxin exposure in utero causes growth faltering in Gambian infants. *Int. J. Epidemiol.* **36**, 1119–1125 (2007)
32. Turner, P.C., Van Der Westhuizen, L., Nogueira Da Costa, A.: Biomarkers of exposure: mycotoxins – aflatoxin, deoxynivalenol and fumonisins. In: Knudsen, L.E., Merlo, D.F. *Biomarkers and Human Biomonitoring*. Royal Society of Chemistry, Cambridge (2011, in press)
33. Turner, P.C., Moore, S.E., Hall, A.J., et al.: Modification of immune function through exposure to dietary aflatoxin in Gambian children. *Environ. Health Perspect.* **111**, 217–220 (2003)
34. Turner, P.C., Mendy, M., Whittle, H., et al.: Hepatitis B infection and aflatoxin biomarker levels in Gambian children. *Trop. Med. Int. Health* **5**, 837–841 (2000)
35. Ahsan, H., Wang, L.Y., Chen, C.J., et al.: Variability in aflatoxin-albumin adduct levels and effects of hepatitis B and C virus infection and glutathione S-transferase M1 and T1 genotype. *Environ. Health Perspect.* **109**, 833–837 (2001)
36. Wang, P., Afriyie-Gyawu, E., Tang, Y., et al.: NovaSil clay intervention in Ghanaians at high risk for aflatoxicosis: II. Reduction in biomarkers of aflatoxin exposure in blood and urine. *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.* **25**, 622–634 (2008)
37. Wang, J.S., Qian, G.S., Zarba, A., et al.: Temporal patterns of aflatoxin-albumin adducts in hepatitis B surface antigen-positive and antigen-negative residents of Daxin, Qidong County, People's Republic of China. *Cancer Epidemiol. Biomarkers Prev.* **5**, 253–261 (1996)
38. Kensler, T.W., He, X., Otieno, M., et al.: Oltipraz chemoprevention trial in Qidong, People's Republic of China: modulation of serum aflatoxin albumin adduct biomarkers. *Cancer Epidemiol. Biomarkers Prev.* **7**, 127–134 (1988)
39. Turner, P.C., Loffredo, C., El-Kafrawy, S., et al.: A survey of aflatoxin-albumin adducts in sera from Egypt. *Food Addit. Contam.* **5**, 583–587 (2008)

40. Scussel, V.M., Haas, P., Gong, Y., et al.: Study of aflatoxin exposure in a Brazilian population using an aflatoxin-albumin biomarker. In: Njapau, H., Trujillo, S., van Egmond, H.P., Park, D.L. (eds.) *Mycotoxins and Phycotoxins: Advances in Determination, Toxicology and Exposure Management*, pp. 197–202. Wageningen Academic Publishers, Wageningen (2006)
41. Cupid, B.C., Lightfoot, T.J., Russell, D.: The formation of AFB(1)-macromolecular adducts in rats and humans at dietary levels of exposure. *Food Chem. Toxicol.* **42**(4), 559–569 (2004)
42. Johnson, N.M., Qian, G., Xu, L., et al.: Aflatoxin and PAH exposure biomarkers in a U.S. population with a high incidence of hepatocellular carcinoma. *Sci. Total Environ.* **408**, 6027–6031 (2010)
43. Merrill Jr., A.H., Wang, E., Vales, T.R., et al.: Fumonisin toxicity and sphingolipid biosynthesis. *Adv. Exp. Med. Biol.* **392**, 297–306 (1996)
44. Shephard, G.S., Thiel, P.G., Sydenham, E.W.: Initial studies on the toxicokinetics of fumonisin B1 in rats. *Food Chem. Toxicol.* **30**, 277–279 (1992)
45. Prelusky, D.B., Miller, J.D., Trenholm, H.L.: Disposition of 14C-derived residues in tissues of pigs fed radiolabelled fumonisin B1. *Food Addit. Contam.* **13**, 155–162 (1996)
46. Norred, W.P., Plattner, R.D., Chamberlain, W.J.: Distribution and excretion of 14C-fumonisin B1 in male Sprague–Dawley rats. *Nat. Toxins* **1**, 341–346 (1993)
47. Shephard, G.S., Thiel, P.G., Sydenham, E.W., et al.: Fate of a single dose of the 14C-labelled mycotoxin, fumonisin B1, in rats. *Toxicon* **30**, 768–770 (1992)
48. Shephard, G.S., Thiel, P.G., Sydenham, E.W., et al.: Distribution and excretion of a single dose of the mycotoxin fumonisin B1 in a non-human primate. *Toxicon* **32**, 735–741 (1994)
49. Dilkin, P., Direito, G., Simas, M.M., et al.: Toxicokinetics and toxicological effects of single oral dose of fumonisin B1 containing *Fusarium verticillioides* culture material in weaned piglets. *Chem. Biol. Interact.* **185**, 157–162 (2010)
50. Van der Westhuizen, L., Shephard, G.S., Van Schalkwyk, D.J.: The effect of repeated gavage doses of fumonisin B1 on the sphinganine and sphingosine levels in vervet monkeys. *Toxicon* **39**, 969–972 (2001)
51. Riley, R.T., An, N.H., Showker, J.L., et al.: Alteration of tissue and serum sphinganine to sphingosine ratio, an early biomarker for exposure to fumonisin-containing feeds in pigs. *Toxicol. Appl. Pharmacol.* **118**, 105–112 (1993)
52. Riley, R.T., Torres, O., Showker, J.L., Zitomer, N.C., Matute, J., Voss, K.A., Gelineau-van Waes, J., Maddox, J.R., Gregory, S.G., Ashley-Koch, A.E.: The kinetics of urinary fumonisin B1 excretion in humans consuming maize-based diets. *Mol Nutr Food Res.* **56**(9):1445–55 (2012)
53. Shephard, G.S., Van der Westhuizen, L., Sewram, V.: Biomarkers of exposure to fumonisin mycotoxins: a review. *Food Addit. Contam.* **24**, 1196–1201 (2007)
54. Turner, P.C., Nikiema, P., Wild, C.P.: Fumonisin contamination of food: progress in development of biomarkers to better assess human health risks. *Mutat. Res.* **443**, 81–93 (1999)
55. Merrill Jr., A.H., Sullards, M.C., Wang, E., et al.: Sphingolipid metabolism, roles in signal transduction and disruption by fumonisins. *Environ. Health Perspect.* **109**(Suppl 2), 283–289 (2001)
56. Gong, Y.Y., Torres-Sanchez, L., Lopez-Carrillo, L., et al.: Association between tortilla consumption and human urinary fumonisin B1 levels in a Mexican population. *Cancer Epidemiol. Biomarkers Prev.* **17**(3), 688–694 (2008)
57. Xu, L., Cai, Q., Tang, L., et al.: Evaluation of fumonisin biomarkers in a cross-sectional study with two high-risk populations in China. *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.* **27**, 1161–1169 (2010)
58. Zitomer, N.C., Mitchell, T., Voss, K.A., et al.: Ceramide synthase inhibition by fumonisin B1 causes accumulation of 1-deoxysphinganine: a novel category of bioactive 1-deoxysphingoid bases and 1-deoxydihydroceramides biosynthesized by mammalian cell lines and animals. *J. Biol. Chem.* **284**(8), 4786–4795 (2009)
59. Meko, F.A., Turner, P.C., Ashcroft, A.E., et al.: Development of a urinary biomarker of human exposure to deoxynivalenol. *Food Chem. Toxicol.* **41**, 265–273 (2003)

60. Turner, P.C., Burley, V.J., Rothwell, J.A., et al.: Dietary wheat reduction decreases the level of urinary deoxynivalenol in UK Adults. *J. Exp. Sci. Environ. Epidemiol.* **18**, 392–399 (2008)
61. Turner, P.C., Rothwell, J.A., White, K.L.M., et al.: Urinary deoxynivalenol is correlated with cereal intake in individuals from the United Kingdom. *Environ. Health Perspect.* **116**, 21–25 (2008)
62. Turner, P.C., White, K.L.M., Burley, V., et al.: A comparison of deoxynivalenol intake and urinary deoxynivalenol in UK adults. *Biomarkers* **15**, 553–562 (2010)
63. Turner, P.C., Burley, V.J., Rothwell, J.A., et al.: Deoxynivalenol: rationale for development and application of a urinary biomarker. *Food Addit. Contam.* **25**, 864–871 (2008)
64. Turner, P.C.: Deoxynivalenol and nivalenol occurrence and exposure assessment. *World Mycotoxin J.* **3**, 315–321 (2010)
65. Warth, B., Sulyok, M., Berthiller, F., et al.: Direct quantification of deoxynivalenol glucuronide in human urine as biomarker of exposure to the *Fusarium* mycotoxin deoxynivalenol. *Anal. Bioanal. Chem.* **401**, 195–200 (2011)
66. IARC: Some traditional herbal medicines, some mycotoxins, naphthalene and styrene. *IARC Monogr. Eval. Carcinog. Risks Hum.* **82**, 1–556 (2002)
67. Black, R.E., Morris, S.S., Bryce, J.: Where and why are 10 million children dying every year? *Lancet* **361**, 2226–2234 (2003)
68. Campbell, D.I., Elia, M., Lunn, P.G.: Growth faltering in rural Gambian infants is associated with impaired small intestinal barrier function, leading to endo toxemia and systemic inflammation. *J. Nutr.* **133**, 1332–1338 (2003)
69. Lunn, P.G., Northro-Clewes, C.A., Downes, R.M.: Intestinal permeability, mucosal injury, and growth faltering in Gambian infants. *Lancet* **338**, 907–910 (1991)
70. Prentice, A.: Nutrient requirements for growth, pregnancy and lactation: the Keneba experience. *S. Afr. J. Clin. Nutr.* **6**, 33–38 (1993)
71. Gong, Y.Y., Egal, S., Hounsa, A., et al.: Determinants of aflatoxin exposure in young children from Benin and Togo, West Africa: the critical role of weaning. *Int. J. Epidemiol.* **32**(4), 556–562 (2003)
72. Williams, J.H., Phillips, T.D., Jolly, P.E., et al.: Human aflatoxicosis in developing countries: a review of toxicology, exposure, potential health consequences, and interventions. *Am. J. Clin. Nutr.* **80**, 1106–1122 (2005)
73. Bouhet, S., Oswald, I.P.: The intestine as a possible target for fumonisin toxicity. *Mol. Nutr. Food Res.* **51**, 925–931 (2007)
74. Gratz, S., Wu, Q.K., El-Nezami, H.: Alteration of intestinal transport, metabolism and toxicity of aflatoxin B1 by *Lactobacillus rhamnosus* strain GG in vitro in Caco-2 cells. *Appl. Environ. Microbiol.* **73**, 3958–3964 (2007)
75. Cullen, J.M., Newberne, P.M.: Acute hepatotoxicity of aflatoxins. In: Eaton, D.L., Groopman, J.D. (eds.) *The Toxicology of Aflatoxins: Human Health, Veterinary and Agricultural Significance*, Academic Press, Inc, San Diego, CA, pp. 3–26 (1994)
76. Lunn, P.G.: The impact of infection and nutrition on gut function and growth in childhood. *Proc. Nutr. Soc.* **59**(1), 147–154 (2000)
77. Campbell, D.I., Lunn, P.G., Elia, M.: Age-related association of small intestinal mucosal enteropathy with nutritional status in rural Gambian children. *Br. J. Nutr.* **88**(5), 499–505 (2002)
78. Campbell, D.I., McPhail, G., Lunn, P.G., et al.: Intestinal inflammation measured by fecal neopterin in Gambian children with enteropathy: association with growth failure, *Giardia lamblia*, and intestinal permeability. *J. Pediatr. Gastroenterol. Nutr.* **39**(2), 153–157 (2004)
79. Jiang, Y., Jolly, P.E., Ellis, W.O.: Aflatoxin B1 albumin adduct levels and cellular immune status in Ghanaians. *Int. Immunol.* **17**(6), 807–814 (2005)
80. Plum, L.M., Rink, L., Haase, H.: The essential toxin: impact of zinc on human health. *Int. J. Environ. Res. Public Health* **7**, 342–365 (2010)
81. Imdad, A., Bhutta, Z.A.: Effect of preventive zinc supplementation on linear growth in children under 5 years of age in developing countries: a meta-analysis of studies for input to the lives saved tool. *BMC Public Health* **11**(Suppl 3), S22 (2011)

82. Mocchegiani, E., Corradi, A., Santarelli, L., et al.: Zinc, thymic endocrine activity and mitogen responsiveness (PHA) in piglets exposed to maternal aflatoxicosis B1 and G1. *Vet. Immunol. Immunopathol.* **62**, 245–260 (1998)
83. Yarru, L.P., Settivari, R.S., Antoniou, E., et al.: Toxicological and gene expression analysis of the impact of aflatoxin B1 on hepatic function of male broiler chicks. *Poult. Sci.* **88**, 360–371 (2009)
84. Zamboni, G., Dufillot, D., Antoniazzi, F., et al.: Growth hormone-binding proteins and insulin-like growth factor-binding proteins in protein-energy malnutrition, before and after nutritional rehabilitation. *Pediatr. Res.* **39**, 410–414 (1996)

Pharmaceutical Risk Assessment and Predictive Enrichment to Maximize Benefit and Minimize Risk: Issues in Product Life Cycle Evaluation

Robert T. O'Neill

Abstract Use of pharmaceutical products provides many benefits to patients but there is also a need to assess the risks associated with the benefits afforded by these exposures. This paper focuses on FDA's response to a variety of safety issues in the pharmaceutical arena that have received broad public attention in the last few years and on the extremely important role that statistics is playing and will play in shaping the nation's future systematic approach to addressing the many facets of medical product quantitative risk assessment and management. The emphasis will be on the life cycle evaluation of risk which includes the pre-market assessment of risk and the post-market or post-approval period of medical products as they are used by populations of all ages and ethnicities, alone and in combination, for short and for very long periods of time. Additionally, with the advent of the genomic revolution, the interest in personalized medicine and the search for predictive biomarkers to aid patient selection and treatment strategies has opened up many methodological challenges for new prediction strategies. Clinical trial study designs to evaluate targeted therapy and to enrich study populations are a major focus of this area. This paper touches on these areas from the perspective of pharmaceutical development and evaluation.

Disclaimer: This paper represents the views of the author and not necessarily those of the Food and Drug Administration.

R.T. O'Neill (✉)

Office of Translational Sciences, Center for Drug Evaluation and Research,
10903 New Hampshire Avenue, Silver Spring, MD 20993, USA
e-mail: Robert.Oneill@fda.hhs.gov

Introduction

The goal of this article is to provide a perspective on the role of statistical prediction in several aspects of pharmaceutical development, product evaluation and regulatory decision making. In particular, the benefits as well as the risks associated with pharmaceutical products are now evaluated across a product life cycle continuum, and prediction of which patients are likely to benefit with minimum risk is increasingly important. The continuum includes the time during which evidence for the efficacy and safety of a pharmaceutical is developed to support its approval and marketing, sometimes called the pre-approval or pre-marketing time; and the time when the pharmaceutical is on the market, called post-marketing time, during which the evidence for its safety as well as its efficacy is further developed, often encompassing its expanded use for many different diseases, for different indications or claims, and for different subpopulations of patients. The benefits and the risks observed across this broad spectrum of patient conditions of use may very well be different depending upon the patient subpopulation characteristics, thus setting the stage for the importance of understanding predictive factors associated with differential treatment response.

There are many new demands and challenges for the modern pharmaceutical development and regulatory evaluation process, and we will describe those which have an impact on risk evaluation and prediction and on the identification of factors that maximize the benefits and minimize the risk of pharmaceuticals. This latter effort emphasizes the individual patient predictive and prognostic factors that might allow such a goal to be achieved. We will focus on three major themes in the current regulatory/public health arena that are associated with prediction and risk: the first is that of personalized medicine or targeted therapy for individual patients, where the use of biomarker information to predict who might respond to therapy is being investigated; the second is risk assessment and prediction in the evaluation of the safety of pharmaceuticals, an area receiving increased public attention and scrutiny, especially when maximizing benefit and minimizing risk is a goal; and finally the role of prediction in assuring there is reproducibility and replication of research results that can robustly support the safety and efficacy decisions made during the product life cycle evaluation.

In section “[Personalized Medicine or Targeted Therapy](#)”, we describe what we mean by personalized or individualized medicine. In section “[Statistical Approaches to Modeling Prognostic Factors](#)”, we discuss the difference between a prognostic and a predictive classifier as it relates to demonstration of differential treatment response, providing a brief summary of the considerable statistical history in evaluating prognostic factors. In section “[Some Clinical Trial Designs or Strategies Proposed to Evaluate Predictive Treatment Effects: Sometimes Called Enrichment or Targeted Therapy Designs](#)”, we discuss some of the recent clinical trial strategies that have been developed to demonstrate that a marker is predictive of a treatment effect or an enhanced treatment effect. In section “[Predictive Clinical Trial Approaches Helping to Manage Risk of New Drugs by Identifying and Screening](#)

Patients” we focus on how predictive markers are being used to minimize exposure of patients to serious life threatening adverse reactions, reinforcing some of the major messages in a recent Institute of Medicines report [1] on the future of drug safety in the United States in which the report urged FDA to enhance the nation’s ability to manage the safety of new drugs. Section “[Predicting Whether Research Results Can Be Replicated: Concern for Reproducible Study Findings – A Theme Connecting Predictions to True Findings](#)” considers emerging controversies on the importance of replication of research findings and the limitations of statistical methods to assure that occurs. Section “[Two Examples of a Framework to Evaluate the Performance Characteristics of Replicable Research Findings](#)” provides two examples where a consortium of scientists was convened to establish principles and practices to evaluate the performance of predictive models for microarray based prediction of clinical and pre-clinical outcomes and to propose best practices that might improve replicability. We conclude in section “[Concluding Remarks](#)” with summary remarks.

Personalized Medicine or Targeted Therapy

With the advent of the genomic revolution and the advances in understanding the differential pathways to disease progression and the use of biomarkers or other classifiers to evaluate the magnitude of patient level response to targeted therapies, there has been increased interest in and demands on clinical trials to provide the evidence to support decisions for therapies that are more tailored to and specific to a patient’s likelihood to respond to a treatment. Concurrent with this interest has been the interest in the development and evaluation of biomarkers of treatment response, especially in the area of oncology [2, 3] and for the use of various patient enrichment strategies in the design of clinical trials. A way to portray personalized medicine in the context of the selection of a pharmaceutical for a given patient is to consider the following situations. If a patient cannot metabolize a drug because his or her genetic makeup lacks a gene to do so, despite taking the drug, they will not experience its intended pharmacological beneficial effect, yet may share the risk of side effects because of such exposure. If a patient’s genetic makeup is such that they are either a slow, intermediate or fast metabolizer of a drug, that patient may need a different dose of the drug to experience a comparable beneficial effect or to reduce the chances of a serious side effect due to overdosing. If in any given patient, the organ specific target of a drug is resistant or non-responsive to the therapy, then the intended therapeutic effect is neutralized or minimized, as in some cancers (e.g. breast, colorectal). So, if there is an identifiable marker, and a patient possesses that marker, the goal is to demonstrate that such a patient should expect a better treatment response in contrast to a patient without the marker. Thus, the prediction of which patients will benefit from any drug may depend upon their genetic make up or whether they possess certain markers that have been identified as predictive of treatment response.

The establishment of a marker that can be used either to select patients for treatment, or perhaps to evaluate a particular treatment in a controlled clinical trial, depends upon the type of evidence needed for that marker's ability to predict an outcome. There is a confusion in the literature as well as in general practice regarding the use of terms to describe prediction of an outcome, or prediction of a treatment effect. Within the regulatory area of biomarker identification and qualification, two terms are used to distinguish between markers that are prognostic of clinical outcome, and markers that are predictive of treatment effect or optimal treatment effect. It is possible that a marker can be both prognostic and predictive at the same time but in order to demonstrate that a marker meets the criteria for each one of these marker types requires certain study designs, usually with prospectively planned criteria. It is important to understand the differences in use of these terms because the evidence needed to support their use and to qualify them as being prognostic or predictive is different.

We define these marker or biomarker terms as follows. A *prognostic* biomarker is a baseline patient or disease characteristic that categorizes patients by degree of risk for disease occurrence or progression. A prognostic biomarker informs about the natural history of the disorder in that particular patient in the absence of a therapeutic intervention. A *predictive* biomarker is a baseline characteristic that categorizes patients by their likelihood for response to a particular treatment. A predictive biomarker is used to identify whether a given patient is likely to respond to a treatment intervention in a particular way. The marker may predict a favorable response or an unfavorable response (i.e., adverse event). In general, characterizing a biomarker as predictive of treatment effect requires a randomized trial with a control group in which the subpopulations identified by the marker can be evaluated for the marker specific treatment effects, and for the relationship of the marker to a clinical outcome in the control group only, thereby separating the prognostic nature of the marker to 'predict' a clinical outcome regardless of treatment, and the predictive nature of the marker to 'predict' a treatment effect that is better in those with the marker than without the marker. In a sense, the former is a single cohort problem, and the latter is a two sample comparative problem.

Prediction of treatment effect on an individual patient level that yields a yes/no answer is a more challenging problem than prediction of treatment effect for an identifiable patient subpopulation. A diagnostic agent that is used to give a yes/no answer as to whether an individual does or does not have a disease is held to a different standard of evidence than is a probabilistic prediction based upon a model. The predictive concept being used in this paper is more probabilistic in nature, essentially quantifying the probability of treatment response and or treatment effect being greater in a subject with the marker in contrast to a subject without the marker. When a diagnostic agent (e.g. PSA screen) is the marker that classifies a subject as eligible for treatment and more likely to benefit, and it is important that that diagnosis be correct, it is then necessary to characterize the performance characteristics of the marker in terms of its misclassification rates, its sensitivity and specificity and the positive predictive value of a patient positively responding to treatment. This subject matter become complex and is not the purpose of this

article, nor can we discuss the issues fully here. However, the efficiency of a clinical trial that employs a patient enrichment strategy using a marker that is poorly characterized is likely to be negatively impacted because of the misclassification, and not likely to be a useful approach.

Statistical Approaches to Modeling Prognostic Factors

Over the last 30–40 years, there is an extensive statistical literature developed for modeling and methods to identify factors associated with an individual's probability of having a clinical outcome, defined here as prognostic factors. This literature has been developed within the context of long-term single cohort follow-up studies and within the context of randomized clinical trials that evaluate differential subgroup responses. Armitage and Gehan [4] in 1974 proposed statistical methods for the identification and use of prognostic factors; in 1980 Byar and Green [5] and Byar and Corle [6] proposed methods to choose treatment for cancer patients based on covariate information, an approach that might be considered predictive, but which is not how the problem is being dealt with today. The Cox proportional hazard model substantially changed how clinical trials of time to event outcomes are designed and analyzed and how covariates and their interactions with treatment assignment are evaluated. In 1989, Gail et al. [7] proposed a model for projecting individualized probabilities of developing breast cancer for white females who are being examined annually, an approach that was eventually used as entrance criteria for clinical trials and eligibility for treatment. In 2003, Pepe et al. [8] considered the limitations of the odds ratio metric in gauging the performance of a diagnostic, prognostic, or screening marker. In 2006, Ware [9] reinforced this concept by discussing the limitations of risk factors as prognostic tools in a comment on Wang et al. [10] who considered whether multiple biomarkers for the prediction (note: term prognostic should have been used) of first major cardiovascular events and death might be an improvement over single biomarkers. For the most part, most all of these approaches and strategies relate a patient's outcomes to a function of that patient's covariates, usually at a baseline untreated status, or at initiation of patient follow-up as in the Framingham study, and are prognostic in the sense we have defined them.

Treatment by covariate interactions that are intended to evaluate the statistical differences in the magnitude of treatment effects as a function of patient baseline covariates, could be considered predictive factors in the sense we have defined them here. In fact statistical models that include covariates and their interaction with treatment are frequently used to evaluate differential treatment effects in clinical trials. But the consideration of treatment by baseline patient covariates interactions alone (i.e., differential treatment effect that is sometimes called effect modification in epidemiology) is not sufficient for the current purpose, though as an exploratory approach it may be useful. The recent interest in evaluating whether a marker is predictive of better treatment effects focuses on evaluating treatment effects in patient subgroups identified by a marker that is pre-specified and that has

a type 1 error allocated to it at the study planning stage. This approach is used in order to control for the multiplicity of treatment comparison hypotheses created by the several marker subgroup evaluations, an approach that has not been routinely considered in the past literature on evaluation of treatment by covariate interactions.

Some Clinical Trial Designs or Strategies Proposed to Evaluate Predictive Treatment Effects: Sometimes Called Enrichment or Targeted Therapy Designs

Clinical trial study designs now exist that can prospectively address the marker predictive hypothesis, but the statistical focus seems to have changed from a regression and interaction evaluation approach to a type 1 error spending hypothesis testing approach which, in some versions, may have a gate-keeper or a sequential testing order pre-specified. In the past, the general strategy in clinical trials was to plan for an overall treatment effect size in the entire trial population, and then if that result was statistically positive and persuasive to proceed to examine pre-specified subgroups. Such examination might consist of evaluation of the marker negative or positive subgroups separately, to examine evidence for equal or comparable treatment effects in each subgroup, or to evaluate whether some subgroups have greater treatment effects than others, assuming all subgroups shared a common minimum treatment effect size. Generally, if no statistically significant treatment effect was observed for the primary endpoint (hypothesis) no further hypothesis testing of a confirmatory nature would be entertained, and any other analyses would be considered exploratory and hypothesis generating. Some of the newer designs for marker subgroups change the goal from an overall statistical test in the all comers population to a subset of patients identified by the marker, or to either possibility being acceptable.

Simon [11–15] and his co-authors proposed several approaches to identify and test for a marker or classifier to target subgroups more likely to respond to treatment. Mandreker and Sargent [16] have provided a useful overview of the many of the currently proposed designs including a discussion of the challenges faced in deciding which design to choose from. As the challenges to such designs include dealing at the protocol sample size stage with the unknown population prevalence of the marker and a lack of any real information on the anticipated treatment effect in the marker subgroups, adaptive designs have been proposed as a way to accommodate what is learned during the course of the trial [17, 18]. It is not the intention in this article to exhaustively catalogue the different available study design choice and strategies, but rather to describe some of the current proposals which have yet to have much experience associated with them.

The key challenges in the choice of these designs are around the decision to include all comers or only a selected marker positive subgroup at the protocol stage, or to adapt later on in the trial to other information on response to treatment in the

various marker groups. A particular concern is what has been called a retrospective versus prospective approach to subgroup identification and analysis of differential treatment effects within those subgroups. The concept of retrospective evaluation comes about as a result of a strategy to assess baseline biomarker status on all subjects randomized into a trial, but with no prospective plan for which markers are associated with benefit. One then uses a post randomization treatment analysis strategy to evaluate the treatment differential effects according to marker status, the knowledge for which did not exist prior to conducting the trial. A variation of this strategy that has been used in some studies, but not recommended, is to only have a subset of the initial full randomized study population assessed for the marker, possibly a convenience sample. Recently, the evaluation of two cancer therapies for metastatic colorectal cancer and the relationship of treatment benefit to KRAS status, a marker for response to treatment, was evaluated in this retrospective manner [2]. If the performance of a predictive marker to correctly identify the marker positive and negative subgroups is of equal importance, then co-development of both the classifier and the drug simultaneously in the same clinical trial is a more complex challenging task.

An interesting clinical trial design of the anti-viral drug Abacavir, called 'PREDICT 1' [19] evaluated whether individualized screening of patients prior to therapy would successfully reduce the incidence of a very serious life threatening hypersensitivity adverse reaction that made the drug not one of first choice. This successful trial randomized all subjects entered to the active marketed drug Abacavir, but the randomized arms differed by whether the screening strategy for the hypersensitivity gene was used, the idea being that patients screened and found positive would be excluded from treatment. This trial of 1956 patients from 19 countries successfully confirmed that screening will reduce severe hypersensitivity reactions, and at the same time provided estimates of the sensitivity, specificity and positive predictive value of the screening diagnostic. This study design confirmed the predictive nature of a screening tool to manage the safety profile of a drug that otherwise would not have been used as widely in the anti-viral area, thereby demonstrating the value to predictive clinical trial approaches.

Predictive Clinical Trial Approaches Helping to Manage Risk of New Drugs by Identifying and Screening Patients

The 'PREDICT 1' Abacavir trial described above is an example of how predictive tools can benefit the optimal use of new pharmaceutical therapies that have potentially great benefit but which may also have high toxicity that goes along with its use and exposure, and thereby limits its use. The recent interest by the Institute of Medicine [1] in improving the approach to evaluating and managing the safety risks of new pharmaceuticals highlights the potential value of predictive tools and innovative study designs to maximize benefit and minimize risk to patients. The

KRAS story [2] described above also illustrates the interest in minimizing exposure of patients to potent anti-cancer therapies when it appears that patients will not benefit yet will still be subject to the adverse effects of such therapies. As targeted therapies are increasingly developed with this philosophy in mind the importance of good statistical predictive tools and diagnostic classifiers with good predictive capabilities will become more evident.

Predicting Whether Research Results Can Be Replicated: Concern for Reproducible Study Findings – A Theme Connecting Predictions to True Findings

Recently, there have been several publications [20–22] that have challenged whether published research, medical or otherwise, is credible in the sense that it can be repeated and demonstrate reproducible findings. Statistics has long had a role in assuring that study conclusions are not likely due to chance, but the recent interest goes beyond that concern into the area of the statistics of replication which has received less attention, yet to this author seems a fertile part of statistical prediction. Ioannidis [23] in a highly cited article on the topic of ‘Why most published research findings are false’ described a framework for evaluating whether true positive and true negative research findings can be distinguished and whether the statistical framework of hypothesis testing from a frequentist perspective can help one understand the limitations of the predictability properties of published research. The subject matter areas that have received attention in this regard are SNP evaluations in genome wide association studies [23] and observational studies [24, 25], where concern has been raised that research results are not replicable and thus the credibility of conclusions based upon such research might be in question. It is not a new concept, as Ottenbacher [26], Lee and Zelen [27] and Goodman [28–30] raised these concerns within the context of frequentist and Bayesian testing of hypotheses and interpretation of whether a statistically significant study finding in a single study can inform about whether that finding is a true positive. More recently, some editors [20, 21] have criticized the use of frequentist based statistics for fostering false research, although qualified statisticians have responded and made the case that it is the misuse and misunderstanding of the statistical approaches that cause the confusion. To remedy the overemphasis of p-values research findings and to promote a better understanding of its limitations, several authors [31–33] have proposed new approaches to provide a more realistic way to interpret the predictiveness of p-values, were one interested in considering the p-value as predictive of future successful studies of the same type.

Two Examples of a Framework to Evaluate the Performance Characteristics of Replicable Research Findings

The MicroArray Quality Control II Study (MACQ-II) [34] is a consortium that was conceived with the primary goal of examining various aspects of model development practices for generating binary classifiers in pre-clinical and clinical data sets based upon gene expression data from microarrays. The motivation for this effort was the recognition that much microarray research which purports to predict outcomes based upon gene expression links is not replicable and especially predictive models that purport to functionally express the predictive nature of microarray expression signatures. To set some standards in place, a mechanism was set up to archive all appropriate data sets that link gene expression and certain clinical and pre-clinical outcomes, and then 36 independent teams analyzed six microarray data sets where the analysis teams were asked to submit models from two stages of analysis. For the first stage, each team built prediction models for up to 13 different coded endpoints using six training data sets. Models were then tested on blinded validation data sets not available to the analysis teams during training. Teams were then allowed to repeat the model building and validation process by using the training models on the original training data set. Extensive performance characteristics of the teams, the models, and the data sets were published [34] that provided metrics on the predictability of the various models. This extensive collaborative effort of government, academic and industry scientists was intended to improve the infrastructure for assuring that prediction models for clinical and preclinical outcomes based upon microarray profiles were in some sense performed as reported and claimed and could be considered credible with a reasonable chance of replication. The conclusions drawn from this study with regard to best practices should be useful to readers, but it is not the goal of this paper to go into detail about them.

The second example is an extensive collaborative project with similar goals, namely to evaluate the performance characteristics of a variety of signaling methods intended to identify drug adverse event associations in electronic health care records and medical claims records. The Observational Medical Outcomes Partnership (OMOP) [35] was established to inform the appropriate use of observational healthcare data bases for active surveillance of adverse reaction/drug exposure pairs. This project, information for which is available at the website www.omop.fnih is a collaboration of FDA, NIH, the pharmaceutical industry and academia in order to better understand the reproducibility characteristics of different signaling methods to inform the large national effort of the Sentinel project mandated by the congress. Many different statistical signaling methods are available and they are applied to different data bases for surveillance signaling purposes. Using a common data model, the OMOP characterized the performance characteristics of different available methods by using simulated and real data sets and candidate signal methods and produced metrics of performance, such as sensitivity and specificity and positive predictive value of different signaling methods. The evaluation considered different

circumstances of the data models in an attempt to characterize true positive and true negative signals from each other, a systematic attempt that is currently undergoing review.

One interesting example provided by OMOP was that of two studies, one published in the *Journal of the American Medical Association* [36] and the other in the *British Medical Journal* [37] on the topic of oral bisphosphonates and the risk of esophageal cancer. Each published study addressed the same question, used different observational study designs but used virtually the same health claims data base with patient records over approximately the same chronological time frame. Each study came to different conclusions and so the question of why they did so was of key interest as well as the additional concern of lack of reproducibility of observational research that it raised. Of course even at times randomized clinical trials do not confirm results. But the topic of replicable research is receiving considerable scrutiny especially within the area of comparative effectiveness research (CER) that uses observational data.

Concluding Remarks

The goal of this article is to tie together three themes that relate to risk prediction and regulatory decision making. The first theme is the role of prediction in individualized or personalized medicine and drug development, especially using biomarkers as classifiers to help attain that goal. The second is the development and use of predictive biomarkers to support study enrichment strategies and to manage the safety and risk profiles of new drugs. And the third is the role of prediction in characterizing whether research findings are replicable, an area of recent concern, especially in observational research. Statistical concepts are obviously critical to linking these themes together, and it is the contributions of statistical thinking that has helped clarify the strengths and limitations of prediction in these areas.

References

1. Psaty, B.M., Burke, S.P.: Protecting the health of the public-Institute of Medicine recommendations on drug safety. *N. Engl. J. Med.* **355**, 1753–1755 (2006)
2. Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., et al.: K-ras Mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* **359**, 1757–1765 (2008)
3. De Roock, W., Claes, B., Bernasconi, D., et al.: Effects of *KRAS*, *BRAF*, *NRAS*, and *PIK3CA* mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol.* **11**(8), 753–762 (2010)
4. Armitage, P., Gehan, E.A.: Statistical methods for the identification and use of prognostic factors. *Int. J. Cancer* **13**, 16–36 (1974)
5. Byar, D.P., Green, S.B.: The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bull. Cancer (Paris)* **67**(4), 477–490 (1980)

6. Byar, D.P., Corle, D.K.: Selecting optimal treatment in clinical trials using covariate information. *J. Chronic Dis.* **30**(9), 445–459 (1977)
7. Gail, M., Byar, D.P., et al.: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**(24), 1879–1886 (1989)
8. Pepe, M.S., Janes, H., et al.: Limitation of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* **159**(9), 882–890 (2004)
9. Ware, J.H.: The limitations of risk factors as prognostic tools. *N. Engl. J. Med.* **355**(25), 2615–2617 (2006)
10. Wang, T.J.: Multiple biomarkers for the prediction of first major cardiovascular events and death. *N. Engl. J. Med.* **355**(25), 2361–2639 (2006)
11. Simon, R., Maitournam, A.: Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.* **10**, 6759–6763 (2004)
12. Maitournam, A., Simon, R.: On the efficiency of targeted clinical trials. *Stat. Med.* **24**, 329–339 (2005)
13. Freidlin, B., Simon, R.: Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin. Cancer Res.* **11**, 7872–7878 (2005)
14. Freidlin, B., Jiang, W., Simon, R.: The cross-validated adaptive signature design. *Clin. Cancer Res.* **16**(2), 691–698 (2010)
15. Jiang, W., Freidlin, B., Simon, R.: Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J. Natl. Cancer Inst.* **99**, 1036–1043 (2007)
16. Mandrekar, S.J., Sargent, D.J.: Clinical trial designs for predictive biomarker validation: theoretical consideration and practical challenges. *J. Clin. Oncol.* **27**, 4027–4034 (2009)
17. Wang, S.J., O'Neill, R.T., Hung, H.M.J.: Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. *Clin. Trials* **7**, 525–536 (2010)
18. Wang, S.J., O'Neill, R.T., Hung, H.M.J.: Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Stat.* **6**, 227–244 (2007)
19. Mallal, S., Phillips, E., Carosi, G., et al.: HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**, 568–579 (2008)
20. Siegfried, T.: Odds are it's wrong. *Sci. News* 177, #10 (2010); see letter in response by Pantula, Teugels, Stefanski, letters for May 8, 2010.
21. Lash, T.L., Vandenbroucke, J.P.: Should preregistration of epidemiologic study protocols become compulsory. *Epidemiology* **23**(2), 184–188 (2012)
22. Ioannidis, J.P.A.: Why most published research findings are false. *Plos Med.* **2**(8), e124 (2005)
23. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., Rothman, N.: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, 434–442 (2004)
24. Shafer, S.L., Dexter, F.: Publication bias, retrospective bias, and reproducibility of significant results in observational studies. *Anesth. Analg.* **114**(5), 931–932 (2012)
25. Greenberg, R.S., Bembea, M., Heitmiller, E.: Rainy days for the Society of Pediatric Anesthesia. *Anesth. Analg.* **114**, 1102–1103 (2012)
26. Ottenbacher, K.J.: The power of replications and replications of power. *Am. Stat.* **50**(3), 271–275 (1996)
27. Lee, S.J., Zelen, M.: Clinical trials and sample size considerations: another perspective. *Stat. Sci.* **15**(2), 95–110 (2000)
28. Goodman, S.N.: A comment on replication, p-values and evidence. *Stat. Med.* **11**, 875–879 (1992)
29. Goodman, S.N.: Toward evidence-based medical statistics. 1. The P value fallacy. *Ann. Intern. Med.* **130**, 995–1004 (1999)
30. Goodman, S.N.: Toward evidence-based medical statistics. 2. The Bayes factor. *Ann. Intern. Med.* **130**, 1005–1013 (1999)

31. Cumming, G.: Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect. Psychol. Sci.* **3**, 286–300 (2008)
32. Boos, D.D., Stefanski, L.A.: P-value precision and reproducibility. *Am. Stat.* **65**(4), 213–221 (2011)
33. Madigan, D.: Beyond p-values: computing the probability of a true association. Presentation at the Observational Medical Outcomes Partnership. July, 2012. Available at www.omop.fnih.org
34. The MAQC Consortium: The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray based predictive models. *Nat. Biotechnol.* **28**(8), 827–838 (2010)
35. The Observational Medical Outcomes Partnership (OMOP); www.omop.fnih.org; 2011 and 2012 Symposium Presentations
36. Cardwell, C.R., Abnet, C.C., Cantwell, M.M., Murray, L.J.: Exposure to oral bisphosphonates and risk of esophageal cancer. *J. Am. Med. Assoc.* **304**(6), 657–663 (2010)
37. Green, J., Czanner, G., Reeves, G., Watson, J., Wise, L., Beral, V.: Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *Br. Med. J.* **341**, 1–8 (2010)

A Multiple Imputation Approach for the Evaluation of Surrogate Markers in the Principal Stratification Causal Inference Framework

Xiaopeng Miao, Xiaoming Li, Peter B. Gilbert, and Ivan S.F. Chan

Abstract The concept of principal surrogate developed in the causal inference framework (Frangakis and Rubin (Biometrics 58:21–29, 2002); Gilbert and Hudgens (Biometrics 64:1146–1154, 2008)) has drawn much attention in the field of biomarker research. Principal surrogates are defined based on the causal treatment effects in principal strata, which are constructed based on the joint distribution of the potential surrogate markers when a patient receives either the placebo or the treatment. The challenge of evaluating principal surrogates lies in the fact that half of these potential surrogate markers cannot be observed in most clinical trials. Therefore assessing the principal surrogacy of biomarkers is essentially a missing data problem. In this article, we propose a multiple imputation approach to evaluate candidate principal surrogate markers. The proposed method employs baseline variables to impute the missing potential surrogate markers. The stratum-specific causal treatment effects on the clinical endpoint are then estimated for each imputed dataset and the inference for surrogacy of a biomarker is based on the

X. Miao (✉)

Data and Statistical Sciences, AbbVie, North Chicago, IL 60064, USA

e-mail: xiaopeng.miao@abbvie.com

X. Li

Biostatistics, Gilead Sciences, Seattle, WA 98102, USA

e-mail: xiaoming.li@gilead.com

P.B. Gilbert

Fred Hutchinson Cancer Research Center and Department of Biostatistics,
University of Washington, Seattle, WA 98109, USA

e-mail: pgilbert@scharp.org

I.S.F. Chan

Late Development Statistics, Merck Research Laboratories, North Wales, PA 19454, USA

e-mail: ivan_chan@merck.com

combined results over multiple imputations. Simulation studies are performed to evaluate the performance of the proposed method and the implementation of the method is illustrated using a vaccine study.

Introduction

The enormous benefits of substituting the clinical endpoints with surrogate markers (also referred to as “surrogate endpoints” or “surrogate outcomes”) that can be measured before the realization of the clinical endpoints have led to increasing efforts in searching for surrogate markers. Assessing the surrogate value of biomarkers would not only help identify more cost-effective outcome measurements for future clinical trials, but also shed light on the mechanism through which the treatment works. For example, in vaccine research, an important goal is to understand the causative role of vaccine-induced immune responses in reducing the risk of diseases (Halloran [26]). Surrogate markers (sometimes called “correlates of protection” in vaccine research) are extremely useful in appraising the consistency of vaccine manufacturing and in predicting long-term effectiveness of vaccines [1].

In the past two decades, there has been extensive literature on statistical methods for surrogacy evaluation. Joffe and Greene [2] reviewed the strengths and limitations of four major mathematical frameworks for evaluating surrogate markers, based on conditional independence (also known as the Prentice framework), direct and indirect effects, meta-analysis and principal stratification. Here we focus on the principal stratification framework, using the concept that a good surrogate marker is a biomarker that provides accurate prediction of the treatment effect on the clinical endpoint based on the treatment effect on the biomarker. Such a predictive surrogate may or may not be a mechanistic cause of the clinical treatment effect, and may or may not mediate the clinical treatment effect. Assessment of mediation is a separate concept that may be best addressed using natural indirect effect estimands [3–5], which are not considered here. The relationship between the principal stratification framework and other major mathematical frameworks are discussed in Joffe and Greene [2]. Validity and implementation of the Prentice framework requires that all subject characteristics predictive of both the surrogate marker and clinical endpoint are correctly controlled for in the regression model used to assess conditional independence, and that there is a large degree of overlap in the support of the surrogate marker distribution across the treatment groups. In contrast, the principal stratification framework can be used without requiring either of these conditions; however, it faces a different challenge for validity, that the causal estimands of interest are only partially identified, necessitating the use of additional identifiability assumptions.

The novel “principal surrogate” definition was developed in the principal stratification framework by Frangakis and Rubin (henceforth FR) [6]. FR proposed constructing principal strata based on the joint distribution of potential surrogate markers and assessing the surrogate value of biomarkers via the stratum-specific

causal treatment effects. Following this approach, several methods have been proposed for the evaluation of candidate principal surrogate markers. For example, Gilbert and Hudgens (henceforth GH) [7] proposed to evaluate principal surrogate markers via the causal effect predictiveness surface for binary clinical endpoints. Qin et al. [8] employed Follmann's [9] augmented trial designs and likelihood-based approaches to validate principal surrogates for discrete failure time endpoints. Li, Taylor and Elliott [10] developed a Bayesian approach to evaluate a binary surrogate for a binary clinical endpoint. Huang and Gilbert [11] proposed a graphical approach to compare the principal surrogate values of different continuous biomarkers for binary clinical endpoints. While these methods provide feasible approaches to assess the principal surrogate value of biomarkers, their applications are limited to certain types of clinical endpoints (e.g., binary) or to certain simplified contexts such as the case of "constant biomarkers" in HIV vaccine trials [7, 8, 11]. A constant biomarker, as defined by GH [7], refers to a biomarker that has a constant value for all subjects who receive placebo.

This article is directly motivated by a phase III trial of ZOSTAVAX conducted by Merck Research Laboratories [12]. ZOSTAVAX is a single dose vaccine that helps prevent herpes zoster in the older adult population. It protects subjects from herpes zoster by boosting the immune system. One of the objectives of the trial is to assess whether the vaccine-induced reduction in the disease incidence rate is associated with the vaccine-induced elevation in immunogenicity. One of the clinical endpoints in this trial is the time from vaccination to disease development, and a candidate surrogate marker is an immune response measured at 6 weeks post-vaccination. In this article, we develop a multiple imputation approach to evaluate candidate principal surrogate markers by incorporating baseline covariates to predict potential surrogate markers. The proposed method can accommodate various types of clinical endpoints, including continuous, discrete and time-to-event measurements. It is also applicable to a general setting including the scenarios with constant biomarkers and scenarios with the biomarkers having arbitrary variability in the placebo group. In addition, the inference for the surrogacy of biomarkers based on the proposed method accounts for the uncertainty due to the unobserved potential surrogate markers. Simulation studies are conducted to examine the performance of the proposed method and it is applied to the motivating the herpes zoster vaccine trial.

The article is organized as follows. In section "[Methods](#)", we first present a general framework of potential surrogate markers and introduce the concept of principal surrogacy. We then discuss the non-identifiability issues inherent in the evaluation of principal surrogates and propose a multiple imputation-based approach that incorporates baseline biomarkers. In section "[Misclassification in Principal Strata Membership and the Estimation of Average Causal Treatment Effects](#)", we investigate the impact of misclassification in principal strata membership on the estimation of causal treatment effects. Results of the simulation studies to assess the performance of the proposed MI method are summarized in section "[Simulation Study](#)". In section "[Application to a Herpes Zoster Vaccine Trial](#)", we apply the proposed method to evaluate an immunological marker as a principal surrogate using data from the motivating herpes zoster vaccine trial. Finally a summary of the findings and some discussions are given in section "[Discussion](#)".

Methods

Potential Outcomes and Principal Surrogate Endpoints

We consider a two-arm vaccine trial with N subjects ($i = 1, \dots, N$) randomly assigned to placebo ($Z = 0$) or vaccine ($Z = 1$). For subject i , S_i denotes the post-baseline immune response, which is a candidate surrogate marker measured at a fixed time t_0 after randomization; B_i denotes the baseline immune response that may correlate with the candidate surrogate marker S_i ; T_i denotes the time from randomization to disease development, which is the clinical endpoint; C_i denotes the censoring time; Y_i defined as $\min(T_i, C_i)$, denotes the observed time to disease development or censoring, whichever happens first; δ_i , defined as $I(T_i < C_i)$, is an indicator of whether a patient had disease ($\delta_i = 1$) or not ($\delta_i = 0$). We assume the post-baseline immune response S_i is measured prior to disease ($T_i > t_0$). This assumption approximately holds in vaccine trials where the immune response is measured shortly (e.g., a few weeks) after vaccination and is well before disease development. Let $T_i(Z)$ denote the potential clinical endpoint if subject i is assigned to treatment Z ($Z = 0$ or $Z = 1$). Similarly, let $S_i(Z)$ denote the potential surrogate marker if subject i is assigned to treatment Z ($Z = 0$ or $Z = 1$). The observed surrogate marker for subject i is a function of the two potential outcomes: $S_i = Z_i S_i(1) + (1 - Z_i) S_i(0)$. FR [6] proposed partitioning all subjects into principal strata based on the joint values of their potential surrogate markers, $S_i(0)$ and $S_i(1)$, such that all subjects within each stratum have the same value of $\{S_i(0), S_i(1)\}$. The principal stratification approach essentially assesses how the vaccine effect on T varies with subgroups defined by fixed levels of $\{S_i(0), S_i(1)\}$.

The vaccine effects on T may be measured by contrasting survival curves or by contrasting hazard functions; we focus on the latter approach. Define

$$\text{hazard}_{(1)}(s_1, s_0) = h\left(T = t \mid Z = 1, S(1) = s_1, S(0) = s_0\right) \quad (1)$$

$$\text{hazard}_{(0)}(s_1, s_0) = h\left(T = t \mid Z = 0, S(1) = s_1, S(0) = s_0\right) \quad (2)$$

where $\text{hazard}_{(1)}(s_1, s_0)$ is the hazard of disease at time t for vaccine recipients in the stratum with $S(1) = s_1$ and $S(0) = s_0$; and $\text{hazard}_{(0)}(s_1, s_0)$ is for placebo recipients in the same stratum. The ratio of these hazards measures the clinical treatment effect in the subgroup with $\{S(1) = s_1, S(0) = s_0\}$. This hazard ratio is not completely a causal estimand, because $\text{hazard}_{(1)}(s_1, s_0)$ and $\text{hazard}_{(0)}(s_1, s_0)$ condition on different sets ($\{T(1) > t, S(1) = s_1, S(0) = s_0\}$ and $\{T(0) > t, S(1) = s_1, S(0) = s_0\}$), respectively [8, 13]. However, for rare disease outcomes the sets may tend to be similar, and the critical feature holds that the biomarker levels $\{S(1) = s_1, S(0) = s_0\}$ are the same for both hazards, such that the principal strata are independent of the treatment assignment [6]. An alternative approach would focus on a contrast of completely causal survival functions in subgroups defined by $\{S_i(0), S_i(1)\}$,

for which very similar statistical methods could be used. In particular, the Cox model proposed below can be fit using the methods in the article, and fitted values used to estimate a causal conditional vaccine efficacy parameter defined as one minus the ratio of the cumulative probability of $T(1)$ below t versus the cumulative probability of $T(0)$ below t , with both terms conditional on the subgroup with $\{S(1) = s_1, S(0) = s_0\}$.

A principal surrogate marker may be defined based on the comparison of $hazard_{(1)}(s_1, s_0)$ and $hazard_{(0)}(s_1, s_0)$ according to the following two criteria:

1. Average Causal Necessity (ACN) [6]

$$hazard_{(1)}(s_1, s_0) = hazard_{(0)}(s_1, s_0) \text{ for all } s_1 = s_0.$$

2. Average Causal Sufficiency (ACS) [7]

$$hazard_{(1)}(s_1, s_0) \neq hazard_{(0)}(s_1, s_0) \text{ for all } s_1 - s_0 > C \text{ for some constant } C \geq 0.$$

ACN states that within the principal strata where the vaccine induces no change in the surrogate marker ($s_1 = s_0$), there is no vaccine-induced reduction in the hazard of disease. ACS states that within the principal strata where the vaccine induces a sufficient change in the surrogate marker ($s_1 - s_0 > C$), there are corresponding vaccine-induced changes on the hazard of disease. We refer to the strata with $S(1) = S(0)$ as the “causal necessity” strata and the strata with $S(1) > S(0)$ as the “causal sufficiency” strata. According to FR [6], the average treatment effect within the “causal necessity” strata is called the “Average Causal Dissociative Effect (ACDE)”, and the average treatment effect within the “causal sufficiency” strata is called the “Average Causal Associative Effect (ACAE)”. Note that the ACDE is equivalent to the “average principal strata indirect effect” and the ACAE is equivalent to the “average principal strata direct effect” as defined in VanderWeele [14]. The criteria ACN and ACS are not the only criteria for judging the predictive value of a surrogate marker; another useful criterion is examination of how the point and confidence interval estimates of vaccine efficacy vary across subgroups $\{S(1) = s_1, S(0) = s_0\}$, with the most useful surrogate markers having widely varying efficacy across the subgroups.

Following the existing literature on principal stratification, we make the following three assumptions throughout the article:

- (A1) the stable unit treatment value assumption (SUTVA; Rubin [27, 28]) that one subject’s treatment assignment does not affect another subject’s outcomes, and consistency, that the observed outcomes equal the potential outcomes under the assignment received;
- (A2) the ignorable treatment assignment assumption that Z_i is independent of the potential surrogate markers and the potential clinical outcomes;
- (A3) there is no population-level treatment effect on the clinical endpoint in the strata with negative effect on the surrogate marker ($S(1) < S(0)$).

As discussed in Wolfson and Gilbert [15], (A1) is plausible in trials where subjects do not interact with one another and (A2) is valid in randomized clinical trials. In addition, (A2) may be relaxed to the assumption that the treatment or

vaccine has no causal effect on the clinical outcome for any individual before the marker is measured, and this is reasonable given that some time is needed for protective immunity to develop. (A3) is a simplifying assumption for practical purposes. Theoretically, (A3) may be violated as a treatment effect can exist in the $S(1) < S(0)$ subgroup if the candidate surrogate is poor. In practice, however, very few subjects are expected to fall into this category for a reasonable candidate surrogate (as in the ZOSTAVAX example to be discussed in section “[Application to a Herpes Zoster Vaccine Trial](#)”), and thus the impact on the inference is likely small. In most vaccine applications, for example, the only reason why the immune response to vaccine would be lower than the immune response to placebo is noise in the immunological assay. In this scenario, the principal stratum with $S(1) < S(0)$ may be expected to have the same vaccine efficacy as the principal stratum with $S(1) = S(0)$; and if one wishes to account for this, (A3) can be relaxed to the more realistic assumption that the population-level vaccine efficacy for the subgroup with $S(1) < S(0)$ equals that for the subgroup with $S(1) = S(0)$.

Under the above assumptions, we define the *ACDE* for the stratum with $\{S(1) = s_1, S(0) = s_0 \mid s_1 = s_0\}$ in a log hazard ratio scale as

$$ACDE(s_1, s_0) = \log \left(\frac{\text{hazard}_{(1)}(s_1, s_0 \mid s_1 = s_0)}{\text{hazard}_{(0)}(s_1, s_0 \mid s_1 = s_0)} \right), \quad (3)$$

and the *ACAE* for the stratum with $\{S(1) = s_1, S(0) = s_0 \mid s_1 > s_0\}$ as

$$ACAE(s_1, s_0) = \log \left(\frac{\text{hazard}_{(1)}(s_1, s_0 \mid s_1 > s_0)}{\text{hazard}_{(0)}(s_1, s_0 \mid s_1 > s_0)} \right). \quad (4)$$

A biomarker with $ACDE(s_1, s_0) = 0$ and $ACAE(s_1, s_0) \neq 0$ satisfies both ACN and 1-sided ACS, which indicates that subjects with no vaccine-induced immune response are not protected and subjects with a positive vaccine-induced immune response receive some protection. While ACN and ACS provide some information about the utility of a biomarker for predicting vaccine efficacy, much more information is gained by estimating $ACDE(s_1, s_0)$ and $ACAE(s_1, s_0)$ over the support of $\{S_i(0), S_i(1)\}$, where the most useful biomarkers will have $ACDE(s_1, s_0)$ near zero, and large variability in $ACAE(s_1, s_0)$ implying that some subgroups receive a large amount of protection.

The Causal Model

Let $P^0 = \{P_k^0\}_{k=1, \dots, q}$ denote the collection of q possible principal strata constructed based on $S(0)$ and $S(1)$, with subjects sharing the same combination of $S(0)$

and $S(I)$ within each distinct stratum. The total number of possible principal strata (q) depends on the levels of the surrogate marker (S). For example, if S is a binary measurement, then q is $2^2 = 4$; if S has four levels, then q is $4^2 = 16$. Assuming the hazard if assigned placebo is constant across principal strata (after adjusting for baseline covariates that can explain the difference in baseline hazard), we consider a Cox proportional hazard model for estimating $ACDE(s_I, s_0)$ and $ACAE(s_I, s_0)$:

$$\text{hazard}_{(z)}(P_k^0) = \text{hazard}_{(0)}(P_1^0) \exp\left(\sum_{k=1}^q \gamma_k \times Z \times P_k\right) \quad (5)$$

where $\text{hazard}_{(z)}(P_k^0)$ is the hazard of disease for subjects assigned $Z = z$ in stratum P_k^0 for $k = 1, \dots, q$; P_k is a dummy variable indicating whether the subject is in stratum P_k^0 . The average (almost) causal treatment effect (ACE in the form of a log hazard ratio) in stratum P_k^0 is given by

$$ACE(P_k^0) = \gamma_k. \quad (6)$$

If the stratum P_k^0 is the “causal necessity stratum”, then $ACE(P_k^0)$ measures $ACDE$. Similarly, if stratum P_k^0 is the “causal sufficiency stratum”, then $ACE(P_k^0)$ measures $ACAE$. As we will show in Section “[Application to a Herpes Zoster Vaccine Trial](#)”, there could be multiple causal necessity strata as well as multiple causal sufficiency strata depending on the number of levels of the surrogate marker. Note that, in practice, some principal strata are more frequently observed than others. In such cases, restrictions on model parameters can be imposed in model estimation. We will elaborate on this in Section “[Application to a Herpes Zoster Vaccine Trial](#)” where we provide an empirical implementation of our method.

Identifiability of the Causal Model and Imputation of Missing Potential Surrogate Markers

The principal strata are constructed based on potential surrogate markers. However, in most clinical trials the two potential surrogate outcomes $S_i(0)$ and $S_i(1)$ for subject i cannot be observed simultaneously. In Table 1, we present the observable and unobservable potential surrogate markers in a standard clinical trial. It can be seen that in the placebo arm ($Z = 0$), $S(1)$ cannot be observed for any subjects; while in the vaccine arm ($Z = 1$), $S(0)$ are missing for all subjects. Assuming the surrogate marker (S_i) is measured for every subject in the study, then every subject would have one missing potential surrogate marker and the observed surrogate marker for subject i is $S_i = Z_i S_i(1) + (1 - Z_i) S_i(0)$.

Given the unobserved potential surrogate markers, the causal model defined in Eq. 5 and the related estimands for the $ACEs$ as specified in Eq. 6 are not identifiable based on the observed data. As Rubin [16] noted, “all problems of causal inference

Table 1 The observed and missing potential surrogate markers in a standard clinical trial

Treatment	Subject index	Potential surrogate	
		S(0)	S(1)
Z=0 (Placebo)	1	$S_1^{obs}(0)$	Missing
	2	$S_2^{obs}(0)$	Missing

	g	$S_g^{obs}(0)$	Missing
Z=1 (Vaccine)	g+1	Missing	$S_{g+1}^{obs}(0)$
	g+2	Missing	$S_{g+2}^{obs}(0)$

	n	Missing	$S_n^{obs}(0)$

should be viewed as a problem of missing data”, thus the evaluation of principal surrogates can be cast as a problem of missing potential surrogate markers.

We propose to impute the missing potential surrogate markers using the baseline surrogate marker measure (B). For biomarkers that are normally distributed, the relationships between the potential surrogate outcomes and the baseline biomarker can be described by two normal models as follows:

$$S_i(0) = \alpha_0 + \alpha_1 B_i + \varepsilon_0, \text{ where } \varepsilon_0 \sim Normal(0, \sigma_0^2), \tag{7}$$

$$S_i(1) = \beta_0 + \beta_1 B_i + \varepsilon_1, \text{ where } \varepsilon_1 \sim Normal(0, \sigma_1^2). \tag{8}$$

Model (7) can be estimated from subjects in the placebo group, and the missing value of $S(0)$ for subjects in the vaccine group can be subsequently imputed based on the fitted models. If the surrogate marker is a binary measurement, then logistic regression or probit model can be employed to impute the missing values. Similarly, log-linear models can be used to impute the missing multinomial data. We can use the model fitted from the placebo group to predict the missing $S(0)$ in the vaccine group because randomization makes the two groups equivalent and the potential surrogate marker $S(0)$ is independent of the treatment assignment. Or, in other words, $S(0)|B, Z = 0 \stackrel{D}{=} S(0)|B, Z = 1$, where D denotes equality in distribution. Similarly, missing values of $S(1)$ for subjects in the placebo group can be imputed from the fitted Model (8) using subjects in the vaccine group.

With the imputed potential surrogate markers $S(0)$ and $S(1)$, the causal model (5) can be identified and the surrogate value of a biomarker can be assessed based on the estimated treatment effects as defined in Eq. 6. The validity and precision of the inferences will depend on the accuracy of the imputed values. To account for the uncertainty associated with the imputed values, we use a multiple imputation (MI)

approach [17], which substitutes the missing values with a set of plausible values of the potential surrogate markers. As illustrated in Table 1, the potential surrogate markers are missing at random (MAR), as the probability that a potential surrogate marker is missing depends only on the observed treatment assignment. This is very important because the validity of the multiple imputation technique relies on the MAR assumption. The MI procedure consists of three stages: imputation stage, analysis stage and repeated-inference stage. In the imputation stage, the missing potential surrogate markers are filled in m times to generate m “complete” datasets in which both $S(0)$ and $S(1)$ are available for all subjects. Assuming the baseline immune response (B) is measured for every subject in the study cohort, the potential surrogate markers have a monotone missing pattern, and we adopt a Bayesian imputation procedure [17] to impute the missing potential outcomes. In the analysis stage, the imputed completed datasets are analyzed using the standard methods. Specifically, in each complete dataset, potential surrogate markers are categorized based on pre-specified thresholds. Then subjects are partitioned into principal strata according to their categorical potential surrogate markers. The $ACE(P_k^0)$ within each stratum is estimated by fitting Model (5) using the complete data. In the repeated-inference stage, the results from the analysis of the m multiply-imputed datasets are combined using Rubin’s rule [17] to obtain an overall inference about the $ACE(P_k^0)$.

Misclassification in Principal Strata Membership and the Estimation of Average Causal Treatment Effects

A caveat of the MI method described in the previous section is that subjects could potentially be misclassified into wrong principal strata based on the imputed data. Due to the uncertainty in the missing potential surrogate markers, the misclassification in principal strata membership is inevitable. In this section, we examine the unique pattern of the misclassification in principal strata (PS) membership, and investigate how such misclassification affects the estimation of the stratum-specific $ACEs$. We focus on two situations: one is the so called “constant biomarker” case and the other is a more general case.

Consider a dichotomous surrogate marker with two levels (0 = low immunogenicity and 1 = high immunogenicity). A constant biomarker means that $S(0)$ can only take the value of 0 while $S(1)$ can take the values of 0 or 1. As a result, there are a total of 2 possible principal strata based on the possible values of $\{S(0), S(1)\}$, denoted as $\{00\}$ and $\{01\}$. Stratum $\{00\}$ consists of subjects who have no change in immunogenicity if vaccinated, and stratum $\{01\}$ contains subjects who have increased immunogenicity if vaccinated. In the case of constant biomarker, $S(0)$ is known for all subjects whereas $S(1)$ are observed only for subjects in the vaccine group. For vaccine recipients, the observed principal strata are constructed based on the observed $S(0)$ and $S(1)$; while for placebo recipients, the observed principal

Table 2 The observed versus the true principal strata in the general case

True principal strata	Observed principal strata based on imputed data				Total
	{00}	{11}	{01}	{10}	
{00}	n_{00_00} Correctly classified	0	n_{00_01} Placebo recipients from 00	n_{00_10} Vaccine recipients from 00	n_{00}^{true}
{11}	0	n_{11_11} Correctly classified	n_{11_01} Vaccine recipients from 11	n_{11_10} Placebo recipients from 11	n_{11}^{true}
{01}	n_{01_00} Placebo recipients from 01	n_{01_11} Vaccine recipients from 10	n_{01_01} Correctly classified	0	n_{01}^{true}
{10}	n_{10_00} Vaccine recipients from 10	n_{10_11} Placebo recipients from 10	0	n_{10_10} Correctly classified	n_{10}^{true}
Total	n_{00}^{obs}	n_{11}^{obs}	n_{01}^{obs}	n_{10}^{obs}	

strata are constructed based on the observed $S(0)$ and the imputed $S(1)$. Therefore misclassification in PS membership may take place only among placebo recipients due to the uncertainty in the imputed $S(1)$.

In the general scenario, $S(0)$ is no longer a constant and it can take the value of either 0 or 1. Similarly $S(1)$ can take the value of either 0 or 1. Therefore there are four possible principal strata corresponding to the possible combinations of $S(0)$ and $S(1)$; {00} and {11} are the strata in which subjects have no change in immunogenicity if vaccinated; {01} is the stratum in which subjects have increased immunogenicity if vaccinated; and stratum {10} contains subjects who have decreased immunogenicity if vaccinated. For placebo recipients, the observed principal strata are constructed based on the known $S(0)$ and the imputed $S(1)$; while for vaccine recipients, the observed principal strata are constructed based on the known $S(1)$ and the imputed $S(0)$. In Table 2, we list the number of subjects in the true PS membership vs. the number of subjects in the observed PS membership based on imputed potential surrogate markers. Note that the first two digits in n 's subscript indicate the true stratum to which the subjects belong and the trailing two digits specify the stratum to which the subjects are classified. It can be seen that

every observed principal stratum consists of correctly classified subjects, misclassified placebo recipients from one stratum and misclassified vaccine recipients from another stratum.

Under the assumptions of (A1) to (A3) and the causal model (5), the observed ACE in a specific stratum is essentially a weighted average of the true ACE in this specific stratum and the true ACE in the stratum to which the misclassified vaccine recipients originally belong. Of note the misclassified placebo recipients have no effect on the estimation of $ACEs$ because the baseline hazard is constant across principal strata in model (5). More specifically,

$$ACE_{00}^{obs} = p_{10_00}ACE_{10}^{true} + (1 - p_{10_00})ACE_{00}^{true} = (1 - p_{10_00})ACE_{00}^{true} \quad (9)$$

$$ACE_{11}^{obs} = p_{01_11}ACE_{01}^{true} + (1 - p_{01_11})ACE_{11}^{true}, \quad (10)$$

$$ACE_{01}^{obs} = p_{11_01}ACE_{11}^{true} + (1 - p_{11_01})ACE_{01}^{true}, \quad (11)$$

$$ACE_{10}^{obs} = p_{00_10}ACE_{00}^{true} + (1 - p_{00_10})ACE_{10}^{true} = p_{00_10}ACE_{00}^{true}, \quad (12)$$

where ACE_{00}^{true} , ACE_{11}^{true} , ACE_{01}^{true} and ACE_{10}^{true} denote the true $ACEs$ in strata $\{00\}$, $\{11\}$, $\{01\}$ and $\{10\}$, respectively. Note $ACE_{10}^{true} = 0$ as a result of assumption (A3). The weight is the proportion of misclassified vaccine recipients in the stratum. The first two digits in p 's subscript indicate the true stratum to which the subjects belong and the trailing two digits specify the stratum to which subjects are misclassified. Equations 9, 10, 11, and 12 show that the estimated $ACEs$ in the observed strata may be biased in the presence of misclassification in PS membership. However, the direction of bias can be determined from Eqs. 9, 10, 11, and 12 and the effect of misclassification in PS membership on the estimation of the causal treatment effects can be assessed. For example, if a biomarker is a good surrogate that satisfies both ACN and ACS, then $ACE_{00}^{true} = 0$, $ACE_{11}^{true} = 0$ and $ACE_{01}^{true} < 0$ (assuming the vaccine has a protective effect on the hazard of disease). From Eqs. 9, 10, and 11, we can see that $ACE_{00}^{obs} = 0$; $ACE_{11}^{obs} = p_{01_11}ACE_{01}^{true} < 0$; and $ACE_{01}^{obs} = (1 - p_{11_01})ACE_{01}^{true}$. Hence misclassification of PS membership may result in an underestimation of the surrogate value of a good surrogate marker. If a biomarker has no surrogate value, then $ACE_{01}^{true} = 0$ and ($ACE_{00}^{true} < 0$ and/or $ACE_{11}^{true} < 0$). An inspection of Eqs. 9, 10, and 11 shows that $ACE_{00}^{obs} = (1 - p_{10_00})ACE_{00}^{true}$; $ACE_{11}^{obs} = (1 - p_{01_11})ACE_{11}^{true}$; and $ACE_{01}^{obs} = p_{11_01}ACE_{11}^{true}$. Hence misclassification in PS membership may result in an overestimation of the surrogate value for a biomarker that has completely no surrogate value. If there are similar $ACEs$ in strata $\{00\}$, $\{11\}$ and $\{01\}$ (i.e., a biomarker with partial surrogate value), then the misclassification in PS membership has little effect on the estimation of the $ACEs$ in these strata.

Simulation Study

Data Generation and Simulation Design

We evaluated the performance of the proposed MI approach through simulation studies. We considered a continuous surrogate marker that can be dichotomized into two levels: low ($S=0$) and high ($S=1$). The data were generated to mimic the structure of a placebo-controlled vaccine trial, which motivated our proposed method and will be discuss more in section “Application to a Herpes Zoster Vaccine Trial”. The clinical endpoint is the time from vaccination to disease development. The candidate surrogate marker is an immune response at 6 weeks post-vaccination and is measured as the natural logarithm of the antibody fold rise from baseline. We examined the performance of the proposed method in (a) the constant biomarker case with $S(0)=0$; and (b) the general case in which $S(0)$ has arbitrary variability.

Data were generated for N subjects ($i = 1, \dots, N$) with 1:1 randomization to placebo ($Z=0$) or vaccine ($Z=1$). The baseline immune response (B) was measured for every subject and was randomly drawn from $B_i \sim Normal(5.56, 1.00)$. For the general case, the potential post-baseline immune responses $P(0)$ and $P(1)$ were generated as functions of B by two linear models: $P_i(0) = 0.30 + 0.93B_i + \varepsilon_{0i}$, where $\varepsilon_{0i} \sim Normal(0, 0.34^2)$; and $P_i(1) = 3.60 + 0.56B_i + \varepsilon_{1i}$, where $\varepsilon_{1i} \sim Normal(0, 0.68^2)$. The potential surrogate markers were the changes in the immune response from baseline, i.e., $S_i(0) = P_i(0) - B_i$, $S_i(1) = P_i(1) - B_i$. The principal strata were constructed by dichotomizing the potential surrogate markers at a pre-specified threshold value c : $P_{00i} = I(S_i(0) < c, S_i(1) < c)$, $P_{01i} = I(S_i(0) < c, S_i(1) \geq c)$, $P_{11i} = I(S_i(0) \geq c, S_i(1) \geq c)$ and $P_{10i} = I(S_i(0) \geq c, S_i(1) < c)$. We considered $c=0.36$ in the simulation. The true survival time was generated from an exponential-Cox model (Bender et al. [29]) with proportional hazards as follows:

$$T_i = - \frac{\log(U_i)}{8 \times 10^{-4} \times \exp(Z_i \times (\gamma_{00}P_{00i} + \gamma_{01}P_{01i} + \gamma_{11}P_{11i} + \gamma_{10}P_{10i}))},$$

where $U_i \sim Uniform(0,1)$. The censoring time was $C_i = 365 + Uniform(0,1) \times 183$. The observed survival or censoring time was $Y_i = \min(T_i, C_i)$ and the indicator of whether the subject had the event was $\delta_i = I(T_i < C_i)$. The average cumulative probability of disease in the placebo group is 29% and the overall censoring percentage during the study is about 75%. For the case of constant biomarker, we set $P_i(0) = B_i$ in the above data generating process such that $S_i(0) = P_i(0) - B_i = 0$ for subject $i = 1, \dots, N$. The true ACDE in the causal necessity strata $\{00\}$ and $\{11\}$ were γ_{00} and γ_{11} , respectively. The true ACAE in the causal sufficiency stratum $\{01\}$ was γ_{01} . We set γ_{10} to 0 such that there was no vaccine effect in stratum $\{10\}$ [consistent with (A3)].

In the generated data, $S(1)$ for subjects with $Z=0$ were set to missing in the constant biomarker case; and $S(0)$ for subjects with $Z=1$ and $S(1)$ for subjects with $Z=0$ were set to missing in the general case. The MI approach was applied

to fill in the missing values. The imputed $S(0)$ for subjects with $Z = 1$ was drawn from its predictive distribution obtained from the estimated model of Eq. 7 using subjects in the placebo group. Similarly, the imputed $S(1)$ for subjects with $Z = 0$ was drawn from its predictive distribution based on the model of Eq. 8 estimated using subjects in the vaccine group. For every simulation setting, we considered 1,000 replications. In each replication, we considered a trial of $N = 2500$ subjects (1:1 allocation). For every trial, we conducted $m = 10$ imputations for the potential surrogate markers. The parameters in the data generating process were selected such that there were significant overall vaccine effects on T in all the generated data. The performance of the proposed MI approach was evaluated by different metrics, including (i) the probability of rejecting the hypothesis that the ACE in a stratum equaling zero (i.e., the type I error rate when the true ACE is zero and the power when the true ACE not equal to zero); (ii) the absolute bias of the ACE estimate in terms of log hazard ratios; and (iii) the coverage rate of the 95% CIs for the ACE s. We varied the parameters γ_{00} , γ_{11} and γ_{01} to generate data with the candidate surrogate marker having different surrogate values. Three scenarios were considered where the biomarker has (a) no surrogate value ($ACAE = 0$, $ACDE \neq 0$); (b) moderate surrogate value ($ACAE \neq 0$, $ACDE \neq 0$); and (c) high surrogate value ($ACAE \neq 0$, $ACDE = 0$ such that ACN and ACS hold).

Simulation Results for the Case of Constant Biomarker

Table 3 presents, for each principal stratum, the average number of subjects, the absolute bias, the standard error, the coverage probability of the 95% CI of the estimated ACE , the empirical type I error rate under the null hypothesis of no ACE , and the power to reject the null when the true ACE is not equal to 0. We present both the results based on fully observed $S(1)$ (as an unattainable benchmark) and the results based on multiply imputed $S(1)$. In the analysis based on fully observed $S(1)$, we assume the potential surrogate marker $S(1)$ was observed for all subjects such that there was no misclassification in the PS membership. In the analysis based on multiply imputed $S(1)$, we set the $S(1)$ in the placebo group to missing and applied the MI approach. Compared with the analysis based on the fully observed $S(1)$, the bias of the MI method appeared to be negligible. The coverage probabilities were close to but slightly above the target level. Furthermore, the test for the principal stratum treatment effect is conservative in terms of type I error rate, which was consistent with findings reported in other MI-based approaches (e.g., [18, 19]). The simulation results show that the proposed MI approach is a valid approach for the evaluation of principal surrogacy of constant biomarkers. As we discussed in section “[Misclassification in Principal Strata Membership and the Estimation of Average Causal Treatment Effects](#)”, in the constant biomarker case, only placebo recipients can be misclassified, the simulation results also showed that misclassification of placebo recipients has negligible effect on the estimation of the causal treatment effects.

Table 3 Simulation results based on 1000 replications for the constant biomarker case

Average overall		Based on fully observed S(1)						Based on multiply-imputed S(1)								
		γ_{00}	γ_{01}	VE	PS	Type I error rate (%)	Mean # of subjects	Bias	SE	CP (%)	Power (%)	Type I error rate (%)	Mean # of subjects	Bias	SE	CP (%)
No surrogate value	0	-1.4	0.61	{00}	5.3	524	-0.01	0.16	94.7	5.3	523	0.00	0.14	99.2	0.8	100.0
				{01}	4.2	1,976	-0.00	0.12	95.8	4.2	1,976	-0.01	0.12	96.6	0.7	100.0
Moderate surrogate value	0	-0.8	0.44	{00}	4.2	525	0.01	0.16	95.8	4.2	524	0.00	0.14	99.3	0.7	100.0
				{01}	5.3	1,975	-0.01	0.10	95.3	5.3	1,976	-0.01	0.10	95.9	0.8	100.0
High surrogate value	-0.8	-0.8	0.55	{00}	4.7	524	0.00	0.20	95.2	4.7	524	-0.01	0.18	98.2	0.8	99.2
				{01}	5.3	1,976	-0.00	0.10	95.3	5.3	1,976	-0.00	0.10	97.0	0.7	100.0
High surrogate value	-0.8	-0.4	0.38	{00}	4.6	525	-0.01	0.20	96.2	4.6	525	-0.01	0.17	98.9	0.7	99.2
				{01}	5.3	1,976	-0.00	0.09	94.7	5.3	1,975	-0.00	0.08	96.7	0.8	99.5
High surrogate value	-2.0	0	0.20	{00}	4.6	524	-0.05	0.33	94.8	4.6	523	-0.04	0.31	97.5	0.8	100.0
				{01}	5.3	1,976	0.01	0.08	95.3	4.7	1,977	0.00	0.08	96.9	3.1	99.9
High surrogate value	-2.5	0	0.22	{00}	4.6	525	-0.05	0.42	96.4	4.6	525	-0.08	0.70	97.0	0.8	99.9
				{01}	5.3	1,975	0.00	0.08	95.4	4.6	1,975	0.00	0.08	96.3	3.7	99.9

CP coverage probability for 95% confidence interval, PS principal stratum; the overall VE is the overall vaccine efficacy, measured as one minus the marginal hazard ratio; Type I error and power are the empirical probabilities of rejecting the null hypothesis that there is no average causal treatment effect in the principal stratum (testing performed at two-sided 5% level)

We also conducted simulation studies to evaluate other factors that might affect the performance of the proposed method. For example, we found that the results are similar across different numbers of imputations ($m = 10$ and 100). But in small studies ($N = 600$) there was slight improvement as m increases in the method in terms of test power and coverage probability. Therefore we recommend using a large number of imputations if the computational time is not a concern. We compared the performance of the Bayesian imputation model and the Frequentist imputation model as described in Lu et al. [18]. The results from the two imputation procedures are very similar, with slightly better performance for the Bayesian imputation. Detailed results are not included here for brevity.

Simulation Results for the General Case

As discussed in section “[Misclassification in Principal Strata Membership and the Estimation of Average Causal Treatment Effects](#)”, the PS membership can be misclassified based on the imputed data and the estimated *ACEs* may be biased in the general case. The simulation study conducted in the general setting aims to investigate the direction of the bias and the effect of misclassification on the estimation of the principal surrogate value. In the general case, we have four possible principal strata: $\{00\}$, $\{11\}$, $\{01\}$ and $\{10\}$. There are very few subjects falling into the stratum $\{10\}$ in our simulation study, therefore we focused on the strata $\{00\}$, $\{11\}$ and $\{01\}$. Table 4 presents, for each principal stratum, the average number of subjects, the absolute bias, the standard errors, the coverage probability of the 95% CI of the estimated *ACE*, the empirical type I error rate under the null hypothesis of no *ACE*, and the power to reject the null when there is principal stratum specific *ACE*. In general, the proposed method provides reasonable power to detect average causal effects when they exist. However, the MI estimates were biased in some scenarios. The bias is caused by the misplacement of vaccine recipients into wrong principal strata and the direction of the bias is consistent with our previous conjecture as described in section “[Misclassification in Principal Strata Membership and the Estimation of Average Causal Treatment Effects](#)”. Note that the type I error rates of the tests for no *ACE* in some strata are high mainly because of the biased estimates of *ACEs* in these strata. As shown in the simulation study, we find that the misclassification of PS membership results in an underestimation of the surrogate value for good surrogates and overestimation of the surrogate value for poor surrogates.

Summary of Simulation Studies

To summarize, our simulation studies showed that misclassification of placebo recipients has little effect on the estimation of the treatment effects and the proposed

Table 4 Simulation results based on 1000 replications for the general case with S(0) having arbitrary variability

Average overall VE	Based on fully observed S(0) and S(1)										Based on multiply-imputed S(0) and S(1)						
	γ_{00}	γ_{11}	γ_{01}	PS	Mean # of subjects			Type I error rate (%)			Power (%)	Mean # of subjects	Bias	SE	CP (%)	Type I error rate (%)	Power (%)
					{00}	{11}	{01}	Bias	SE	CP (%)							
No surrogate value	0	0	-1.4	0.52	{00}	461	0.00	0.18	94.3	5.7	464	0.00	0.14	99.7	0.3	100.0	
					{11}	319	0.00	0.20	96.2	3.8	319	-1.01	0.18	5.2	94.8		
					{01}	1,659	-0.01	0.14	94.7		1,659	0.36	0.11	15.8			
Moderate surrogate value	0	0	-0.8	0.37	{00}	461	0.01	0.17	94.4	5.6	464	0.00	0.13	99.9	0.1	1.0	
					{11}	319	0.00	0.22	93.3	6.8	319	-0.62	0.17	44.0	56.0		
					{01}	1,659	0.00	0.11	95.7		1,659	0.16	0.10	71.8			
High surrogate value	-0.3	-0.3	-0.8	0.45	{00}	461	-0.01	0.19	94.7		39.3	464	0.03	0.15	99.3	14.4	
					{11}	319	-0.01	0.22	94.9		26.5	319	-0.40	0.18	90.0	66.9	
					{01}	1,659	0.00	0.11	95.1		100.0	1,659	0.09	0.10	90.0	100.0	
	-0.8	-0.8	-0.8	0.54	{00}	461	-0.01	0.21	94.7		98.2	464	0.12	0.18	96.4	90.5	
					{11}	319	0.00	0.26	94.8		89.8	319	-0.01	0.18	100.0	82.4	
					{01}	1,659	0.00	0.11	95.3		100.0	1,659	0.00	0.10	98.2	100.0	
	-0.8	-0.8	-0.4	0.39	{00}	461	0.00	0.22	94.2		97.5	464	0.12	0.17	97.4	91.0	
					{11}	319	-0.01	0.25	95.5		91.4	319	0.34	0.16	91.7	24.8	
					{01}	1,659	0.00	0.10	94.9		98.5	1,659	-0.06	0.09	94.8	99.6	
	-2.5	-2.5	0	0.32	{00}	461	-0.09	0.72	96.2		99.9	464	0.73	0.27	37.6	100.0	
					{11}	319	-0.23	1.58	96.3		99.3	319	2.29	0.16	0.0	2.4	
					{01}	1,659	0.00	0.09	94.4	5.6	1,659	-0.18	0.09	57.1	42.9		
	-2.0	-2.5	0	0.30	{00}	461	-0.04	0.34	95.7		100.0	464	-0.49	0.24	60.4	100.0	
					{11}	319	-0.20	1.37	96.5		99.5	319	2.30	0.16	0.0	2.3	
					{01}	1,659	0.00	0.09	94.5	5.5	1,659	-0.18	0.09	58.1	41.9		

CP coverage probability for 95% confidence interval, PS principal stratum; the overall VE is the overall vaccine efficacy, measured as one minus the marginal hazard ratio; Type I error and power are the empirical probabilities of rejecting the null hypothesis that there is no average causal treatment effect in the principal stratum (testing performed at two-sided 5% level)

MI approach provides valid assessment of principal surrogacy in the case of constant biomarker. In the general case, the estimated treatment effects may be biased in some strata. However, the proposed method provides insightful assessment on surrogate value of biomarkers given the fact that the direction of the bias is known. Specifically, if the results based on the proposed MI approach show the biomarker has a high surrogate value, then the true surrogate value should be even higher given a likely downward bias in surrogacy estimation caused by misclassification of vaccine recipients. In contrast, an estimate of low surrogate value from the proposed method would suggest that the true surrogate value is actually even lower given a likely upward bias caused by misclassification of vaccine recipients.

Application to a Herpes Zoster Vaccine Trial

We applied the proposed MI method to evaluate an immunological marker using data from a phase III herpes zoster vaccine trial conducted by Merck Research Laboratories. The trial was a multicenter, double-blinded, placebo-controlled randomized study to examine the effectiveness of ZOSTAVAX in the prevention of herpes zoster among 50–59 years old individuals [12]. The clinical endpoint of interest is the time from randomization to herpes zoster development and the candidate surrogate is the antibody response at 6 weeks post-vaccination measured by the glycoprotein enzyme-linked immunosorbent assay (gpELISA). The study used a case-cohort design [20] where the antibody responses at baseline and 6 weeks were measured in the case-cohort population (all the herpes zoster cases occurred during the course of study plus a prospectively selected, random subset of 10% study participants). A Kaplan-Meier plot of herpes zoster events in subjects randomized in the herpes zoster vaccine trial is presented in Fig. 1.

The objective of the analysis was to examine whether the vaccine-induced reduction in the rate of herpes zoster is caused by the vaccine-induced elevation in antibody titers as measured by gpELISA at 6 weeks post-vaccination. For this purpose, 13 of the 22,439 (0.06%) trial participants who experienced the herpes zoster event prior to the antibody response measurement at 6 weeks postvaccination were excluded from the analysis. In previous analysis (not published), we found that a categorical combined score of fold rise and titer captures about 65% of the treatment effect on protection of herpes zoster using Freedman's method [21]. Based on a preliminary assessment that suggests a noticeable threshold effect of both 6-week antibody titer and fold rise on the hazard of herpes zoster, antibody titer and fold-rise from baseline were both dichotomized into "low" and "high" levels in this analysis. Therefore, for both $S(0)$ and $S(1)$, we have four ordered categories, denoted as 1, 2, 3 and 4, where 1 denotes "low titer and low fold rise"; 2 denotes "high titer and low fold rise"; 3 denotes "low titer and high fold rise"; and 4 denotes "high titer and high fold rise". The order of the number corresponds to the relative level of immunogenicity, with 4 indicating the highest level of immunogenicity.

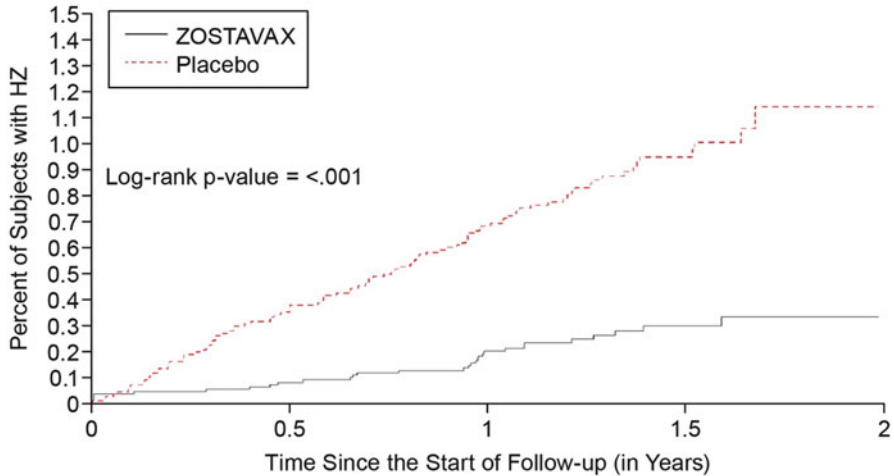


Fig. 1 Kaplan-Meier plot of herpes zoster (HZ) events in subjects randomized in the herpes zoster vaccine trial (Intention-to-treat population)

We cross-classified the subjects in the “complete” data set according to the combination of $S(0)$ and $S(1)$ to form 16 distinct principal strata. We found that some principal strata were more common than others. For example, Table 5 presents the number of subjects and the corresponding potential surrogate markers in each distinct principal stratum based on the single deterministic imputation (the imputed value is exactly the predicted value based on the Bayesian imputation model and the predictive uncertainty is ignored). Since this study was a multi-center randomized trial and only 1 out of the 2,276 subject included in the case-cohort analysis was classified into the negative surrogate effect stratum $S(0) > S(1)$ (Table 5), the assumptions (A1), (A2), and (A3) needed for the causal analysis appear reasonable and practical for this application.

Table 6 presents the estimated stratum-specific ACEs based on 100 random imputations for the three major principal strata as well as the combined causal necessity and causal sufficiency strata. The estimated ACEs in stratum {22} suggest that no significant vaccine effect is observed on the hazard of herpes zoster when there is no vaccine effect on the surrogate marker. The estimated ACEs in strata {24} and {14} show that when the vaccine elevates immunogenicity, there is a corresponding vaccine effect on the hazard of herpes zoster, and this effect becomes more pronounced as the vaccine effect on the surrogate markers increases. These results show that the vaccine effect on the immune response measured by gpELISA antibody clearly predicts the vaccine effect to protect against herpes zoster, with greater vaccine-induced immunogenicity associated with greater protection. Furthermore, the estimated ACE based on subjects from all causal necessity strata (e.g., stratum {11}, {22}, {33} and {44}) and the estimated ACE based on subjects from all causal sufficiency strata (e.g., ({12}, {13}, {14}, {23}, {24}, {34}) show

Table 5 Herpes zoster vaccine trial: name, number of subjects and the corresponding potential surrogate markers for each principal stratum based on a single deterministic imputation

	Name of principal stratum	Number of subjects	Potential surrogate markers		
			S(0)	S(1)	
Causal necessity strata	{11}	5	1	1	
	{22}	569	2	2	
	S(0) = S(1)	{33}	5	3	
	(N = 579)	{44}	0	4	
Causal sufficiency strata	{12}	3	1	2	
	{13}	28	1	3	
	S(0) < S(1)	{14}	355	1	4
	(N = 1696)	{23}	0	2	3
		{24}	1,310	2	4
		{34}	0	3	4
Negative surrogate effect strata	{21}	1	2	1	
	{31}	0	3	1	
	S(0) > S(1)	{32}	0	3	2
	(N = 1)	{41}	0	4	1
		{42}	0	4	2
		{43}	0	4	3

that the antibody response satisfies both ACN and ACS. Therefore, the 6-week antibody response as measured by gpELISA may be used as a principal surrogate for protection of herpes zoster in trials of ZOSTAVAX.

Discussion

The validation of principal surrogate markers is based on the estimation of the average causal treatment effects in the principal strata of interest. However, given the fact that every subject has one missing potential surrogate marker, it is challenging to assess the principal surrogacy of biomarkers. In this article, we proposed a multiple imputation approach for the evaluation of principal surrogate markers by incorporating a baseline predictor. The proposed method provides remarkable advantages as it is applicable to various types of clinical endpoints and to general settings in which $S(0)$ has arbitrary variability.

To our knowledge, this paper is the first to investigate the unique pattern of misclassification in the PS membership and its impact on the assessment of a biomarker's surrogate value. Simulation studies showed that the proposed MI approach provides valid inference about principal surrogacy in the case of constant biomarker. In the general case, although the MI approach may generate biased ACE estimates in some principal strata due to misclassification in PS membership,

Table 6 Herpes zoster vaccine trial: results of the estimated stratum-specific causal treatment effects for the three major strata, the combined causal necessity strata and the combined causal sufficiency strata ($m = 100$ imputations)

	Principal stratum	The estimated causal treatment effects in hazard ratio (HR)	95% CI for HR	p-value
Causal necessity strata	{22}	0.72	0.28, 1.86	0.499
	All causal necessity strata	0.64	0.27, 1.54	0.323
Causal sufficiency strata	{24}	0.17	0.05, 0.53	0.003
	{14} ^a	0.04	0.01, 0.27	0.001
		0.00	0.00, 0.00	<0.001
	All causal sufficiency strata	0.12	0.05, 0.31	<0.001

^aTwo estimates for the average causal treatment effect (ACE) in stratum {14} are presented because the distribution of the estimated ACE from the 100 multiply imputed datasets has two modes. This is due to the fact that the vaccine is very effective in stratum {14}, and in some of the imputed datasets, no subjects in the vaccine group had the event in this stratum. Hence we use Rubin's rule to combine the results for each distribution and present two estimates, both suggesting there was a strong vaccine effect in {14}

the direction of the bias can be predicted based on the correlations between the baseline predictor and the surrogate markers, and the estimated surrogate value. Unfortunately, the inference can only be made if the estimated surrogate value is low or high. It would be difficult to infer the direction of the bias when the estimated surrogate value is moderate. A further extension of the proposed method is to correct the bias in the MI approach for the general case, and further research in this area is ongoing.

In this paper, we focused on continuous biomarkers and employed parametric models for the imputation of potential surrogate outcomes (models (7) and (8)). Misspecification of the imputation model may lead to bias or reduced precision in the estimation of the analysis model; hence it is important to check the model assumptions. Remedies include transforming the data (i.e., taking logarithm of heavily skewed variables) or using some less parametric imputation approaches. Simulation studies have shown that the less parametric approaches (e.g., the predictive mean matching method [22]) are more robust to model misspecification [23]. For the analysis model, we employed Cox proportional hazard model (model (5)) to estimate the average causal treatment effects. The key assumptions for the Cox model are the proportional hazard assumption and no omission of relevant covariates. Consequences of model misspecifications and techniques to handle model misspecification have been extensively studied by Lin and Wei [24] and many others.

Almost all of the literature on statistical methods for assessing principal surrogates has assumed that the treatment/vaccine has no effect on the clinical endpoint before the surrogate marker is measured. The proposed method utilizes the same assumption by requiring that the post-baseline immune response is measured prior to the development of the clinical endpoint. This simplifying assumption makes the methodology development relatively simple and implementable. If this assumption is violated, then the proposed method may be biased and may provide overly-precise inferences [15]. For the ZOSTAVAX application, only a very small fraction (0.06%) of trial participants experienced the clinical endpoint within 6 weeks post-vaccination, and any vaccine protection conferred during this early period is expected to be small. As a result, any potential violation of the assumption would only have a small impact on the overall result.

The proposed method provides a general framework that accommodates various types of clinical endpoints. However, its application is limited to categorical surrogate markers or continuous surrogate markers that can be categorized in order to construct principal strata. For direct evaluation of continuous surrogate markers, GH [7] proposed the causal effect predictiveness (CEP) surface for the assessment of candidate principal surrogates. They provided likelihood-based approaches to estimate the CEP surface for constant biomarkers, but not for the general case. Our proposed MI approach can be employed to estimate the CEP surface for continuous markers in both the constant biomarker case and the general case.

The surrogacy of a biomarker defined by principal surrogacy is a treatment-specific concept. As Pearl [25] pointed out, a biomarker may be a good principal surrogate when a certain treatment is used but fails to be a good surrogate under new conditions. One of the biggest challenges of surrogate markers evaluation using data from a single trial is that there is always risk associated with extrapolating the information gathered from one study to a new setting. The evaluation of surrogacy of biomarkers should be a process of collecting consistent evidences across treatments, populations and even disease stages.

References

1. Chan, I., Wang, W., Heyse, J.: Vaccine clinical trials. In: *Encyclopedia of Biopharmaceutical Statistics*, 2nd edn. Dekker, New York (2003)
2. Joffe, M.M., Greene, T.: Related causal frameworks for surrogate outcomes. *Biometrics* **65**, 530–538 (2009)
3. Robins, J.M., Greenland, S.: Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155 (1992)
4. Pearl, J.: Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, pp. 411–420. Morgan Kaufmann, San Francisco (2001)
5. VanderWeele, T.: Principal stratification – uses and limitations. *Int. J. Biostat.* **7**(1), Article 28: 1–14 (2011)
6. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. *Biometrics* **58**, 21–29 (2002)

7. Gilbert, P.B., Hudgens, M.G.: Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154 (2008)
8. Qin, L., Gilbert, P.B., Follmann, D., Li, D.: Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *Annu. Appl. Stat.* **2**, 386–407 (2008)
9. Follmann, D.: Augmented designs to assess immune response in vaccine trials. *Biometrics* **62**, 1161–1169 (2006)
10. Li, Y., Taylor, J.M.G., Elliott, M.R.: A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**, 523–531 (2010)
11. Huang, Y., Gilbert, P.B.: Comparing biomarkers as principal surrogate endpoints. *Biometrics* (2011). doi:[10.1111/j.1541-0420.2011.01603.x](https://doi.org/10.1111/j.1541-0420.2011.01603.x) [Epub ahead of print]
12. Schmader, K.E., Levin, M.J., Gnann Jr., J.W., McNeil, S.A., Vesikari, T., Betts, R.F., Keay, S., Stek, J.E., Bundick, N.D., Su, S.C., Zhao, Y., Li, X., Chan, I.S.F., Annunziato, P.W., Parrino, J.: Efficacy, safety, and tolerability of herpes zoster vaccine in persons aged 50 to 59 years. *Clin. Infect. Dis.* **54**(7), 922–928 (2012)
13. Hernán, M.A.: The hazard of hazard ratios. *Epidemiology* **21**, 13–15 (2010)
14. VanderWeele, T.: Simple relations between principal stratification and direct and indirect effects. *Stat. Probab. Lett.* **78**, 2957–2962 (2008)
15. Wolfson, J., Gilbert, P.B.: Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics* **66**, 1153–1161 (2010)
16. Rubin, D.B.: Direct and indirect causal effects via potential outcomes. *Scand. J. Stat.* **31**, 161–170 (2004)
17. Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York (1987)
18. Lu, K., Jiang, L., Tsiatis, A.A.: Multiple imputation approaches for the analysis of dichotomized responses in longitudinal studies with missing data. *Biometrics* **66**, 1202–1208 (2010)
19. Wang, T., Wu, L.: Multiple imputation methods for multivariate one-sided tests with missing data. *Biometrics* (2011). doi:[10.1111/j.1541-0420.2011.01597.x](https://doi.org/10.1111/j.1541-0420.2011.01597.x)
20. Prentice, R.L.: A case-cohort design for epidemiological cohort studies and disease prevention trials. *Biometrika* **73**, 1–11 (1986)
21. Freedman, L., Graubard, B., Schatzkin, A.: Statistical validation of intermediate endpoints for chronic disease. *Stat. Med.* **11**, 167–178 (1992)
22. Little, R.J.A.: Missing-data adjustments in large surveys. *J. Bus. Econ. Stat.* **6**, 287–301 (1988)
23. Lazzeroni, L.C., Schenker, N., Taylor, J.M.G.: Robustness of multiple-imputation techniques to model misspecification. In: *American Statistical Association Proceedings of the Section on Survey Research Methods*, Anaheim, California, Aug 6–9, pp. 260–265. (1990)
24. Lin, D.Y., Wei, L.J.: The robust inference for the Cox proportional hazard model. *J. Am. Stat. Assoc.* **84**, 1074–1078 (1989)
25. Pearl, J.: Principal stratification – a goal or a tool? *Int. J. Biostat.* **7**, 1–13 (2011)
26. Halloran, M.E.: Vaccine studies. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*. Vol. 6. New York: Wiley; pp. 4687–4694 (1998)
27. Rubin, D.B.: Randomization analysis of experimental data: the fisher randomization test. *J. Am. Stat. Assoc.* **75**, 591–593 (1980)
28. Rubin, D. B.: Statistics and causal inference: comment: which ifs have causal answers. *J. Am. Stat. Assoc.* **81**, 961–962 (1986)
29. Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24**, 1713–1723 (2005)

Mapping Return Values of Extreme Wind Speeds

Adam L. Pintar and Franklin T. Lombardo

Abstract Structures subjected to wind loads must be designed to perform adequately from the points of view of stress and serviceability. Wind loading specified for design is based in part on the wind speeds affecting the site of interest. A particular quantity of interest in design is the N -year extreme wind speed, regardless of its direction, at a location of interest, defined by its longitude and latitude. Wind maps consisting of isotachs for N -year extreme wind speeds defined in building codes and standards are therefore required for structural design purposes. Alternatively, numerical versions of maps can be developed wherein automatic interpolations are performed that yield the N -year speeds at points defined by longitude and latitude. The raw data to be analyzed to develop the map are irregular time series of wind speeds above a specified threshold at multiple wind reporting stations. This work presents a two-stage approach to creating the map. The first stage involves the estimation of the parameters of an extreme value distribution at each station. In the second stage an interpolant based on the estimated parameters is created so that the N -year extreme wind speeds may be estimated at the geographical coordinates of interest. Standard errors and confidence bounds for the estimates are calculated using a non-parametric bootstrap algorithm. Results are presented for a region within Kansas, and those results are compared to the ASCE 7-10 Standard over the same region.

A.L. Pintar (✉)
National Institute of Standards and Technology, 100 Bureau Drive,
Gaithersburg, MD 20899, USA
e-mail: adam.pintar@nist.gov

F.T. Lombardo
Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180
e-mail: lombaf@rpi.edu

Introduction

An important consideration in designing structures is determining the effects induced by wind on that structure and its components. Those effects are functions of the wind speed, at or near the structure's location, that will be exceeded with some probability, p , in any one year. In building codes and standards these probabilities are expressed in terms of mean recurrence intervals (or mean return periods) $N = 1/p$ in years. Wind speeds with mean return period N are referred to in this chapter as N -year return values. Maps of estimates of N -year return values are typically used in codes and standards, including [16]. In this chapter, an algorithm for creating such maps will be described, and the algorithm will be applied to a region within Kansas. The map from [16], which covers the entire United States, is compared, over the appropriate region in Kansas, to the map developed in this work.

The algorithm described here attempts to address some potential limitations of the map in [16]. The map in [16] employs the Gumbel distribution which fixes the so called tail length parameter at zero. In addition, a single set of Gumbel parameters is assumed to describe the extreme wind climate of Kansas and the rest of the non-hurricane regions of the US. This assumption leaves the N -year return values displayed in [16] uniform over most of these non-hurricane regions, including Kansas. In this chapter, a different extreme value model is employed, but most importantly, the tail length parameter of that model is not fixed but estimated from the data. The algorithm in general also allows for the inclusion of non-uniform (i.e., regional) climatology. Finally, the map in [16] treats all extreme wind events the same. This chapter recognizes and takes into account that extreme winds can arise from different sources (e.g., thunderstorms).

Estimating return values of extreme wind speeds (at a single observing station) is an application of extreme value theory, and since their accurate estimation is important for building safety, the subject has received substantial attention. For example, [10] compares two estimation methods based on the generalized Pareto distribution (GPD). In [24], an argument in favor of the reverse Weibull distribution over the Gumbel distribution for describing extreme winds is presented, and [8] makes a similar assertion specifically for hurricane generated winds. More recently, [29] presents a new procedure for the estimation of the extreme-value index of the GPD, and then applies the procedure to data from stations in Belgium. This introduction does not present a complete review of the literature on applying extreme value theory to the estimation of extreme wind speeds. A novel feature of the work presented in this chapter is an algorithm for interpolating estimates of N -year return values to geographical coordinates of interest where data have not been observed, that is, an algorithm for creating maps.

The raw data are irregular time series of wind speeds (nominally 3 s gusts) above some threshold at 26 observation stations in Kansas. The time series are irregular because an observation is recorded only when the wind speed crosses a threshold, not at regular intervals such as daily, weekly or monthly. The time series range in number of observations from 4,161 to 29,825 and in time span from approximately

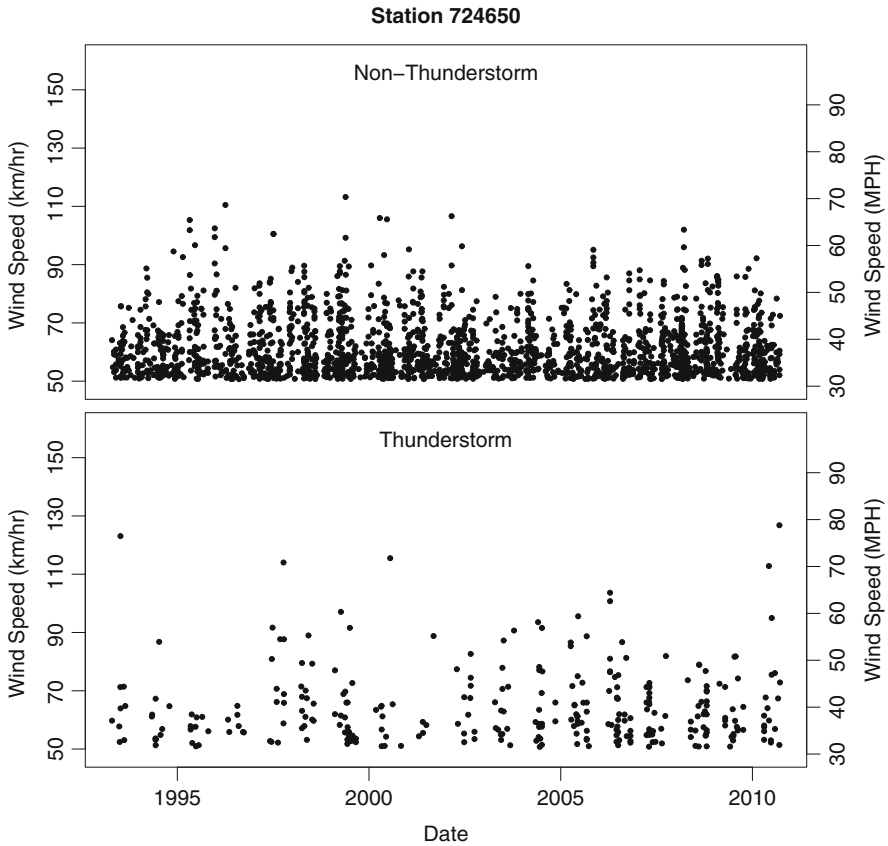


Fig. 1 A random sample of the raw data for observing station 724650 in western Kansas separated by thunderstorm and non-thunderstorm winds

7 years (Nov. 2003–Oct. 2010) to approximately 31 years (Jan. 1980–Dec. 2010). Note that the shortest time span does not correspond to smallest number of observed wind speeds. Also, the dataset used here is a subset of a larger dataset. The subset was chosen such that within a station the threshold used in data collection was constant over time. The details for how the larger dataset was obtained can be found in [13]. An example of the raw data for a single station in western Kansas, number 724650, is shown in Fig. 1. For presentation purposes, Fig. 1 actually shows a random sample of about 10% of the observations from station 724650 so that individual observations of wind speed may be distinguished. Observed wind speeds due to thunderstorms are separated from other observed wind speeds in Fig. 1 because a distinction is drawn between the two types of winds in the remainder of this chapter.

In this chapter, the wind speeds described as non-thunderstorm may also be described as synoptic because they do not include observations from tornadoes,

tropical storms, or hurricanes. It is with this understanding that the term non-thunderstorm is used. The observed wind speeds are separated in this way because thunderstorm and non-thunderstorm winds arise from different underlying environmental conditions. Due to this fact, perhaps not surprisingly, the distribution of extreme thunderstorm wind speeds may exhibit different properties than the distribution of extreme non-thunderstorm wind speeds. This was the case in, for example, [13]. Another important point is that not all of the raw time series will be used to estimate the N -year return value at a given observing station. Only wind speeds above a threshold higher than that used in the data collection process will be considered. The reasons for this as well as a procedure for choosing the high threshold are described in the next section.

The remainder of the chapter proceeds as follows. The section “Estimating N -Year Return Values” describes a procedure for creating maps of estimated N -year return values. The section “Bootstrap for Estimating Uncertainty” presents a bootstrap algorithm for assessing the uncertainty in those estimates. The section “Results” shows the results of applying the estimation and uncertainty assessment procedures to data from 26 observing stations in Kansas. The last section revisits key points from the chapter.

Estimating N -Year Return Values

The procedure described in this chapter for estimating N -year return values of extreme wind speeds proceeds in two stages. In the first stage, a two-dimensional non-homogeneous Poisson process is fitted to extreme wind speeds, for each station separately, using maximum likelihood (ML). This statistical technique for analyzing extreme values was first used in [26] and further considered in [27]; although, as noted in [27] the mathematical foundations for this technique can be found in [9] and [22]. In the second stage, the fitted values of the Poisson process parameters are spatially smoothed using local polynomial regression (also known as LOESS or LOWESS) [3,4]. Lambert conformal conic projections of longitude and latitude (see page 104 of [28]) are the covariates in the regression, which allows interpolation to longitudes and latitudes where data have not been observed. A separate regression is fitted to each parameter. Once the Poisson process parameters are estimated at each longitude and latitude of interest, N -year return values can be calculated and maps drawn. To assess the uncertainty in the estimated N -year return values, point-wise bootstrap [6] confidence intervals are constructed for the estimated N -year return values. Note that [27] also used a two stage analysis to evaluate whether or not extreme rainfall events are becoming more prevalent. The first stage of that analysis is similar to the one presented here, but the second stage spatial analysis is different.

Before continuing note that in the first stage of the procedure, when a non-homogeneous Poisson process model is fitted to each station separately, not all of the raw data from an observing station is used. Instead, only a declustered and

thresholded subset of the raw data is used. The details of how the declustering and thresholding are carried out are found in the sections “Clusters of Extreme Wind Speeds” and “Choice of Thresholds”, respectively.

Poisson Process

Consider the irregular time-series of wind speeds above some threshold, u , at one station. Denote the time-series by $\{(T_i, Y_i)\}_{i=1}^I$ where T_i is the time of the threshold exceedance, Y_i is the wind speed, and I is the number of observations. In [26] it is proposed that $\{(T_i, Y_i)\}_{i=1}^I$ can be modeled as a two-dimensional non-homogeneous Poisson process (henceforth referred to simply as a Poisson process) on $\mathcal{D}_1 = [m_T, M_T] \times [u, \infty)$ with intensity function

$$\lambda_1(t, y) = \frac{1}{\sigma} \left(1 + \zeta \frac{y - \mu}{\sigma} \right)_+^{-1/\zeta - 1} \tag{1}$$

when u is suitably large. The notation m_T and M_T refers to the minimum and maximum observed times, respectively. The notation $(\cdot)_+$ is defined as

$$(\cdot)_+ = \begin{cases} \cdot & \text{if } \cdot > 0 \\ 0 & \text{if } \cdot \leq 0. \end{cases}$$

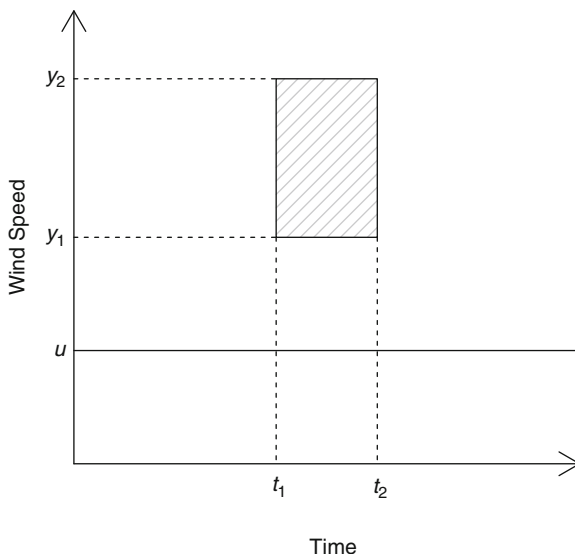
The model parameters, μ and σ do not have nicely self contained interpretations, but the third model parameter ζ can be interpreted as the tail length. For $\zeta \geq 0$, the intensity function has an infinitely long tail (in the wind speed or y direction), and for $\zeta < 0$, the intensity function has a finite tail. Notice also that the right hand side of (1) is free of t , which means that the intensity function is constant in the time or t direction. The justification for this approach lies in convergence theorems, and more details can be found in [26]. Section 2.4 of [18] describes the simplest case of a Poisson process, the one dimensional homogeneous Poisson process, and Chap. 4 of [23] and [21] give general expositions on point processes, of which the Poisson process is one.

To visualize the Poisson process model, consider Fig. 2, which is similar to the figure on page 14 of [27]. In Fig. 2, wind speed is represented by the vertical axis, and time (or date) is represented by the horizontal axis. As previously mentioned, the space on which the Poisson process is defined is $\mathcal{D}_1 = [m_T, M_T] \times [u, \infty)$. The Poisson process model assumes that the number of, (T_i, Y_i) data pairs, that fall within the shaded region in Fig. 2 is Poisson distributed with expected value

$$\Lambda_1 = (t_2 - t_1) \int_{y_1}^{y_2} \frac{1}{\sigma} \left(1 + \zeta \frac{y - \mu}{\sigma} \right)_+^{-1/\zeta - 1} dy$$

which is the amount of volume trapped by the intensity function over the region $(t_1, t_2) \times (y_1, y_2)$.

Fig. 2 Graphical representation of \mathcal{D}_1 on which the Poisson process corresponding to λ_1 is defined. The expected number of points that will fall in the shaded box within \mathcal{D}_1 is Λ_1



One problem with the current formulation of the Poisson process model (as it relates to our data) is that it does not allow the possibility of accounting for the difference between thunderstorm and non-thunderstorm winds. To account for the two types of winds, $\lambda_1(t, y)$ is modified so that the parameters, μ , σ , and ζ are allowed to vary with time. Specifically,

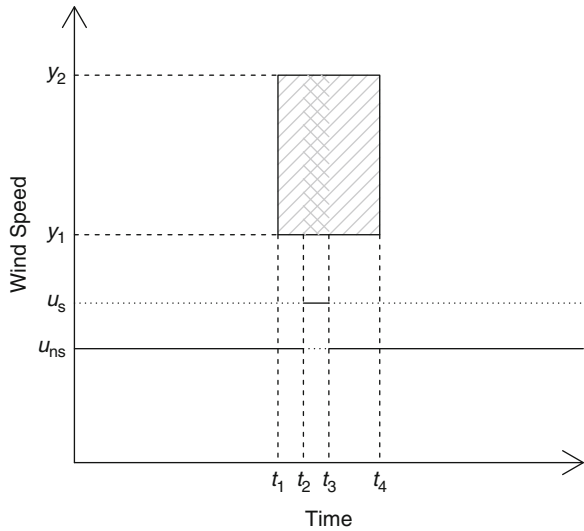
$$\lambda_2(t, y) = \frac{1}{\sigma_t} \left(1 + \zeta_t \frac{y - \mu_t}{\sigma_t} \right)_+^{-1/\zeta_t - 1} \tag{2}$$

Note that the right hand side of (2) relies on t through the parameters, so it is not constant in the time or t direction. While the intensity function in (2) is very general, in this application, it can be simplified because we are only interested in two different types of winds, thunderstorm and non-thunderstorm. Thus, when t is in a thunderstorm period of time, $(\mu_t, \sigma_t, \zeta_t) = (\mu_s, \sigma_s, \zeta_s)$, the thunderstorm set of parameters, and when t is in a non-thunderstorm period of time, $(\mu_t, \sigma_t, \zeta_t) = (\mu_{ns}, \sigma_{ns}, \zeta_{ns})$, the non-thunderstorm set of parameters. Thus we have

$$\lambda_2(t, y) = \begin{cases} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s} \right)_+^{-1/\zeta_s - 1} & \text{for } t \text{ in a thunderstorm period} \\ \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} & \text{for } t \text{ in a non-thunderstorm period} \end{cases} \tag{3}$$

The subscripts s and ns stand for storm and non-storm, respectively. Figure 3 helps to visualize this more general Poisson process model. Figure 3 is similar to Fig. 2;

Fig. 3 Graphical representation of \mathcal{D}_2 on which the Poisson process corresponding to λ_2 is defined. The expected number of points that will fall in the shaded box within \mathcal{D}_2 is Λ_2



however, it adds the complication of a thunderstorm time period within the shaded region. That is, t_1 to t_2 represents a non-thunderstorm time period, t_2 to t_3 represents a thunderstorm time period, and t_3 to t_4 represents a non-thunderstorm time period again. Note from Fig. 3 that thunderstorm and non-thunderstorm time periods are also allowed different thresholds. Since thunderstorm and non-thunderstorm winds are allowed different thresholds, the space on which the Poisson process is defined is a union of two disjoint regions, and it is referred to as $\mathcal{D}_2 = \mathcal{D}_{2,s} \cup \mathcal{D}_{2,ns}$. As before, the Poisson process model assumes that the number of (T_i, Y_i) data pairs that fall within the shaded region in Fig. 3 follows a Poisson distribution, but that the expected value of the Poisson distribution is now

$$\begin{aligned} \Lambda_2 &= (t_2 - t_1) \int_{y_1}^{y_2} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} dy \\ &+ (t_3 - t_2) \int_{y_1}^{y_2} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s} \right)_+^{-1/\zeta_s - 1} dy \\ &+ (t_4 - t_3) \int_{y_1}^{y_2} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} dy. \end{aligned}$$

Note that Λ_2 is calculated by summing the amount of volume trapped by $\lambda_2(t, y)$ over the disjoint regions $(t_1, t_2) \times (y_1, y_2)$, $(t_2, t_3) \times (y_1, y_2)$, and $(t_3, t_4) \times (y_1, y_2)$. The Poisson process model that accounts for both thunderstorm and non-thunderstorm winds is used throughout the remainder of this chapter.

The Likelihood

The Poisson process model described in the section “Poisson Process” provides the means to estimate N -year return values for each station at which data were collected; however, the parameters of the Poisson process must first be estimated, and the method of maximum likelihood (ML) will be used for that task (see for example [2], pp. 315–323).

To use the method of ML, a likelihood, which is typically the joint probability density of the observed data conditional on the parameters of interest, must first be constructed. To do so, note that conditional on observing I points in \mathcal{D}_2 and on the parameters of interest, the I points are independently and identically distributed with the following density function (page 342 of [23]):

$$\begin{aligned}
 f(t, y | \mu_s, \sigma_s, \zeta_s, \mu_{ns}, \sigma_{ns}, \zeta_{ns}) &= \frac{\lambda_2(t, y)}{\iint_{\mathcal{D}_2} \lambda_2(t, y) dt dy} = \\
 &= \begin{cases} \frac{\frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1}}{\iint_{\mathcal{D}_{2,s}} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1} dy dt + \iint_{\mathcal{D}_{2,ns}} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1} dy dt} & \text{thunder } t \\ \frac{\frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1}}{\iint_{\mathcal{D}_{2,s}} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1} dy dt + \iint_{\mathcal{D}_{2,ns}} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1} dy dt} & \text{non-thunder } t. \end{cases} \quad (4)
 \end{aligned}$$

Then, since I follows a Poisson distribution, the joint probability density of the I observations is

$$\begin{aligned}
 &\left(\prod_{i=1}^I f(T_i, Y_i | \mu_s, \sigma_s, \zeta_s, \mu_{ns}, \sigma_{ns}, \zeta_{ns}) \right) \\
 &\times \exp \left\{ - \iint_{\mathcal{D}_{2,s}} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1} dy dt \right. \\
 &\left. - \iint_{\mathcal{D}_{2,ns}} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1} dy dt \right\} \\
 &\times \left(\left[\iint_{\mathcal{D}_{2,s}} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1} dy dt \right. \right. \\
 &\left. \left. + \iint_{\mathcal{D}_{2,ns}} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1} dy dt \right]^I \right) \times \left(\frac{1}{I!} \right) \quad (5)
 \end{aligned}$$

However, (5) is proportional to

$$\begin{aligned}
 L(\mu_{ns}, \sigma_{ns}, \zeta_{nt}, \mu_s, \sigma_s, k_s) &= \left(\prod_{i=1}^I \lambda_2(T_i, Y_i) \right) \\
 &\times \exp \left\{ - \iint_{\mathcal{D}_{2,s}} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s} \right)_+^{-1/\zeta_s - 1} dy dt \right. \\
 &\left. - \iint_{\mathcal{D}_{2,ns}} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} dy dt \right\}
 \end{aligned} \tag{6}$$

which is used as the likelihood. Recall that

$$\lambda_2(t, y) = \begin{cases} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s} \right)_+^{-1/\zeta_s - 1} & \text{for } t \text{ in a thunderstorm period} \\ \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} & \text{for } t \text{ in a non-thunderstorm period} \end{cases}$$

The ML estimate of the Poisson process parameters is the value of $\eta = (\mu_{ns}, \sigma_{ns}, \zeta_{ns}, \mu_s, \sigma_s, \zeta_s)$, say $\hat{\eta} = (\hat{\mu}_{ns}, \hat{\sigma}_{ns}, \hat{\zeta}_{ns}, \hat{\mu}_s, \hat{\sigma}_s, \hat{\zeta}_s)$, that maximizes (6). An analytic expression for $\hat{\eta}$ does not exist, and (6) must be maximized using numerical methods. In particular, we have used the `optim` function in R [19], which implements by default the method described in [17].

Equation for the N -Year Return Value

Recall from the section “Introduction” that the N -year return value is the wind speed, y_N , where the probability that y_N will be exceeded in any 1 year is $1/N$. Without loss of generality, since the number of observations in the region $[0, 1] \times (y_N, \infty)$ is assumed to be distributed as a Poisson random variable with mean $\int_{y_N}^{\infty} \int_0^1 \lambda_2(t, y)$, the N -year return value is the solution to the equation

$$1 - \exp \left\{ - \int_{y_N}^{\infty} \int_0^1 \lambda_2(t, y) dt dy \right\} = \frac{1}{N} \tag{7}$$

for y_N since $1 - \exp \left\{ - \int_{y_N}^{\infty} \int_0^1 \lambda_2(t, y) dt dy \right\}$ is the probability that one or more wind speeds above y_n occur in 1 year. Letting A_s and A_{ns} represent the typical amount of thunderstorm and non-thunderstorm time, respectively in 1 year, (7) becomes

$$\begin{aligned}
 1 - \exp \left\{ - A_s \int_{y_N}^{\infty} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s} \right)_+^{-1/\zeta_s - 1} dy \right. \\
 \left. - A_{ns} \int_{y_N}^{\infty} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}} \right)_+^{-1/\zeta_{ns} - 1} dy \right\} = \frac{1}{N}
 \end{aligned}$$

Since $1 - \exp\{-1/N\}$ is approximately $1/N$ for large N , the simpler

$$A_s \int_{y_N}^{\infty} \frac{1}{\sigma_s} \left(1 + \zeta_s \frac{y - \mu_s}{\sigma_s}\right)_+^{-1/\zeta_s - 1} dy \\ + A_{ns} \int_{y_N}^{\infty} \frac{1}{\sigma_{ns}} \left(1 + \zeta_{ns} \frac{y - \mu_{ns}}{\sigma_{ns}}\right)_+^{-1/\zeta_{ns} - 1} dy = \frac{1}{N}$$

is solved instead. This is also the approach taken in [14]. The N -year return value is then the solution to

$$A_s \left(1 + \zeta_s \frac{y_N - \mu_s}{\sigma_s}\right)^{-1/\zeta_s} + A_{ns} \left(1 + \zeta_{ns} \frac{y_N - \mu_{ns}}{\sigma_{ns}}\right)^{-1/\zeta_{ns}} = \frac{1}{N} \quad (8)$$

for y_N . There is no analytic expression for the N -year return value, but a numerical solver can be used. In particular, we use the R function `uniroot`, which leverages a Fortran subroutine based on an algorithm described in [1]. Estimates of A_s and A_{ns} are discussed in the section on interpolating to spatial coordinates where observations have not been made.

Clusters of Extreme Wind Speeds

Since extreme wind speeds at a particular station are driven by the underlying environmental conditions, they tend to cluster together in time. This calls the Poisson process model into question. More specifically, the independence assumption is less tenable. So that the independence assumption is more tenable, clusters are identified, and only cluster maximums are considered in the fitting process. To identify clusters, consecutive observations are considered to be from different clusters if they are separated by some specified period of time. In this chapter, that period of time for non-thunderstorm winds is taken to be 1 day, and for thunderstorm winds it is taken to be one hour. Thunderstorm and non-thunderstorm winds are de-clustered separately. This is similar to the de-clustering procedure used in [26].

Choice of Thresholds

An important consideration when fitting the Poisson process model for extreme values is the choice of the threshold, and here it will be necessary to choose two thresholds. The threshold pair can have a substantial impact on the estimated N -year return value, due in part to the fact that the thresholds determine which observations in the time series are included (and excluded) for fitting. Sometimes, a high quantile of the observed wind speeds is used. The authors of [14] use the 95th quantile.

Here, a different approach is taken, which is to find the pair of thresholds that produce the best fit, in a sense, of the model to the data. The process begins by

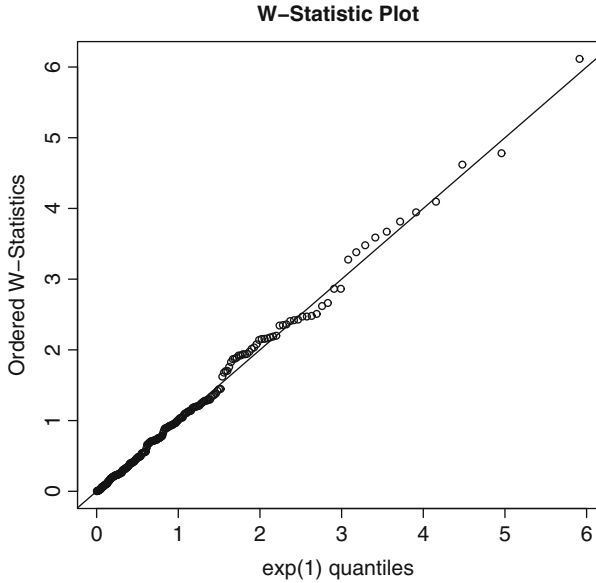


Fig. 4 Q-Q plot of the W_i 's for the threshold pair, thunderstorm = 51.5 and non-thunderstorm = 54.5, at station 724650

specifying a regular grid of threshold pairs. In this case, the observed wind speeds are rounded to integer values, so the grid of potential thresholds are always taken to be of the form $xx.5$. The matter of generating this grid is returned to shortly. For each grid point or pair of potential thresholds, the fit of the model is judged using the W statistic on page 31 of [27]. Specifically,

$$W_i = \frac{1}{\tilde{\zeta}_{T_i}} \log \left\{ 1 + \frac{\tilde{\zeta}_{T_i} Y_i}{\tilde{\sigma}_{T_i} + \tilde{\zeta}_{T_i} (u_{T_i} - \tilde{\mu}_{T_i})} \right\}.$$

If the Poisson process model (and the parameter estimates) were exactly correct, the W_i 's would be independent and identically distributed (*iid*) exponential random variables with mean one. So the pair of thresholds that lead to the W_i 's most closely resembling *iid* exponential random variables with mean one is chosen, and quantile-quantile (Q-Q) plots are used to judge this resemblance. Figure 4 presents such a Q-Q plot of the W_i 's for the best threshold pair at station 724650. The horizontal axis in Fig. 4 depicts quantiles of the exponential mean one distribution, and the vertical axis depicts the ordered W_i 's. Notice in Fig. 4 that the points follow the 45° line very well. This implies that the W_i 's closely resemble *iid* observations from an exponential distribution with mean one.

When selecting a threshold pair, it would be most desirable to visually examine quantile-quantile plots, such as the one in Fig. 4, for all potential threshold pairs in the grid. Since this is impractical, the information contained in the plots must be

numerically summarized. One sensible choice is to calculate the vertical distances from each point to the 45° line and then summarize those distances. We chose to examine the maximum distance, but the mean or median distance could have been used too. Then, the threshold pair that minimizes the distance summary is selected. The selected threshold pair was not very sensitive to the choice of distance summary, but some differences did exist. For station 724650, the selected threshold pair is 82.9 kilometers per hour (km/h) (51.5 miles per hour (MPH)) for thunderstorm time periods and 87.7 km/h (54.5 MPH) for non-thunderstorm time periods. Note that the calculations are done in MPH since the original units of the data are MPH, but the results here and in the following sections are presented in both MPH and km/h. One may be surprised that the selected threshold for thunderstorm winds is smaller than the selected threshold for non-thunderstorm winds. This is not always the case, but it is also not rare. To understand why, consider that a very small proportion of the total observation time is occupied by thunderstorms (for station 724650 less than 1%). So that enough observations are available to provide good estimates of the thunderstorm parameters, a slightly lower threshold may be useful. The de-clustered and thresholded observations that are used to fit the Poisson process model for station 724650 are depicted in Fig. 5. From Fig. 5, it is clear that very high thunderstorm wind speeds are much larger than very high non-thunderstorm winds. For instance, around 1995, station 724650 experienced a thunderstorm wind speed that was just shy of 161 km/h (100 MPH). The largest non-thunderstorm observation is less than 129 km/h (80 MPH).

The first step in the process for selecting a threshold pair is to construct a grid of potential threshold pairs. To do this, an upper and lower bound on the thunderstorm and non-thunderstorm thresholds is chosen. To choose an upper and lower bound for the thresholds, a minimum and maximum average number of observations per year is selected. For the minimum, three, four, or five are reasonable. It is undesirable to use only one or two because an advantage of the Poisson process approach to extreme values over the classical approach is its ability to use more data than just the yearly maxima. In this chapter, the minimum is taken to be four so that the amount of data used for fitting the extreme value model parameters is at least four times what would be used with the classical approach. There is more flexibility in choosing the maximum. It should be large enough so as not to miss good threshold pairs, but not so large that it adds undue computational burden. In this chapter, the maximum is taken to be fifteen, which in this case fits those two qualitative criteria nicely. Then, the upper bound for the thunderstorm threshold is the largest thunderstorm threshold that leads to no less than an average of four thunderstorm observations per year. The lower bound for the thunderstorm threshold is the smallest thunderstorm threshold that leads to no more than an average of fifteen thunderstorm observations per year. The bounds on the non-thunderstorm thresholds are calculated similarly. For station 724650, on which observations spanning almost eighteen years are considered, this leads to a lower and upper bound on the thunderstorm threshold to be 71.6 km/h (44.5 MPH) and 89.3 km/h (55.5 MPH), respectively. The lower and upper bound on the non-thunderstorm threshold for station 724650 is 78.1 km/h (48.5 MPH) and

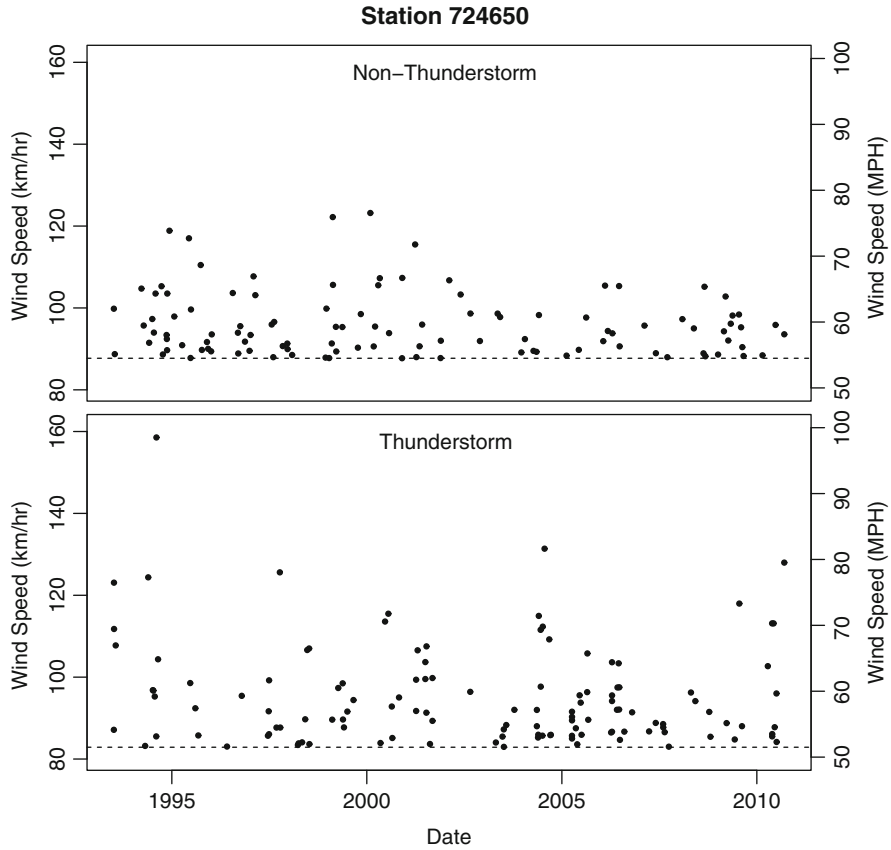


Fig. 5 The thresholded and declustered data for station 724650 that will be used to fit the Poisson process model

89.3 km/h (55.5 MPH), respectively. Recall that the selected thunderstorm threshold was 82.9 km/h (51.5 MPH), which allows 108 total thunderstorm observations, and the selected non-thunderstorm threshold was 87.7 km/h (54.5 MPH), which allows 123 total non-thunderstorm observations.

Interpolating

Recall that the goal of this analysis is the estimation of the N -year return value at an arbitrary longitude (θ) and latitude (ϕ), but so far only the estimation of the N -year return values at observing stations has been discussed. To interpolate to (θ, ϕ) pairs where data have not been collected, LPR, also known as locally weighted scatter plot smoothing (LOESS or LOWESS), is leveraged [3, 4].

Suppose that at (θ, ϕ) there exists some “true” Poisson process producing the irregular time-series of wind speeds. Then, we can denote the parameters of those “true” Poisson processes as $\zeta_s(\theta, \phi)$, $\zeta_{ns}(\theta, \phi)$, $\sigma_s(\theta, \phi)$, $\sigma_{ns}(\theta, \phi)$, $\mu_s(\theta, \phi)$, and $\mu_{ns}(\theta, \phi)$. It is reasonable to assume that the spatially varying functions are smooth, but that their exact form is unknown. Under these assumptions, we may model the estimated Poisson process parameters at station j , for example $\tilde{\mu}_s(j)$, as

$$\tilde{\mu}_s(j) = \mu_s(\theta_j, \phi_j) + \varepsilon_j^{\mu_s}, \quad (9)$$

where the $\varepsilon_j^{\mu_s}$ has zero expectation and variance $\tau_{\mu_s}^2(j)$. Similar spatial models are assumed for the other Poisson process parameters. The error variance is allowed to vary from station to station because the stations have been in use for different periods of time. Thus, differing amounts of uncertainty in the ML estimates is expected. In fact, we assume that error variances are proportional to the estimated variances from the ML estimation procedure, so $\tau_{\mu_s}^2(j) = \tau_{\mu_s}^2 \widehat{\text{var}}(\tilde{\mu}_s(j))$, where $\widehat{\text{var}}(\tilde{\mu}_s(j))$ is the estimated variance of $\tilde{\mu}_s(j)$ from the ML estimation procedure.

The spatial models for the Poisson process parameters can be estimated by LPR, and the function `locfit` in the R package `locfit` [12] is used to produce the estimates of $\mu_s(\theta, \phi)$, $\hat{\mu}_s(\theta, \phi)$ and the other Poisson process parameters at (θ, ϕ) of interest. Another good general reference to LPR, which is written specifically for using the R package `locfit` is [11]. The estimated values of the Poisson process parameters due to LPR can then be transformed to an estimated N -year return value at (θ, ϕ) using (8) and a numerical solver.

To use (8) and a numerical solver to estimate a N -year return value at (θ, ϕ) , a value for A_s and A_{ns} must be available, and the following procedure is used to get a plug-in estimate for them. At each observing station, the average number of thunderstorms per year during the observation period was calculated, and the mean of those averages is taken. This number can be interpreted as the typical number of thunderstorms occurring in one year, and it is multiplied by the typical length of a thunderstorm, which is taken to be one hour, to get A_s . Lastly, if the units of time are days, $A_{ns} = 365 - A_s$.

When using LPR, the value of the smoothing parameter, or bandwidth, must be chosen. For this task, the generalized cross-validation score [5] is utilized. Generally speaking, larger bandwidths will lead to biased predictions, and smaller bandwidths will lead to predictions with higher variability. The generalized cross-validation score attempts to balance the trade-off between bias and variance, and low generalized cross-validation scores are preferred. The `gcv` function of the R package `locfit` is used to calculate the generalized cross-validation scores. A second LPR parameter, the degree of the local polynomial, also has to be chosen, and one is used since it is a typical choice.

Since the Euclidean distance between observations plays an important role in LPR, the (θ, ϕ) pairs are not directly used as covariates. Rather, (θ, ϕ) are transformed to their Lambert Conformal Conic Projections (x, y) using the function `mapproject` from the R package `mapproj` [20]. For our example, the parameters for the projection are `lat0 = 37.044` and `lat1 = 39.549`, the minimum and

maximum of the observed latitudes. Thus, for example, $\mu_s(\theta, \phi)$ is more accurately described as $\mu_s[h_{xy}(\theta, \phi)]$, where $h_{xy}(\cdot, \cdot)$ represents the projection from (θ, ϕ) to (x, y) . However, we continue to use the original notation for brevity.

Summary of the Estimation Procedure

Since the estimation procedure described in this section is complex, a summary is given here so that the reader may see a concise but still complete view of the algorithm.

1. Fit a two-dimensional non-homogeneous Poisson process model to each station separately
 - Create a grid of reasonable threshold pairs
 - Decluster the raw time series for each threshold pair
 - Fit a Poisson process model to the declustered data using ML for each threshold pair
 - Quantify the fit of Poisson process model to the data for each threshold pair using the quantile-quantile plot approach
 - Select the threshold pair that leads to the best fit
 - The final parameter estimates for a station are the ones that correspond to the selected threshold pair
2. Interpolate the Poisson process parameters to all longitude and latitude coordinates of interest separately
 - Use LPR
 - Choose a LPR bandwidth for each parameter using generalized cross-validation
 - Use Lambert conformal conic projections instead of longitude and latitude directly
3. Calculate the N -year return value at each longitude and latitude of interest using the interpolated values

Bootstrap for Estimating Uncertainty

To assess the uncertainty in our estimates of the N -year return values, point-wise bootstrap percentile upper bounds are calculated. The bootstrap was introduced in [6], and a general introduction to it can be found in [7]. This bootstrap algorithm is based on re-sampling the residuals from the fitted spatial models for each Poisson process parameter. The raw residuals are not directly sampled. They are first scaled so that they will have a common variance. After the residuals are re-sampled with replacement, they are again scaled by the appropriate standard error from the ML

estimation procedure so that they have the appropriate variance (since some stations have been in use longer than others). To create a bootstrap data set for each of the Poisson process parameters, the re-sampled and scaled residuals are added to the value of the fitted surface at the appropriate coordinate, (θ_j, ϕ_j) for station j . The bootstrap data sets are then used in the same manner as described above to estimate N -year return values at any coordinate of interest. When this has been done many times, say $N_B = 1,000$, there exists a bootstrap distribution of N -year return values at each coordinate of interest. The upper bound is then some high quantile, maybe 90%, of that bootstrap distribution. More details on the bootstrap algorithm are provided in the appendix.

Results

In this section, the methods of the sections “Estimating N -Year Return Values” and “Bootstrap for Estimating Uncertainty” are applied to data from 26 observing stations in Kansas. A map of Kansas with the locations of all 26 observing stations as well as the locations of some of the larger cities in Kansas is displayed in Fig. 6. The stations are marked by the dots, and the cities are marked by their names. Note that most of the cities have at least one reporting station nearby. For example, Wichita has three. The station that was used as an example in much of this chapter is marked by its number, 724650. The parameter estimates and their ML standard errors for all 26 stations are given in Table 1.

In Figs. 7 and 8, contour maps of the estimated 50-year return values (a) and the point-wise (not simultaneous) 90% upper bounds for those estimates (b) are presented.

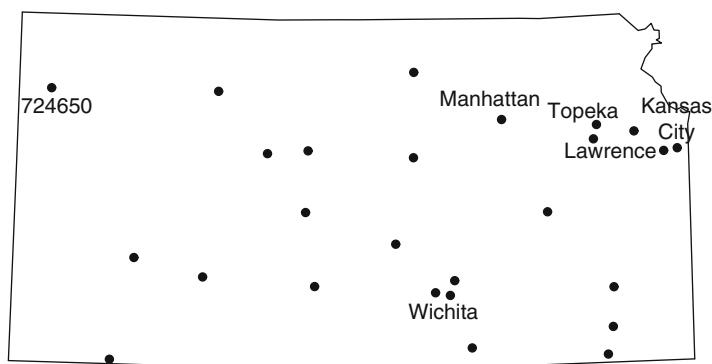


Fig. 6 Map of Kansas that includes all 26 observing stations (marked by *dots*) and some of Kansas’s larger cities (marked by only by *text*)

Table 1 Parameter estimates and standard errors (in parentheses) for the Poisson process parameters for the 26 stations

Station	μ_s	σ_s	ζ_s	μ_{ns}	σ_{ns}	ζ_{ns}
724468	57.302 (1.477)	6.947 (0.996)	0.076 (0.138)	17.338 (5.173)	9.07 (2.768)	-0.11 (0.067)
724475	51.833 (1.081)	5.952 (0.602)	-0.088 (0.099)	30.106 (3.299)	2.567 (1.231)	0.084 (0.1)
724500	54.601 (1.01)	7.361 (0.722)	0.111 (0.079)	33.204 (4.508)	2.391 (1.403)	0.11 (0.11)
724502	55.773 (1.382)	7.751 (0.677)	-0.173 (0.09)	26.389 (4.161)	4.339 (1.896)	0.004 (0.096)
724504	54.996 (1.137)	6.133 (0.863)	-0.046 (0.152)	33.193 (4.596)	3.638 (1.866)	0.002 (0.102)
724505	51.434 (1.069)	7.809 (0.81)	0.1 (0.058)	34.563 (2.278)	1.525 (0.672)	0.15 (0.084)
724506	54.03 (1.356)	7.103 (0.949)	0.099 (0.151)	30.534 (4.122)	3.488 (1.564)	0.05 (0.091)
724507	54.207 (1.188)	6.713 (0.665)	-0.079 (0.078)	29.961 (7.173)	3.569 (2.579)	0.029 (0.134)
724508	55.41 (1.451)	7.397 (0.953)	0.047 (0.116)	18.456 (13.831)	7.341 (5.8)	-0.053 (0.147)
724510	62.305 (1.238)	8.242 (0.725)	-0.045 (0.09)	9.778 (10.367)	16.348 (5.966)	-0.244 (0.07)
724515	61.161 (1.416)	7.997 (0.931)	-0.176 (0.145)	4.17 (23.004)	18.583 (12.677)	-0.244 (0.124)
724516	57.978 (1.463)	7.714 (0.917)	-0.017 (0.118)	-4.673 (20.328)	21.658 (11.1)	-0.256 (0.09)
724517	51.624 (1.065)	6.442 (0.676)	-0.041 (0.059)	33.588 (4.107)	2.301 (1.28)	0.138 (0.108)
724518	52.684 (1.236)	6.829 (0.791)	-0.043 (0.062)	24.134 (4.762)	5.518 (2.134)	-0.022 (0.081)
724519	53.945 (1.121)	5.84 (0.728)	0.013 (0.121)	21.747 (6.069)	6.373 (2.999)	-0.1 (0.097)
724555	61.09 (1.474)	7.53 (0.867)	-0.146 (0.09)	18.308 (4.545)	9.262 (2.476)	-0.102 (0.061)
724556	55.136 (1.091)	7.077 (0.794)	0.02 (0.08)	22.276 (4.154)	7.441 (2.087)	-0.069 (0.062)
724560	55.687 (1.152)	6.319 (0.784)	0.07 (0.114)	13.579 (4.879)	11.615 (2.92)	-0.193 (0.056)
724565	51.61 (1.136)	6.363 (0.722)	0.004 (0.07)	20.184 (4.385)	5.668 (2.029)	0.038 (0.084)
724580	54.054 (1.041)	6.869 (0.761)	0.09 (0.087)	17.714 (7.398)	10.039 (3.95)	-0.178 (0.078)
724585	58.49 (1.298)	8.518 (0.807)	0.005 (0.099)	34.387 (3.235)	3.289 (1.239)	0.097 (0.082)
724586	56.522 (1.108)	6.977 (0.657)	-0.056 (0.074)	35.933 (6.72)	2.529 (2.181)	0.099 (0.163)
724650	60.945 (1.145)	7.607 (0.713)	0.006 (0.093)	28.095 (9.268)	7.493 (4.125)	-0.076 (0.105)
724655	59.858 (1.377)	7.604 (0.838)	-0.009 (0.134)	25.631 (4.199)	6.065 (1.985)	-0.02 (0.072)
745430	48.919 (0.988)	5.577 (0.798)	-0.175 (0.126)	2.427 (42.747)	11.367 (17.632)	-0.032 (0.294)
745431	49.895 (1.716)	7.554 (1.211)	0.102 (0.149)	20.945 (2.856)	4.886 (1.452)	-0.046 (0.068)

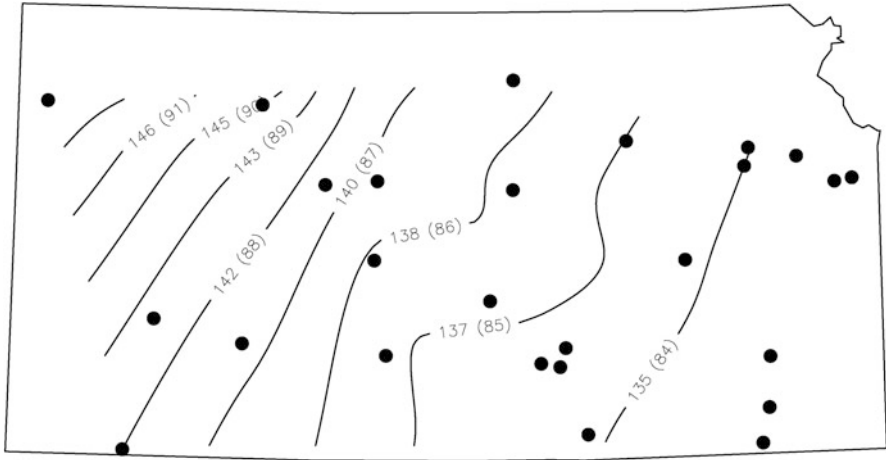


Fig. 7 Contour map of estimated 50-year return values over the convex hull encompassing the 26 observing stations. The results are presented in km/h outside of the parentheses and (MPH) inside of the parentheses

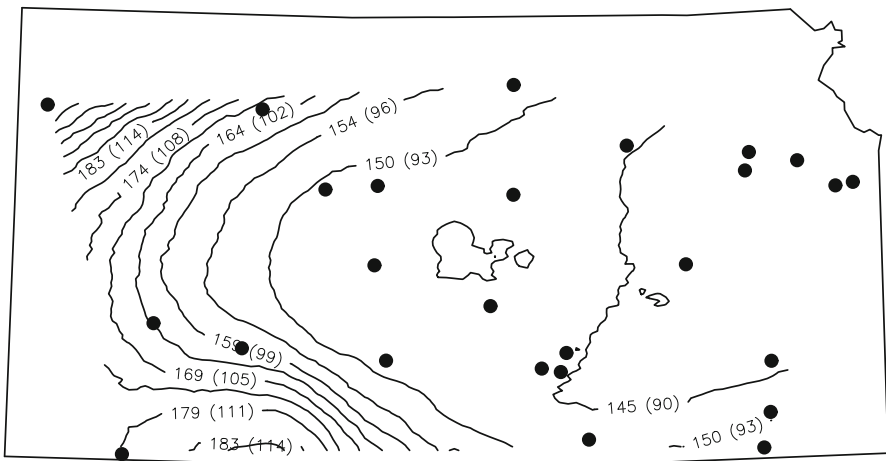


Fig. 8 Contour map of point-wise 90% upper bounds on estimated 50-year return values over the convex hull encompassing the 26 observing stations. The results are presented in km/h outside of the parentheses and (MPH) inside of the parentheses

The contour lines in Figs. 7 and 8 are restricted to the convex hull that encompasses the 26 stations since estimation outside of that convex hull would be extrapolation.

Figure 7 shows that the cities marked in Fig. 6 all have an estimated return value that is less than 137 km/h (85 MPH), and Fig. 8 shows that the 90% upper bound on the estimated 50-year return value is about 145 km/h (90 MPH). Also from Fig. 7,

a positive trend from east to west can be observed, which becomes steeper in the western part of the state. This is indicated by the contour lines being closer together. Figure 8 shows the same positive trend for the 90% point-wise upper bounds, but the increased steepness of the trend in the western part of Kansas is far more pronounced. Thus, the uncertainty in the estimated 50-year return value is larger in the western part of the state. This is due, at least in part, to there being less stations in the western part of the state.

An interesting comparison is between Figs. 7 and 8 and the 50-year return values in the ASCE 7-10 Standard [16]. From the wind speed by location website of the Applied Technology Council [30], the ASCE 7-10 Standard gives the 3-s peak gust 50-year return value at the station 724468 (near Kansas City) and station 724650 to be 145 km/h (90 MPH). Further, [25] states that the ASCE 7-95 Standard [15] gives the 50-year return value to be 145 km/h (90 MPH) for all of Kansas. In Fig. 7, the estimated 50-year return value for station 724468 (eastern Kansas) is below 145 km/h (90 MPH), but for station 724650 (western Kansas), it is slightly above 145 km/h (90 MPH). In Fig. 8, the 90% upper bound for station 724468 (eastern Kansas) is about 145 km/h (90 MPH), but for station 724650 (western Kansas), it is far above 145 km/h (90 MPH). Again the increased uncertainty in the estimated 50-year return values in the western part of the state is at least partly due to the sparseness of stations there. Further, the large uncertainty in the estimated 50-year return value specifically at station 724650 could also be partly due to its position as a vertex on the convex hull encompassing the stations since uncertainty for regression surfaces tends to increase towards the boundaries of the observed independent variables. If perhaps there truly are regional differences as illustrated in Figs. 7 and 8, a constant N -year return value over the entire state, as in [16], would imply risk inconsistency, which is undesirable.

Conclusion

In this chapter, a two-stage procedure for creating a map of N -year return values based on irregular time series of wind speeds at observing stations within the region to be mapped is described. In the first stage, a two-dimensional non-homogeneous Poisson process model is fitted using maximum likelihood to wind speeds above a high threshold at each observing station. In the second stage, local polynomial regression is used to interpolate the parameters of the Poisson process model to any coordinate of interest. The interpolated parameter values are then used to estimate N -year return values at the coordinates of interest. The uncertainty in those estimates is quantified using a bootstrap algorithm to calculate point-wise upper bounds.

The procedure was demonstrated on data from 26 stations within Kansas for $N = 50$, with the results shown in Figs. 7 and 8. The estimated 50-year return values showed a positive east to west trend, and the trend became steeper in the western part of the state. The point-wise upper bounds showed a similar trend; however the increased steepness in the western part of the state was far more pronounced.

The increased uncertainty in the western part of the state is at least partly due to the sparsity of the stations there.

For two of the observing stations (one in the east and one in the west), the results of this procedure were compared to the 3-s peak gust 50-year return values stated in the ASCE 7-10 Standard. The ASCE 7-10 Standard gave both 50-year return values as 145 km/h (90 MPH). The procedure presented in this chapter gives estimates that are below 145 km/h (90 MPH) in the east and slightly above 145 km/h (90 MPH) in the west with 90% upper bounds that are very close to 145 km/h (90 MPH) in the east but far above 145 km/h (90 MPH) in the west.

Acknowledgements This work is part of a project on extreme wind climatology initiated and coordinated by Antonio Possolo of the Information Technology Laboratory, NIST, and Emil Simiu of the Engineering Laboratory, NIST. We thank Antonio Possolo and Emil Simiu for their most helpful conversations and insights on this work, and Emil Simiu for valuable suggestions on improving the style and general presentation of the chapter.

Appendix

The details of the bootstrap algorithm for calculating a percentile upper bound on the N -year return value at a coordinate of interest are presented now. Let $r_j^{\mu_s} = \frac{\tilde{\mu}_s(j) - \hat{\mu}_s(\theta_j, \phi_j)}{\text{se}[\tilde{\mu}_s(j) - \hat{\mu}_s(\theta_j, \phi_j)]}$, where $\text{se}[\tilde{\mu}_s(j) - \hat{\mu}_s(\theta_j, \phi_j)]$ is the estimated standard deviation of $\tilde{\mu}_s(j) - \hat{\mu}_s(\theta_j, \phi_j)$. Similar quantities for the other Poisson process parameters are also defined. Recall that (θ_j, ϕ_j) is the latitude and longitude of station j , so $r_j^{\mu_s}$ is the standardized residual for station j . Now, carry out the following steps.

1. From $(r_1^{\mu_s}, r_2^{\mu_s}, \dots, r_{N_{\text{stations}}}^{\mu_s})$, sample with replacement to obtain $(r_1^{\mu_s,*}, r_2^{\mu_s,*}, \dots, r_{N_{\text{stations}}}^{\mu_s,*})$. Perform a similar procedure for the other Poisson process parameters.
2. Use the re-sampled residuals to create a bootstrap dataset. Specifically, $\tilde{\mu}_s^*(j) = \hat{\mu}_s(\theta_j, \phi_j) + \hat{\tau}_{\mu_s} \sqrt{\widehat{\text{var}}(\tilde{\mu}_s(j))} r_j^{\mu_s,*}$ for $j = 1, 2, \dots, N_{\text{stations}}$ where $\hat{\tau}_{\mu_s}$ is an estimate of τ_{μ_s} . Note that $r_j^{\mu_s,*}$ is scaled by the estimated variance of $\varepsilon_j^{\mu_s}$. Perform a similar procedure for the other Poisson process parameters.
3. Repeat the procedure in the section “Estimating N -Year Return Values” for estimating the N -year return value at all (θ, ϕ) of interest using the bootstrap dataset.
4. Repeat steps 1–3 N_{boot} times to get N_{boot} estimates of the N -year return value at all (θ, ϕ) of interest, and call them $y_N^k(\theta, \phi)$. We take $N_{\text{boot}} = 1,000$.
5. The $(1 - \alpha)100\%$ percentile upper bound for the N -year return value at (θ, ϕ) is then the $(1 - \alpha)100\%$ quantile of $(y_N^1(\theta, \phi), y_N^2(\theta, \phi), \dots, y_N^{N_{\text{boot}}}(\theta, \phi))$.

To use the above procedure, one must be able to calculate $r_j^{\mu_s}$, $\hat{\tau}_{\mu_s}$, and the corresponding quantities from the other Poisson process parameters. The R package `locfit` will calculate $(\hat{\tau}_{\mu_s} r_j^{\mu_s})$ and $\hat{\tau}_{\mu_s}$, which are sufficient to carry out the above algorithm. It is likely that other software packages for LPR will also have the ability to calculate these quantities.

References

1. Brent, R.P.: Algorithms for Minimization Without Derivatives. Prentice-Hall, Englewood Cliffs (1973)
2. Casella, G., Berger, R.L.: Statistical Inference, 2 edn. Duxbury, Pacific Grove (2002)
3. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979)
4. Cleveland, W.S., Devlin, S.J.: Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596–610 (1988)
5. Craven, P., Wahba, T.: Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403 (1979)
6. Efron, B.: Bootstrap methods, another look at the jackknife. *The Ann. Stat.* **7**, 1–26 (1979)
7. Efron, B., Tibshirani, R.: An introduction to the bootstrap, vol. 57. Chapman & Hall/CRC, Boca Raton (1993)
8. Heckert, N., Simiu, E., Whalen, T.: Estimates of hurricane wind speeds by the “peaks over threshold” method. *J. Struct. Eng.* **124**, 445–449 (1998)
9. Leadbetter, M.R., Lindgren, G., Rootzen, H.: Extremes and related properties of random sequences and series. Springer, New York (1983)
10. Lechner, J., Leigh, S., Simiu, E.: Recent approaches to extreme value estimation with application to wind speeds. Part I: The Pickands method. *J. Wind Eng. Ind. Aerodyn.* **41**(1), 509–519 (1992)
11. Loader, C.: Local regression and likelihood. Springer, New York (1999)
12. Loader, C.: locfit: Local regression, likelihood and density estimation. (2010). <http://CRAN.R-project.org/package=locfit>. R package version 1.5-6
13. Lombardo, F.T., Main, J.A., Simiu, E.: Automated extraction and classification of thunderstorm and non-thunderstorm wind data for extreme-value analysis. *J. Wind Eng. Ind. Aerodyn.* **97**, 120–131 (2009)
14. Mannshardt-Shamseldin, E.C., Smith, R.L., Sain, S.R., Mearns, L.O., Cooley, D.: Downscaling extremes: a comparison of extreme value distributions in point-source and gridded precipitation data. *Ann. Appl. Stat.* **4**, 484–502 (2010)
15. American Society of Civil Engineers. Minimum Design Loads for Buildings and Other Structures (1995)
16. American Society of Civil Engineers. Minimum Design Loads for Buildings and Other Structures (2010)
17. Nelder, J.A., Mead, R.: A simplex method for function minimization. *The Comput. J.* **7**, 308–313 (1965)
18. Norris, J.: Markov chains. Cambridge University Press, Cambridge (1998)
19. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2011). <http://www.R-project.org/>. ISBN 3-900051-07-0
20. for R by Ray Brownrigg, D.M.P., Minka, T.P., transition to Plan 9 codebase by Roger Bivand.: mapproj: Map projections (2011). <http://CRAN.R-project.org/package=mapproj>. R package version 1.1-8.3
21. Reiss, R.: A course on point processes. Springer, New York (1993)
22. Resnick, S.I.: Extreme values, point processes and regular variation. Springer, New York (1987)
23. Resnick, S.: Adventures in stochastic processes. Birkhauser, Boston (1992)
24. Simiu, E., Heckert, N.A.: Extreme wind distribution tails: a peaks over threshold approach. *J. Struct. Eng.* **122**(5), 539–547 (1996)
25. Simiu, E., Wilcox, R., Sadek, F., Filliben, J.J.: Wind speeds in the asce 7 standard peak-gust map: an assessment. Technical report, National Institute of Standards and Technology, Gaithersburg (2001)

26. Smith, R.L.: Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Stat. Sci.* **4**, 367–393 (1989)
27. Smith, R.L.: Statistics of extremes, with applications in environment, insurance, and finance. In: Finkenstädt, B., Rootzén, H. (eds.) *Extreme Values in Finance, Telecommunications, and the Environment*, chap. 1, pp. 1–78. Chapman & Hall/CRC (2004)
28. Snyder, J.: Map projections – a working manual. USGPO, Paper 1395 (1987)
29. Van de Vyver, H., Delcloo, A.W.: Stable estimations for extreme wind speeds. an application to Belgium. *Theor. Appl. Climatol.* **105**, 417–429 (2011)
30. Windspeed by Location. <https://www.atcouncil.org/windspeed/>. Accessed 26 Feb 2012

Decision Analysis Methods for Selecting Consumer Services with Attribute Value Uncertainty

Dennis D. Leber and Jeffrey W. Herrmann

Abstract The basic risky decision is defined as a decision for which the outcome of an uncertain event, in addition to the alternative selected, defines the final consequence. Beyond the uncertain event, additional uncertainties can enter a decision including the uncertainty in the attribute values used to assess the decision consequences. When considering the selection of consumer products and services, formal and informal reviews of products and services are often used by consumers to estimate the level of satisfaction that will be received. When developing a decision model based on these data, attribute value uncertainty is often present and should be incorporated. In this chapter, we consider the uncertainty in the attribute values used to describe the possible consequences. We present several approaches to incorporate attribute value uncertainty into the decision analysis for choosing a roofing firm based on customer review data.

Introduction

As the digital age evolves, personal opinions can be found regarding just about anything with only a few clicks of a mouse. While some opinions – fashion, entertainment, political – may be appealing only in the eye of the beholder, others can be very useful for consumers. For example: reviews and ratings from previous

D.D. Leber (✉)

National Institute of Standards and Technology, 100 Bureau Drive MS 8980,
Gaithersburg, MD 20899, USA
e-mail: dennis.leber@nist.gov

J.W. Herrmann

Department of Mechanical Engineering, University of Maryland, 0151B Glenn L. Martin Hall,
Building 088, College Park, MD 20742, USA
e-mail: jwh2@umd.edu

purchasers are shared for products on many retail websites; travel websites often provide a forum for past travelers to convey their experiences and reviews of a hotel or resort; and the dining experiences of past patrons at restaurants nationwide can be found in abundance. Many rely on these reviews to provide, in a qualitative sense, a measure of the satisfaction expected to be received from the product or service being considered.

While these individual qualitative reviews are certainly useful, more quantitative summaries of consumers' opinions are available from non-profit organizations such as Consumers Union and the Center for the Study of Services. These organizations survey their members to gain real world knowledge of consumer products and services. The survey results are analyzed and provided as summary statistics for a variety of performance rating criteria. Consumers can use these results when selecting a product or service provider.

From a decision analysis framework, one might model such a consumer decision as a decision under certainty with either a single or multiple attributes. The products (or service providers) are represented as n alternatives. The m attributes for each alternative are a subset of the available performance rating criteria, and the attribute values are the quantitative values based on the survey results. For each of the n alternatives, the multiple attribute values are combined using a model of the decision-maker's preferences, which yields a decision parameter value that is the basis for the final selection.

The survey attempts to assess the true values of the performance rating criteria. Because the survey is based on limited data, the summary statistics used for the performance rating criterion value is an estimate of the true value and contains uncertainty. We refer to the uncertainty associated with the attribute values as *attribute value uncertainty*. When selecting a product or service provider, the decision-maker should consider how this uncertainty affects the relative desirability of the alternatives. Through an example of selecting a roofing firm based on data published by the Center for the Study of Services, we illustrate a proposed method to incorporate the attribute value uncertainty into the analysis of a decision.

Decision Analysis and Uncertainty

Decision problems can be classified on several dimensions. First, the decision-maker can be either an individual or a group. Second, the number of attributes used to describe the set of consequences can be a single attribute or can consist of multiple attributes. And finally, a decision problem may be classified under conditions of certainty, risk, or uncertainty. These conditions may be defined as follows [1]:

1. *Decisions under certainty*: Each alternative is known to lead invariably to a specific outcome.
2. *Decisions with risk*: Each alternative leads to one of a set of possible outcomes, where each outcome occurs with a probability assumed to be known by the decision-maker. These outcomes may be the result of an uncertain future event, for example.

3. *Decisions under strict uncertainty*: Each alternative leads to one of a set of possible outcomes, though nothing is known or can be stated about the probability of the occurrence of each outcome.

Ron Howard first coined the term decision analysis in a 1966 conference talk [2] where he provided a formal procedure for the analysis of decision problems. Active work in this field had been taking place for more than a decade prior to Howard's introduction of the terminology. Notable contributions during this time include works from von Neumann and Morgenstern [3], Savage [4], and Luce and Raiffa [1]. These works provided the foundation to formally address, through analytical methods, the decision problem for which the consequence of the action cannot be realized until some uncertain event is resolved; i.e., decisions with risk. The method of expected utility theory, first formalized by von Neumann and Morgenstern [3] and later put into practical terms for multiattribute decision analysis in the award winning text of Keeney and Raiffa [5], provides a structured approach to decision analysis when uncertain events exist through the consideration of the probability distributions over the potential outcomes of the uncertain event. Consider the following example of a risky decision for which the method of expected utility theory is applicable. A family is considering one of two outings during the upcoming weekend: a visit to a local museum or attending an outdoor Major League Baseball game. An uncertain event, a weekend rain storm, whose likelihood has been described with some probability by meteorologists, may lead to unfavorable consequences if the family chooses to attend the baseball game.

The driving force behind all decisions is the decision-maker's preference structure. Models to describe one's preference structure include ordinal value functions, measureable value functions, and utility functions. Dyer [6] provides a comprehensive overview of these models, their applications and underlying assumptions, and assessment methods. In brief, ordinal value functions are applicable in decisions under certainty. They lead to a rank ordering of the decision alternatives, but do not indicate magnitude of preference among the alternatives. Measureable value functions, also applicable in decisions under certainty, provide an interval scale of measurement; that is, the decision-maker's strength of preference amongst the alternatives is captured. Finally, utility functions are applicable in decisions with risk. The utility model of one's preference structure not only considers the decision-maker's values of the potential consequences but also incorporates his psychological reactions to taking risks. See Keeney and Raiffa [5], Kirkwood [7], Dyer and Sarin [8], von Winterfeldt and Edwards [9], and Farquhar and Keller [10] for further in-depth discussions of these preference structure models.

A further aspect of uncertainty in decision making was presented in the 1960s by Daniel Ellsberg, best known in the decision analysis community for his now infamous Ellsberg Paradox (see [11] for a well described presentation). The term decision ambiguity in the decision analysis context was first defined by Ellsberg [12] and has since been generalized and elaborated by many. Frisch and Baron [13] present a nice definition: "Ambiguity is uncertainty about probability, created by missing information that is relevant and could be known." Some have used this

idea to challenge the validity of utility theory; particularly as a descriptive theory though most proponents of utility theory argue that the theory was meant only as a normative one [6, 13, 14]. Others have attempted to expand utility theory to include ambiguity (see [15] as an example). In short, decision ambiguity refers to the uncertainty in describing the probability profile of a risky decision. In the previously noted example of the family outing, the decision ambiguity is the uncertainty in the probability of the weekend rainstorm described by the meteorologists.

To summarize, uncertainty in decision making is by no means a new concept. The theory of expected utility addresses the decision problem for which an uncertain future event stands between the decision at hand and the realized consequences. Decision ambiguity considers the uncertainty involved in describing the probability profile of the uncertain event in a risky decision. Other examples of uncertainty in decision making include uncertain decision-maker preference structures [16] and uncertainty in attribute weights [17–19]. Although these methods encompass many aspects of uncertainty in decision making, they all presume that the consequences (described by the attribute values) are precisely defined and neglect any uncertainty that may exist in their assessment. Little work has been published that explicitly considers the uncertainty that may be present in the assessments of the consequences (the attribute value uncertainty). Until now, this work has been limited to the use of the PROMETHEE outranking technique to incorporate attribute value uncertainty, as in Hyde et al. [20] and Zhang et al. [21].

Problem Statement

The consequence associated with any decision is the result of the selected alternative and the outcome of relevant external factors that are outside the control of the decision-maker (e.g., an uncertain future event). To illustrate this perspective, a simple decision may be represented as a decision table (Table 1). The n decision alternatives a_1, a_2, \dots, a_n are the rows in the table. The columns in the table correspond to s_1, s_2, \dots, s_r , the r mutually exclusive and exhaustive possible outcomes of relevant external factors. Associated with each possible outcome is $P(s_j)$, the probability that s_j will be the true outcome. As shown in each cell of the

Table 1 General form of a decision table

		Outcomes			
		s_1	s_2	...	s_r
Decision alternatives	a_1	$x_{1,1,1}, x_{1,1,2}, \dots, x_{1,1,m}$	$x_{1,2,1}, x_{1,2,2}, \dots, x_{1,2,m}$...	$x_{1,r,1}, x_{1,r,2}, \dots, x_{1,r,m}$
	a_2	$x_{2,1,1}, x_{2,1,2}, \dots, x_{2,1,m}$	$x_{2,2,1}, x_{2,2,2}, \dots, x_{2,2,m}$...	$x_{2,r,1}, x_{2,r,2}, \dots, x_{2,r,m}$

	a_n	$x_{n,1,1}, x_{n,1,2}, \dots, x_{n,1,m}$	$x_{n,2,1}, x_{n,2,2}, \dots, x_{n,2,m}$...	$x_{n,r,1}, x_{n,r,2}, \dots, x_{n,r,m}$

table, the consequence that ensues when alternative a_i is selected and s_j is the true outcome is described by m attributes and their associated attribute values x_{ij1} to x_{ijm} .

Table 1 clearly displays the components of a decision: the alternatives, the uncertain possible outcomes, and the resulting consequences described by attribute values. When the decision components are viewed as displayed in Table 1, it becomes evident that uncertainty in expressing the attribute values (that is, the uncertainty in the values of x_{ijk}) is essentially unlike uncertainty about which of the set of possible outcomes, s_1, s_2, \dots, s_r , will occur (risky decision) and uncertainty in defining the probability of each outcome, $P(s_j)$ (decision ambiguity).

While it may be conceivable to model the attribute value uncertainty as an uncertain event in a risky decision, we choose to maintain a decision model that distinguishes the attribute value uncertainty as a unique component of uncertainty. The reason is that a decision-maker can control, to some extent, the amount of uncertainty in an estimate of the true value of an attribute by varying the amount of information observed in its assessment whereas the outcome of a future uncertain event cannot be controlled in this same manner.

Decision-makers often consider decision alternatives that have consequences that are described by uncertain attributes. If a decision problem includes attributes whose values are determined by means of sampling and measurement, attribute value uncertainty exists. For example, the listed fuel mileage of a new car being considered is only an estimate of the true value based on sampling and experimental evaluations which include measurement. When these attribute values are provided only as point values, the decision-maker must move forward under the assumptions that the values are accurate and that the level of uncertainty associated with each alternative is equivalent. Although the value of a new car's fuel mileage may be an innocent example, the Department of Homeland Security's selection of a radiation detection system to be installed at airports based on estimated system performance parameters is a much more serious matter. Consider also the decisions that depend upon the results of the 2010 United States census. The allocation of congressional seats and federal funding will be decided based on the estimated population within each congressional district. There is undeniably uncertainty in these estimates, and the uncertainties are not equivalent from district to district.

The selection of a product or service provider may be modeled as a multiattribute decision under certainty. That is, there are no uncertain future events that stand between the decision at hand and the realized consequences. In this case there is only a single outcome. The alternatives a_1, a_2, \dots, a_n are the n products or service providers being considered. Some of the relevant attributes have attribute value uncertainty. Because the model is a decision under certainty, either a multiattribute ordinal value function or a multiattribute measurable value function may be used to represent the preference structure. The decision-maker's goal in this situation is to select the alternative that, given his preferences, maximizes his satisfaction in the presence of attribute value uncertainty. We shall propose an approach to incorporate the attribute value uncertainty into the decision model and then consider various rules for evaluating the decision-maker's satisfaction with each alternative.

In this situation, the decision-maker faces the risk of selecting an alternative that is not the best one, which could be identified if no attribute value uncertainty existed. The decision-maker may have to make a tradeoff, as in other settings involving risk, between alternatives whose performance is described to range from very well to poor (that is, there is a large amount of uncertainty about their performance) and other alternatives whose performance is described as neither very well nor poor (that is, there is less uncertainty about their performance). The proposed methods should help the decision-maker understand this tradeoff and make a better decision.

Approach

This section describes the approach that we will use to identify the best of a set of alternatives that have attribute value uncertainty. For any alternative, the point estimates for each attribute can be used to evaluate the decision model. The alternative with the largest resulting decision parameter value (e.g., value or measurable value in a decision under certainty, utility in a risky decision [6]) would be considered the alternative most fitting given the decision-maker's preferences. We will call this the *expected value approach*. Because this approach fails to consider the attribute value uncertainty, it may fail to select the alternative that maximizes the decision-maker's satisfaction. Thus, we propose the following approach, which augments the expected value approach by incorporating the attribute value uncertainty as follows:

1. Identify and develop the alternatives a_1, a_2, \dots, a_n and the attributes.
2. Model the uncertainty of each attribute value for each alternative. Let $F_{ij}(x)$ be the probability distribution for alternative i 's value for attribute j , $i = 1, \dots, n$, $j = 1, \dots, m$.
3. Randomly sample the attribute values based on the associated uncertainty models to generate R realizations for each alternative. Each realization has a single value for each of the relevant attributes. Let x_{ijr} be alternative i 's value for attribute j in realization r , $i = 1, \dots, n$, $r = 1, \dots, R$.
4. Define the multiattribute decision model based on the decision-maker's preference structure and the realizations of the attribute values. This includes defining the individual value (or utility) functions and attribute weights.
5. Propagate the attribute value uncertainty through the multiattribute decision model and onto the decision parameter. That is, for each alternative and each of its R realizations, calculate the corresponding decision parameter value. The result is a distribution of R decision parameter values for each alternative. Let y_{ir} be alternative i 's value for the decision parameter in realization r , $i = 1, \dots, n$, $r = 1, \dots, R$.
6. Use a decision rule to identify the most desirable alternative based upon distributions of the decision parameter values.

By appropriately adjusting the definition (Step 4) and evaluation (Step 5) of the decision model, this approach can be used in both decisions under certainty and

decisions with risk, including risky decisions with decision ambiguity. As Steps 1 and 4 above are the basis for developing any decision model [5] our discussion in the following Sections will focus on approaches to model the attribute value uncertainty (Step 2) and selecting an alternative based on a collection of decision parameter distributions (Step 6).

Modeling Attribute Value Uncertainty

The ideal attribute value input to the decision model is the true, but often unknown, value. As previously discussed, when attribute values are obtained based on sampling, such as surveys or measurements, the value obtained is merely an estimate of the true attribute value. This estimate contains some uncertainty that depends upon the experimental technique used to estimate the value. In this section, we describe several approaches to modeling this uncertainty. Our general decision analysis approach to incorporate attribute value uncertainty can be used with any of these approaches for modeling attribute value uncertainty.

Uncertainty and its assessment has become a popular topic in recent years. Lindley [22] suggests the reason for the peak in interest is that the rules for assessing and applying uncertainty are now understood and that past tendencies of suppressing uncertainties are no longer necessary. Of the various methods that could be leveraged to model attribute value uncertainty, we will discuss two approaches: a bootstrap approach and a Bayesian approach.

A Bootstrap Method for Modeling Attribute Value Uncertainty

The non-parametric bootstrap method relies upon resampling of the observed data to model the attribute value uncertainty. Introduced by Efron [23], the bootstrap is “a computer-based method for assigning measures of accuracy to statistical estimates” [24].

Given observed data x_1, x_2, \dots, x_n , a typical application of the non-parametric bootstrap technique involves generating a sample of size n with replacement from the observed empirical distribution. Denoted by \mathbf{x}^* , this sample is called a bootstrap sample. From the bootstrap sample, we compute the value of the parameter of interest, denoted by θ^* . We repeat this process b times to create an approximation for the distribution of θ^* and obtain statistical properties such as the standard error, which are directly related to the parameters of the distribution that underlies the original observations.

A number of variants of the bootstrap exist beyond the non-parametric bootstrap method such as the parametric bootstrap and the Bayesian bootstrap, each of which could also be used to develop a model of the attribute value uncertainty. These variants utilize the same general resampling approach to create an approximation for the distribution of θ^* , though the Bayesian bootstrap uses a posterior probability

distribution for resampling the observed data rather than the uniform distribution used in the non-parametric bootstrap resampling. The parametric bootstrap samples from an assumed parametric distribution with parameter values estimated based upon the observed data. Chernick [25] provides a summary of these and other bootstrap techniques as well as a variety of applications.

A Bayesian Model of Attribute Value Uncertainty

Another alternative in developing a model of the attribute value uncertainty is to leverage the Bayesian paradigm of inference, which describes where the likely attribute values are found using the posterior probability distribution [26]. Generally speaking, an initial degree of belief, described in terms of a probability distribution called the prior, is updated using Bayes' Theorem when new data are observed to produce a new degree of belief called the posterior distribution.

This approach allows for probabilities to be associated with the unknown parameters. That is, the resulting posterior probability distribution describes what is currently known about the parameters, where the probabilities are interpreted as representing the degree of belief that given values of the parameter is the true value [27].

The posterior distribution provides a method to model one's knowledge of the true value of the attribute. This model captures the uncertainty in the attribute value estimate provided by the sampling or measurements.

Selection of an Alternative

Traditional decision analysis approaches clearly identify the most desirable alternative. This property should not be lost when expanding the model to be more comprehensive by including attribute value uncertainty. The result of propagating uncertainty is a set of decision parameter values that are described by distributions. Thus, selecting an alternative changes from a simple ordering exercise to a comparison of distributions. This section discusses three approaches to compare the resulting decision parameter distributions: Rank 1, Stochastic Dominance, and Majority Judgment.

Rank 1

In each realization, each alternative has one value for the decision parameter. If we consider the realizations one at a time and examine the decision parameters for all of the alternatives in that realization, then the alternatives can be ranked by the decision parameter, and the most desirable alternative (the one ranked first) can be identified. (If multiple alternatives tie for first in a realization, all of those so tied are considered

as ranked first.) The number of realizations in which an alternative is ranked first (its *rank 1 value*) describes the relative desirability of that alternative. An alternative's rank 1 value can vary from 0 (it is never ranked first) to R (it ranked first in every realization). We use this value in the decision rule that selects the alternative with the greatest rank 1 value.

Stochastic Dominance

Our second approach builds upon the concept of stochastic dominance for comparing distributions. In the following discussion Y_i and Y_j represent the decision parameters for alternatives i and j respectively. The distributions of these parameters are the ones generated by the R realizations.

Hadar and Russell [28] discuss stochastic dominance as an approach to predicting a decision-maker's choice between two uncertain events without knowledge of the decision-maker's utility function. They define two types of stochastic dominance: first-degree stochastic dominance and second-degree stochastic dominance which are presented below.

First-degree stochastic dominance: Y_i stochastically dominates Y_j in the *first degree* if and only if

$$P[Y_i \leq y] \leq P[Y_j \leq y] \quad \forall y \quad (1)$$

That is, the value of the cumulative distribution for Y_i never exceeds that of Y_j for all $y \in Y$.

Second-degree stochastic dominance: When the support of Y_i and Y_j is contained within the closed interval $[a, b]$, Y_i stochastically dominates Y_j in the *second degree* if and only if

$$\int_a^t P[Y_i \leq y] dy \leq \int_a^t P[Y_j \leq y] dy \quad \forall t \in [a, b] \quad (2)$$

That is, the area under the cumulative distribution for Y_i is less than or equal to that of Y_j for all $t \in [a, b]$.

First-degree stochastic dominance is relevant in the absence of any restrictions on the unknown utility function other than monotonicity. Second-degree stochastic dominance is more restrictive in that the results apply only when the unknown utility functions are concave, indicating a risk-averse decision-maker. Under these restrictions, if Y_i is found to stochastically dominate Y_j in either the first or second degree then alternative i is preferred to alternative j because alternative i will have a greater expected utility.

If a single Y_i is identified to stochastically dominate (first- or second-degree) Y_j , for all j , $i \neq j$, and, in at least one case, the inequality in Eqs. 1 or 2 is found to be a strict inequality then Hadar and Russell have shown that alternative i can be selected with few underlying assumptions.

We use the idea of stochastic dominance as a decision rule to select the alternative with a set of decision parameter values that stochastically dominates all others. It should be noted, however, that this rule may not produce a solution, and thus an alternative would not be identified for selection.

In particular, consider the decision parameter values for alternatives i and j . Both are sets of R values generated as discussed in Step 5 of the approach. Alternative i dominates alternative j based upon the ideas of first-degree stochastic dominance if, for all values y , the number of values of the decision parameter Y_i that are not greater than y is less than or equal to the number of values of the decision parameter Y_j that are not greater than y .

Let $Z_i = \{y_{i[1]}, y_{i[2]}, \dots, y_{i[R]}\}$ be the ordered set of the R decision parameter values generated as discussed in Step 5 of the approach for alternative i where $y_{i[1]} \leq y_{i[2]} \leq \dots \leq y_{i[R]}$. Let $f_i(y)$ be the number of decision parameter values in Z_i that are less than or equal to y . Note that this is a step function that increases at each value in the set Z_i . Let a and b be the lower bound and the upper bound on the decision parameter values across all of the alternatives. Alternative i dominates alternative j based upon the ideas of second-degree stochastic dominance if the following condition holds:

$$\int_a^t f_i(y)dy \leq \int_a^t f_j(y)dy \quad \forall t \in [a, b] \quad (3)$$

Because $f_i(y)$ and $f_j(y)$ are step functions, it is easy to calculate these integrals for any value of t , and this condition holds for all $t \in [a, b]$ if it holds for all $t \in Z_i \cup Z_j$.

Majority Judgment

By considering the decision parameter value resulting from each of the R realizations as a score assigned by an individual judge or voter, the problem of selecting an alternative based on distributions of decision parameter values may be viewed as one of social choice. A consensus value for each alternative that appropriately represents the message of all judges is sought in comparing and selecting the most desirable alternative. While many models of social choice exist, we consider the method of Majority Judgment.

In an attempt to identify a model of social choice that overcomes the shortcomings displayed by traditional social choice models such as the Borda and Condorcet methods, Balinski and Laraki [29, 30] propose the method of Majority Judgment. The Majority Judgment method relies upon the middlemost interval to identify a social grading function that has desirable functional properties, provides protection against outcome manipulation by individual voters or judges, and overcomes many of the shortcomings of traditional social choice models. When considering $y_{i[1]}, y_{i[2]}, \dots, y_{i[R]}$ ordered scores for a given alternative i , the *majority-grade* is defined to be the median score, $y_{i[(R+1)/2]}$, when R is odd and the lower bound of the middlemost interval, $y_{i[R/2]}$, when R is even where

$y_{i[1]} \leq y_{i[2]} \leq \dots \leq y_{i[R]}$. The Majority Judgment method identifies the alternative with the largest majority-grade as the most desirable alternative in the social choice context. If multiple alternatives have the same largest majority-grade, then a single majority-grade value is removed from the set of scores for each alternative in the tie, and the majority-grade of the new distributions are calculated. If a tie again occurs, this process is repeated until a single alternative has the largest majority-grade. The method extends this concept to provide a complete rank-ordering termed the *majority-ranking*.

As the majority-grade is defined based upon the middle-most interval, it emphasizes the significance of place in order rather than magnitude. That is, it is robust against extreme scores. As the goal of the decision problem at hand is to identify the most desirable alternative given the set of considered alternatives this trait makes the method of Majority Judgment an attractive rule for selecting an alternative. Further, the majority-ranking provides a ranking between any two alternatives that are dependent upon the grades of only those two alternatives. In other words, the majority-ranking is independent of irrelevant alternatives (Arrow's IIA).

To select a decision alternative based on the Majority Judgment method where the decision parameters are described by distributions, a majority-grade is computed for each alternative by considering the decision parameter value resulting from each of the R realizations as an individual score. Specifically, the median (if R is odd) or the lower bound of the middlemost interval (if R is even) of the distribution of decision parameter values is computed for each alternative. The alternative with the largest majority-grade is then identified as the most desirable alternative. If a tie exists, the tie-breaking procedure defined by the method of Majority Judgment is used to identify the single most desirable alternative.

Application

When homeowners require a repairman or other services, they often seek reviews and recommendations for potential service providers. One source for such information is the Center for the Study of Services, who publishes quarterly periodicals in several major metropolitan areas. These periodicals provide ratings for various consumer services. The Spring/Summer 2011 edition of the Washington Consumers' Checkbook [31] provides an extensive review of roofing firms in the Washington, D.C., metropolitan area. We will consider the problem of selecting a roofing firm using the data in the Washington Consumers' Checkbook to illustrate the proposed approaches for making decisions in the presence of attribute value uncertainty. As the purpose of this demonstration is to illustrate the application of the decision analysis method rather than to endorse any particular service provider, the roofing firm names as provided by the Washington Consumers' Checkbook review have been replaced by numeric ID codes.

The Washington Consumers' Checkbook review includes ten performance rating criteria for each of 94 roofing firms. The results of the review were obtained through a survey of the organization's members. The ten performance rating criteria are:

1. Work performed properly on first attempt
2. Began and completed work promptly
3. Provided cost information early
4. Neatness of work
5. Expert advice on service options and costs
6. Overall performance
7. Percentage of customers rating a firm "adequate" or "superior" for "overall performance"
8. Number of complaints (and rate) filed with local government
9. Number of complaints (and rate) filed with the Better Business Bureau
10. Percent of \$5,000 job the firm allows the customer to pay upon completion

For each of criteria 1 through 6, the measure provided is the proportion of customers surveyed who rated the firm as "superior". For criteria 8 and 9, the complaint rate is the ratio of the number of complaints filed to the number of full-time employees performing residential work. This measure is an attempt to account for the exposure of a company, whereas a larger company that performs more work experiences greater exposure to incur complaints. In addition to the performance rating criteria, the number of survey responses is noted for each roofing firm.

Roofing Firm Decision Model

The decision of selecting a roofing firm from the firms reviewed in the Washington Consumers' Checkbook was modeled as a multiattribute decision with certainty. There are no uncertain events in this decision model, but uncertainty is prevalent in the review's estimates of the performance criteria values that are attributes of the decision model.

The alternatives considered in this decision model are the roofing firms. Of the 94 firms included in the survey, seven were removed from consideration due to incomplete data, so $n = 87$ decision alternatives remained.

For this demonstration, we considered the following $m = 4$ attributes:

- X_1 : Work performed properly on first attempt
- X_2 : Began and completed work promptly
- X_3 : Neatness of work
- X_4 : Percent of \$5,000 job the firm allows the customer to pay upon completion

The survey results for the performance rating criteria were used as estimated values of the attributes included in the decision model. The estimates for attributes X_1 , X_2 , X_3 are provided as the proportion of customers surveyed who rated the firm "superior" for each performance criteria. These attributes are random variables consisting of a collection of Bernoulli trials: the performance criterion was rated by

Table 2 Summary statistics for the distribution of data across the 87 firms considered in the decision model

	Mean	Std dev	Min	Median	Max
Number of survey responses	54.17	66.30	10	29	390
X ₁ : Work performed properly on first attempt	0.74	0.16	0.23	0.79	1.00
X ₂ : Began and completed work promptly	0.74	0.16	0.28	0.77	1.00
X ₃ : Neatness of work	0.76	0.16	0.27	0.79	1.00
X ₄ : Percent of job firm allows paid after completion	0.77	0.19	0.33	0.67	1.00

each survey respondent as either superior or not. Thus, for each of the *i* roofing firms (*i* = 1, 2, ..., 87), each of the *X_j* performance criteria (*j* = 1, 2, 3) can be described by a binomial random variable with parameters *p_{ij}* and *k_i*, where *p_{ij}* is the proportion of the *k_i* survey respondents who provided a rating of “superior” for the performance criteria. For these attributes, a larger value is preferred.

Attribute *X₄* is not random; it is provided by the roofing firm. The attribute is the percentage of a \$5,000 job that the firm will allow the customer to pay upon completion of the job. The value is considered to be a constant for each firm. Larger values for *X₄* are preferred because, if the value is small, the customer must pay more upfront, which increases the customer’s financial risk. There were seven firms (of the 94 firms included in the survey) for which no value for this attribute was obtained. These seven firms were removed from the alternatives considered in the decision analysis model. Summary statistics of the distribution for the four attributes and the number of survey responses across the 87 firms considered are provided in Table 2.

A multiattribute measurable value function was used to represent the decision-maker’s preference structure. The preference model used in this demonstration represent the preference structure of the author. We assume the preference structure is such that attributes *X₁*, *X₂*, *X₃*, *X₄* are mutually preference independent and mutually difference independent. Therefore, the multiattribute measurable value function can be represented by the sum of single attribute measurable value functions [8] as displayed in Eq. 4.

$$v(x_1, x_2, x_3, x_4) = \sum_{i=1}^4 \lambda_i v_i(x_i) \tag{4}$$

Here $\sum_{i=1}^4 \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ and the individual measurable value functions *v_i(x_i)* are scaled such that, for *x_i^{*}*, the most preferred outcome, *v_i(x_i^{*})* = 1 and, for *x_i⁰*, the least preferred outcome, *v_i(x_i⁰)* = 0.

Expected Value Approach

One may employ the performance ratings provided by the survey in conjunction with Eq. 4 to evaluate the multiattribute measurable value model. As described

Table 3 Additive individual measureable value functions and weights for the expected value approach

Attribute	$v_i(x_i)$	λ_i
X_1 : Work performed properly on first attempt	$v_1(x_1) = -\frac{1}{3.634} \left(1 - e^{(x_1-0.23)/0.502} \right)$	0.476
X_2 : Began and completed work promptly	$v_2(x_2) = 1 - e^{-(x_2-0.28)/0.866}$	0.190
X_3 : Neatness of work	$v_3(x_3) = x_3/0.73 - 0.37$	0.286
X_4 : Percent of price for a \$5,000 job the firm allows the customer to pay upon completion of job	$v_4(x_4) = \frac{1}{0.909} \left(1 - e^{-(x_4-0.33)/0.280} \right)$	0.048

Table 4 Attribute values and resulting decision parameter value for the top 10 roofing firm alternatives based on the expected value approach

Roofing firm ID	n	x_1	x_2	x_3	x_4	Value
Firm 29	24	1.00	0.92	0.96	1.00	0.9837
Firm 84	82	0.99	0.99	0.99	1.00	0.9835
Firm 57	23	0.95	0.96	0.96	1.00	0.9263
Firm 28	54	0.96	0.92	0.92	1.00	0.9216
Firm 90	36	0.94	0.88	0.97	0.95	0.9183
Firm 8	347	0.95	0.94	0.93	1.00	0.9145
Firm 91	49	0.96	0.73	0.89	1.00	0.9089
Firm 93	13	0.92	1.00	0.92	0.67	0.8679
Firm 71	89	0.93	0.74	0.84	1.00	0.8570
Firm 35	23	0.91	0.83	0.91	0.66	0.8530

by Keeney and Raiffa [5], the alternative with the largest resulting value would be considered to be the alternative most fitting given the decision-maker’s preferences.

The individual measureable value functions $v_i(x_i)$, $i = 1, 2, 3, 4$, that were considered in the value analysis were developed by utilizing an augmentation to the midvalue splitting technique that leverages an analytical exponential form [7] based on the attribute value ranges displayed in Table 2. The swing weighting procedure [32] was used to develop the associated weights λ_i , $i = 1, 2, 3, 4$, for each individual measureable value function. The individual measureable value functions and associated weights are provided in Table 3.

Based on these defined individual measureable functions and associated weights, Eq. 4 was evaluated for each alternative. Table 4 displays the top 10 alternatives resulting from the analysis utilizing the expected value approach. All results are displayed graphically in Fig. 1. Roofing Firm 29, whose value equals 0.9837, is the most desirable alternative. This firm is followed closely by Roofing Firm 84, whose value equals 0.9835.

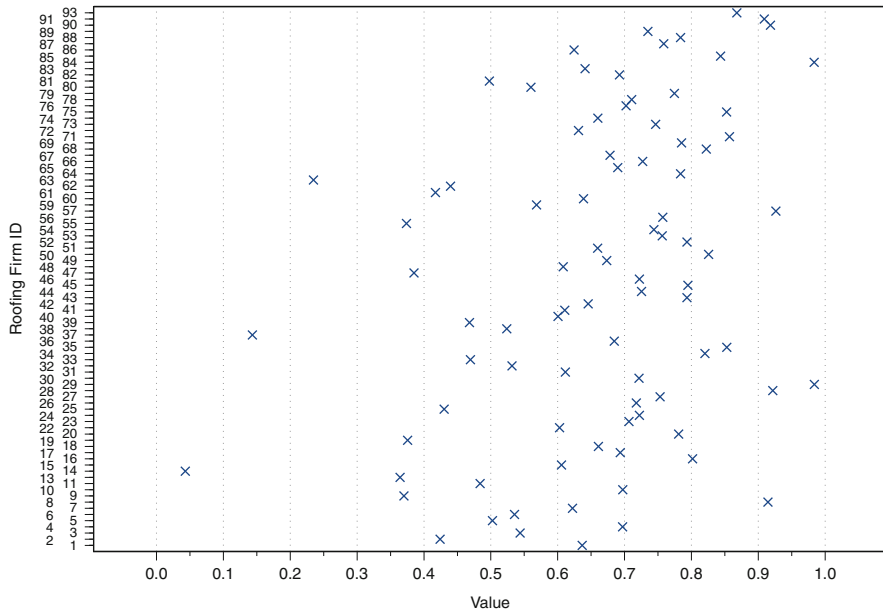


Fig. 1 Decision model results using the expected value approach

Incorporating Attribute Value Uncertainty

To describe the uncertainty in attributes X_1 , X_2 , and X_3 we chose to use a Bayesian approach. For each alternative and each of these attributes, we began with the assumption – or prior knowledge – that the true value of the attribute lies between 0 and 1 with equal likelihood. This is represented by the *Uniform* (0, 1) prior distribution, which is equivalent to a *Beta* (1, 1) distribution. Observations from a *Binomial* (k_i , p_{ij}) distribution were used to update the prior distribution. In this case, the observations were the estimates of the X_j performance criteria ($j = 1, 2, 3$) obtained by the Washington Consumers’ Checkbook review for the $i = 1, 2, \dots, 87$ roofing firms. Given this new information along with the prior distribution, the knowledge about the unknown parameter p was updated to create a posterior distribution. Because the *Beta* (α , β) distribution is the conjugate prior to the *Binomial* (n , p) distribution, the posterior distribution is the *Beta*($1 + k_i p_{ij}$, $1 + k_i(1 - p_{ij})$) distribution. This posterior distribution describes the uncertainty in each attribute for each alternative.

Given the posterior distributions for each attribute for each alternative, we drew $R = 1000$ random samples from each of these distributions. Summary statistics for these random realizations for each attribute across all alternatives are presented in Table 5.

Table 5 Summary statistics for the 1000 random realizations across all alternatives

	Mean	Std dev	Min	Median	Max
X ₁ : Work performed properly on first attempt	0.73	0.17	0.027	0.76	1.00
X ₂ : Began and completed work promptly	0.73	0.17	0.050	0.76	1.00
X ₃ : Neatness of work	0.74	0.17	0.049	0.78	1.00
X ₄ : Percent of job firm allows paid after completion	0.77	0.19	0.33	0.67	1.00

Table 6 Additive individual measureable value functions and weights when considering attribute value uncertainty

Attribute	$v_i(x_i)$	λ_i
X ₁ : Work performed properly on first attempt	$v_1(x_1) = -\frac{1}{4.53} \left(1 - e^{(x_1 - 0.027)/0.570} \right)$	0.476
X ₂ : Began and completed work promptly		0.190
X ₃ : Neatness of work	$v_3(x_3) = x_3/0.951 - 0.05$	0.286
X ₄ : Percent of price for a \$5,000 job the firm allows the customer to pay upon completion of job	$v_4(x_4) = \frac{1}{0.909} \left(1 - e^{-(x_4 - 0.33)/0.280} \right)$	0.048

Based on the distributions of the random realizations for each attribute across all alternatives, summarized by the ranges displayed in Table 5, the individual measureable value functions $v_i(x_i)$, $i = 1, 2, 3, 4$, and associated weights λ_i , $i = 1, 2, 3, 4$, were redefined by again using an augmentation to the midvalue splitting technique and the swing weighting procedure. The redefined individual measureable value functions and weights are provided in Table 6.

Provided these defined individual measureable functions and associated weights, Eq. 4 was evaluated for each alternative for each of the 1000 random realizations. The result is a distribution of 1000 overall decision parameter values for each roofing firm.

When analyzing the resulting 87 distributions of decision parameter values, we first found the minimum value of every alternative’s decision parameter and identified the greatest of these minimum values. We then determined that 59 alternatives were dominated in the following way: for each of these 59 alternatives, the maximum value of its decision parameter was less than the greatest minimum value. This left 28 non-dominated alternatives. The non-dominated alternatives and their associated value distributions are displayed in Fig. 2.

Results

Given the distributions of decision parameter values for the non-dominated roofing firms, we applied the Rank 1, Stochastic Dominance, and Majority Judgment decision rules. Table 7 lists the Rank 1 and Majority Judgment results for the six firms that had the most value in the expected value approach (see Table 4). The

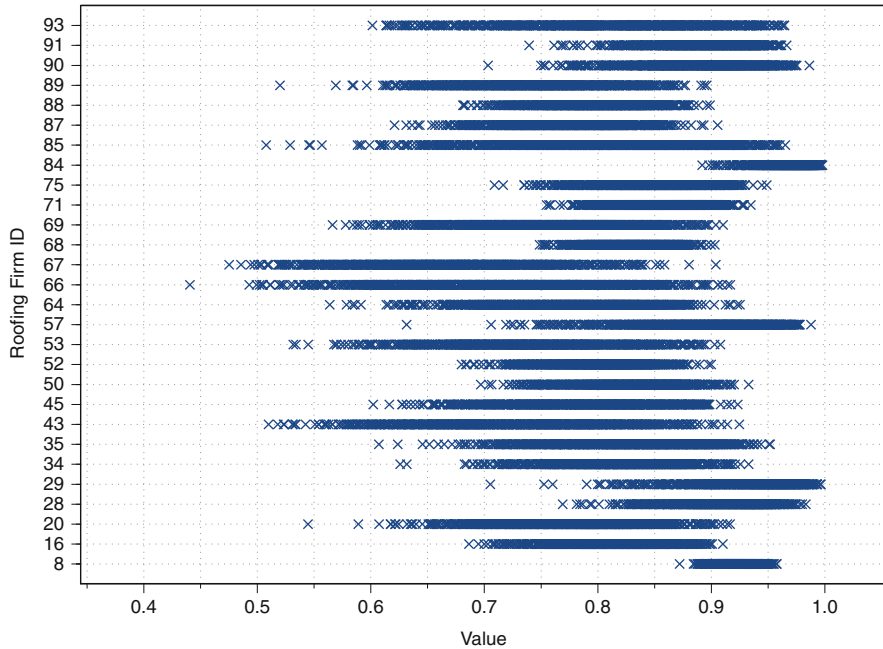


Fig. 2 Decision parameter distributions for the 28 non-dominated alternatives

Table 7 Results for each decision rule considered. For each firm, the table lists its rank when using the different decision rules. For the expected value approach, the value in parenthesis is the overall value. For the rank 1 decision rule, the value in parenthesis is the number of times that firm was ranked first. For Majority Judgment, the value in parenthesis is the majority-grade

Roofing firm ID	n	Expected value	Rank 1	Majority judgment
Firm 29	24	1 (0.9837)	2 (175)	2 (0.9426)
Firm 84	82	2 (0.9835)	1 (740)	1 (0.9724)
Firm 57	23	3 (0.9263)	3 (27)	5 (0.9007)
Firm 28	54	4 (0.9216)	4 (23)	4 (0.9164)
Firm 90	36	5 (0.9183)	5 (13)	6 (0.8980)
Firm 8	347	6 (0.9145)	6 (11)	3 (0.9234)

results for the Stochastic Dominance decision rule are best displayed graphically as empirical cumulative distribution curves, which are displayed in Fig. 3 for the top roofing firms.

As seen in Table 7, when considering the more comprehensive decision model that incorporates the attribute value uncertainty, the Rank 1 and Majority Judgment decision rules identify Roofing Firm 84 as the most desirable alternative with Roofing Firm 29 identified as the second most desirable alternative. (The expected value approach identified Roofing Firm 29 as the most desirable alternative, with Roofing Firm 84 as the second largest value.) In 1000 realizations, Roofing Firm

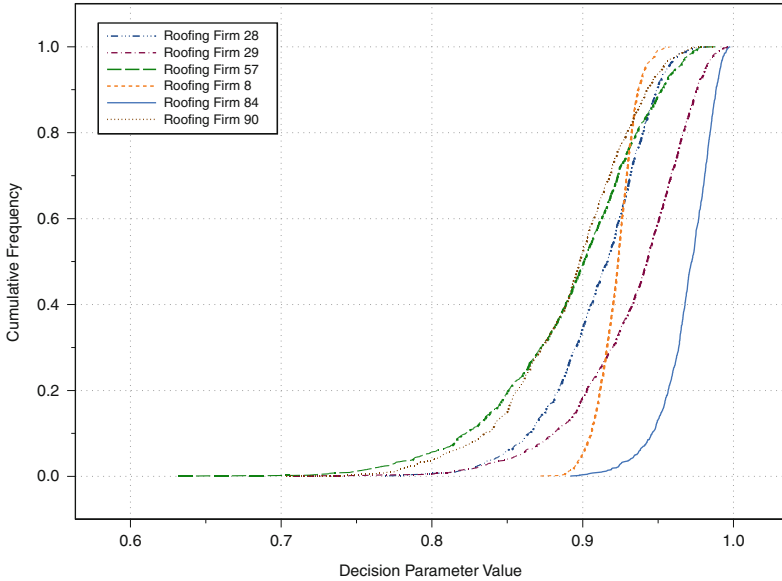


Fig. 3 Empirical cumulative distribution curves for the top roofing firms. Roofing Firm 84 is shown to stochastically dominate all other firms in the first degree

84 was the most desirable option 740 times, and Roofing Firm 29 was the most desirable option only 175 times. The majority-grade in the Majority Judgment selection method for Roofing Firm 84 was 0.9724, while the majority-grade for Roofing Firm 29 was 0.9426.

As shown in Fig. 3, the empirical cumulative distribution curve for Roofing Firm 84 never exceeds that of any other alternative, so Roofing Firm 84 stochastically dominates all other alternatives in the first degree. Thus Roofing Firm 84 is deemed to be the most desirable alternative using the Stochastic Dominance decision rule. This result is consistent with the results obtained by the other decision rules that consider the attribute value uncertainty. Further, the fact that we found one alternative that stochastically dominates all of the other alternatives in the first degree is an extremely powerful result, as Hadar and Russell [28] have shown that this is the decision-maker’s most preferred alternative regardless of his underlying utility function.

Summary and Conclusions

This chapter presented an approach for making decisions when uncertainty exists in the values of the attributes being used to compare the alternatives and derive a decision parameter. This type of uncertainty is different from uncertainty about

future events (risky decision) and uncertainty about the probabilities of future events (decision ambiguity). Ignoring this uncertainty, especially when it varies between alternatives, could lead to poor decisions.

The method presented here requires modeling the uncertainty about the attribute values and then propagating that uncertainty to determine the uncertainty in the decision parameter. The method is a Monte Carlo approach that randomly samples values of the uncertain attributes and computes the corresponding values of the decision parameter. Because it is not limited to specific types of distributions or decision models, it is a very general approach that can be used in a wide variety of settings. This chapter presented three decision rules (Rank 1, Majority Judgment, and Stochastic Dominance) for selecting an alternative based on its decision parameter distribution.

Unfortunately, it does require many samples of the uncertain attributes, which could be computationally expensive. The decision-maker must choose a decision rule to compare the distributions of the alternatives' decision parameters, and different rules may identify different alternatives as the "best".

This chapter has used the example of selecting a roofing firm to demonstrate the approach, but the approach can be used in any setting with attribute value uncertainty, including other models of attribute value uncertainty, other forms of the decision parameter, and other decision rules beyond those presented here.

The next research question to consider is that of experimental design: when planning to initially obtain, or if the decision-maker has the opportunity to get more information about some attributes for some alternatives, which information would be most valuable? In this case, information provides value by removing uncertainty about an alternative's decision parameter estimate.

References

1. Luce, R.D., Raiffa, H.: Games and Decision. Wiley, New York (1957)
2. Howard, R.A.: Decision analysis: applied decision theory. In: Proceedings of the Forth International Conference on Operational Research, pp. 55–71. Wiley, New York (1966)
3. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)
4. Savage, L.J.: The Foundations of Statistics, 1st edn. Wiley, New York (1954)
5. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives – Preferences and Value Tradeoffs, 2nd edn. Cambridge University Press, New York (1993)
6. Dyer, J.S.: MAUT – multiattribute utility theory. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) Multiple Criteria Decision Analysis – State of the Art Surveys, pp. 265–295. Springer, New York (2005)
7. Kirkwood, C.W.: Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets. Duxbury Press, Belmont (1997)
8. Dyer, J.S., Sarin, R.K.: Measurable multiattribute value functions. *Oper. Res.* **27**(4), 810–822 (1979)
9. von Winterfeldt, D., Edwards, W.: Decision Analysis and Behavior Research. Cambridge University Press, Cambridge (1986)

10. Farquhar, P.H., Keller, R.L.: Preference intensity measurement. *Ann. Oper. Res.* **19**, 205–217 (1989)
11. Einhorn, H.J., Hogarth, R.M.: Decision making under ambiguity. *J. Bus.* **59**(4), S225–S250 (1986)
12. Ellsberg, D.: Risk, ambiguity, and the savage axioms. *Q. J. Econ.* **75**(4), 643–669 (1961)
13. Frisch, D., Baron, J.: Ambiguity and rationality. *J. Behav. Decis. Making* **1**, 149–157 (1988)
14. Raiffa, H.: Risk, ambiguity, and the savage axioms: comment. *Q. J. Econ.* **75**(4), 690–694 (1961)
15. Srivastava, R.P.: Decision making under ambiguity: a belief-function perspective. *Arch. Contr. Sci.* **6**(1), 5–27 (1997)
16. Pandey, V., Nikolaidis, E., Mourelatos, Z.: Multi-objective decision making under uncertainty and incomplete knowledge of designer preferences. *SAE Int. J. Mater. Manuf.* **4**(1), 1155–1168 (2011)
17. Kahn, B.E., Meyer, R.J.: Consumer multiattribute judgments under attribute-weight uncertainty. *J. Consum. Res. Inc.* **17**(4), 508–522 (1991)
18. Mustajoki, J., Hamalainen, R.P., Salo, A.: Decision support by interval SMART/SWING-incorporating imprecision in the SMART and SWING methods. *Decis. Sci.* **36**(2), 317–339 (2005)
19. Chambal, S.P., Weir, J.D., Kahraman, Y.R., Gutman, A.J.: A practical procedure for customizable one-way sensitivity analysis in additive value models. *Decis. Anal.* **8**(4), 303–321 (2011)
20. Hyde, K., Maier, H.R., Colby, C.: Incorporating uncertainty in the PROMETHEE MCDA method. *J. Multi-Criteria Decis. Anal.* **12**, 245–259 (2003)
21. Zhang, Y., Fan, Z.-P., Liu, Y.: A method based on stochastic dominance degrees for stochastic multiple criteria decision making. *Comput. Ind. Eng.* **28**, 544–552 (2010)
22. Lindley, D.V.: *Understanding Uncertainty*. Wiley, Hoboken (2006)
23. Efron, B.: Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**(1), 1–26 (1979)
24. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman & Hall, New York (1993)
25. Chernick, M.R.: *Bootstrap Methods: A Practitioner’s Guide*. Wiley, New York (1999)
26. Winkler, R.L.: *Introduction to Bayesian Inference and Decision*. Holt, Rinehart and Winston, New York (1972)
27. Lee, P.M.: *Bayesian Statistics – An Introduction*, 2nd edn. Arnold, London (1997)
28. Hadar, J., Russell, W.R.: Rules for ordering uncertain prospects. *Am. Econ. Rev.* **59**, 25–34 (1969)
29. Balinski, M.L., Laraki, R.: A theory of measuring, electing, and ranking. *Proc. Natl. Acad. Sci. USA* **104**(21), 8720–8725 (2007)
30. Balinski, M.L., Laraki, R.: *Majority Judgment: Measuring, Ranking, and Electing*. The MIT Press, Cambridge (2010)
31. Center for the Study of Services: Who’s on top? Ratings of roofers. *Washington Consumers’ Checkbook*, **15**(4), 50–65 (2011, Spring/Summer)
32. Clemen, R.T., Reilly, T.: *Making Hard Decisions*, 2nd edn. Duxbury, Pacific Grove (2001)

Evaluating Incremental Values from New Predictors with Net Reclassification Improvement in Survival Analysis

Yingye Zheng, Layla Parast, Tianxi Cai, and Marshall Brown

Abstract Developing individualized prediction rules for disease risk and prognosis has played a key role in modern medicine. When new genomic or biological markers become available to assist in risk prediction, it is essential to assess the improvement in clinical usefulness of the new markers over existing routine variables. Net Reclassification Improvement (NRI) has been proposed to assess improvement in risk reclassification in the context of comparing two risk models and the concept has been quickly adopted in medical journals (Pencina et al., *Stat Med* 27:157–172, 2008). We propose both nonparametric and semiparametric procedures for calculating NRI as a function of a future prediction time t with a censored failure time outcome. The proposed methods accommodate covariate-dependent censoring, therefore providing more robust and sometimes more efficient procedures compared with the existing nonparametric-based estimators (Pencina et al., *Stat Med* 30: 11–21, 2011; Uno et al., *Stat Med* 32:2430–42, 2013). Simulation results indicate that the proposed procedures perform well in finite samples. We illustrate these procedures by evaluating a new risk model for predicting the onset of cardiovascular disease.

Y. Zheng (✉) • M. Brown
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North,
Seattle, WA 98109, USA
e-mail: yzheng@fhcrc.org; mdbrown@fhcrc.org

L. Parast
RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, USA
e-mail: parast@rand.org

T. Cai
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
e-mail: tcai@hsph.harvard.edu

Introduction

Developing individualized prediction rules for disease risk and prognosis is fundamental for successful disease prevention and treatment selection. For many diseases, risk prediction models have been developed and incorporated into clinical practice guidelines. For example, the Gail model was developed for predicting individual breast cancer risk [10] and a risk calculator based on that model can be used to assist physicians making screening recommendations. For cardiovascular disease (CVD), prediction models such as the Framingham Risk Score (FRS) are used for stratifying patients into different levels of risks. However, much refinement is needed even for the best of these models because of their limited discriminatory accuracy. For example, the Framingham model, largely based on traditional clinical risk factors, has recognized limitations in its clinical utility [12]. A considerable fraction of patients who experienced CVD events had none of the identified risk factors, indicating a need to explore avenues beyond routine clinical measures for more accurate prediction [15]. This fuels much of the current search for novel biologic markers and genetic factors that, when combined with routine clinical risk factors, may provide accurate prediction at the individual level.

When new genomic or biological markers become available to assist in risk prediction, it is essential to assess the clinical usefulness of these new markers compared to existing routine markers. Careful evaluation of the incremental value is particularly crucial when markers are either expensive or invasive to measure. To quantify the added clinical value of new markers over a conventional risk scoring system for predicting disease risk, one may calculate the difference in the prediction measures for the existing conventional model and the new model, which includes information from the new markers. For example the difference in the areas under the receiver operating characteristic curves (AUC of ROC) are often used to quantify the improvement in discrimination attributable to added markers. Since a risk model is often used to stratify patients into proper risk categories, statistical summaries that depend on clinically meaningful risk thresholds may be more relevant [4, 6, 17]. As an alternative to measuring the difference between AUCs, Net Reclassification Improvement (NRI) has also been proposed to assess improvement in risk reclassification in the context of comparing two risk models constructed with and without novel markers [18]. Using “up” and “down” to denote changes in one or more risk categories in the upward and downward directions, respectively, for a subject between their baseline and augmented risk values, the NRI is defined as

$$\text{NRI} = [\text{Pr}(\text{up}|\text{Diseased}) + \text{Pr}(\text{down}|\text{Healthy})] - [\text{Pr}(\text{down}|\text{Diseased}) + \text{Pr}(\text{up}|\text{Healthy})].$$

Such a measure is appealing because it acknowledges both desirable risk reclassifications (up for diseased and down for healthy subjects) and undesirable risk reclassifications (down for diseased and up for healthy subjects). Due to its simplicity, NRI has been quickly adopted in medical journals. However, compared with many other measures for incremental values, the concept has not received much attention in the statistical literature.

Since a risk model is often used for predicting an individual's future outcome, it is essential to incorporate the additional dimension of time when assessing the performance of a risk model in a cohort study. For both deriving and evaluating risk models, prospective cohort data is often used. In this setting a subject's health status at a future time t is sometimes unknown due to loss of follow-up, termination of a study or the occurrence of a competing risk event. Such censoring poses additional challenges compared with settings previously examined in the literature which focus on incremental value calculation with a dichotomous outcome. Currently there is limited development in methods to estimate the incremental value of novel markers with censored failure time outcomes. Recently Pencina et al. [19] proposed a method for calculating time-dependent NRI, based on nonparametric Kaplan-Meier (KM) estimators in order to account for censoring in cohort data. The asymptotic properties of a similar estimator is studied in detail in [24]. However, the validity of these estimators relies critically on the assumption that censoring is independent of predictors used in the risk models. Furthermore, the nonparametric procedure considered in these estimators may potentially lead to efficiency loss. A more flexible and more efficient estimating procedure is needed in practice.

In this manuscript, we propose quantitative procedures for calculating NRI as a function of a future prediction time t with a censored failure time outcome. Compared with existing nonparametric estimators, our procedures do not require the assumption that censoring is independent of predictors, therefore the methods would be widely applicable to many practical situations. We also consider procedures that aim to improve efficiency while maintaining robustness. This manuscript is organized as follows. In section "Measures of Risk Stratification and Reclassification", we specify models and define NRI suitable for event time outcomes. In section "Estimation", we describe procedures for estimating time-dependent NRI using data obtained from a prospective cohort study with a failure time outcome. We comment on the theoretical properties of our proposed estimators in section "Inference". We then describe simulation studies to evaluate the performance of the proposed estimators. The results are reported in section "Simulation Studies". An application of our procedures for comparing two CVD risk models is presented in section "Example". Concluding remarks are in section "Discussion".

Measures of Risk Stratification and Reclassification

Consider the situation that a vector of predictor \mathbf{Y} measured at baseline is used for predicting the time to event outcome T . Risk models can be built using sub-vectors of \mathbf{Y} . Let $\mathbf{Y}_{(1)}$, a function of \mathbf{Y} , denote a vector of conventional predictor values in the existing model. Let $\mathbf{Y}_{(2)}$, also a function of \mathbf{Y} , denote a vector of predictors used in the new model that contains $\mathbf{Y}_{(1)}$, but also new predictor values. Individual-level risk at a future time t can be derived as $P = \Pr(T \leq t | \mathbf{Y}_{(1)})$, based on the conventional model, and $Q = \Pr(T \leq t | \mathbf{Y}_{(2)})$, the corresponding risk based on the new model, respectively. Since, in practice, risk categories are often uncertain

for many diseases, a more objective and flexible measure of improvement in risk prediction would be based on P or Q in their original continuous scales. Therefore, following the definition of [19], in this manuscript we focus on the time-dependent continuous NRI, which is a more general definition that does not rely on the existence of risk categories. In the time-dependent setting, we further denote an ‘event’ person at time t as those with $T \leq t$, and a ‘nonevent’ person as $T > t$. Here, $\text{NRI}(t)$ is equal to the sum of ‘event NRI’ and ‘nonevent NRI’, which are defined as:

$$\text{event NRI}_u(t) = \Pr(Q - P > u | T \leq t) - \Pr(Q - P \leq u | T \leq t) \equiv 2\Pr(Q - P > u | T \leq t) - 1,$$

and

$$\text{nonevent NRI}_v(t) = \Pr(Q - P \leq v | T > t) - \Pr(Q - P > v | T > t) \equiv 1 - 2\Pr(Q - P > v | T > t).$$

Since, $\text{NRI}_{u,v}(t) = \text{event NRI}_u(t) + \text{nonevent NRI}_v(t)$, it follows that $\text{NRI}_{u,v}(t) = 2\{\Pr(Q - P > u | T \leq t) - \Pr(Q - P > v | T > t)\}$. In practice we may chose u and v such that improvement in risk estimates is meaningful [24]. Setting $u = v = 0$ gives the ‘continuous NRI’ considered in [19]. For the ease of presentation, in the sequel, we’ll omit the subscript u and v from our notations and assume $u = v = 0$, but note that our estimators can be constructed for any arbitrary u and v . In the next section, we show how each component of $\text{NRI}(t)$ can be estimated.

Estimation

Suppose we have a cohort of N individuals from the targeted population followed prospectively. Due to censoring, the observed data consist of N i.i.d copies of vector, $\mathcal{D} = \{\mathbf{D}_i = (X_i, \delta_i, \mathbf{Y}_i)^\top, i = 1, \dots, N\}$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ for T_i and C_i denote failure time and censoring time respectively. \mathbf{Y}_i are predictors from individual i measured at time 0, including subset $\mathbf{Y}_{i(1)}$ used in the existing model (model 1) and $\mathbf{Y}_{i(2)}$ in the new model (model 2) such that $\mathbf{Y}_{i(1)} \in \mathbf{Y}_{i(2)}$. For illustration, we first assume that P and Q both follow the conventional Cox regression models. Specifically, at time t , we assume $P(\theta_1) = 1 - \exp[\Lambda_{01}(t) \exp\{\beta_1^\top \mathbf{Y}_{(1)}\}]$ and $Q(\theta_2) = 1 - \exp[\Lambda_{02}(t) \exp\{\beta_2^\top \mathbf{Y}_{(2)}\}]$, where Λ_{0k} is the baseline cumulative hazard function, β_k are unknown vector of parameters, for model $k = 1, 2$, and $\theta_1 = (\Lambda_{01}(t), \beta_1^\top)^\top$, $\theta_2 = (\Lambda_{02}(t), \beta_2^\top)^\top$. It is important to note that these models are most likely not correctly specified. Nevertheless under a mild regularity condition, the standard maximum partial likelihood estimator $\hat{\beta}_k$ for β_k converges to a constant vector, as $n \rightarrow \infty$ [13]. This provides theoretical ground for our asymptotic studies.

To estimate $\text{NRI}(t)$, Pencina et al. [19] first expressed the two key components as

$$\Pr\{B(\theta) > 0 | T \leq t\} = \frac{\Pr\{T \leq t | B(\theta) > 0\} \Pr\{B(\theta) > 0\}}{\Pr(T \leq t)}$$

and

$$\Pr\{B(\theta) > 0|T > t\} = \frac{\Pr\{T > t|B(\theta) > 0\}\Pr\{B(\theta) > 0\}}{\Pr\{T > t\}},$$

where $B(\theta) = Q(\theta_2) - P(\theta_1)$ and $\theta = (\theta_1, \theta_2)^\top$. To account for censoring, Pencina et al. [19] proposed to use the KM estimator to estimate the survival function using data from all subjects for $\Pr(T \leq t)$ and using subjects with $B(\theta) > 0$ for estimation of $\Pr[T \leq t|\{B(\theta) > 0\}]$. We refer to the resulting estimator as the ‘KM estimator’ hereafter.

Uno et al. [24] considered estimating $NRI(t)$ based on an inverse-probability-of-censoring weighted (IPW) estimator (hereafter referred to as the ‘IPW estimator’), with its key components estimated as

$$\widehat{\Pr}^{IPW}\{B(\theta) > 0|T \leq t\} = \frac{\sum_i I\{B_i(\hat{\theta}) > 0, X_i \leq t\} \hat{W}_i(t)}{\sum_i I(X_i \leq t) \hat{W}_i(t)} \tag{1}$$

$$s\widehat{\Pr}^{IPW}\{B(\theta) > 0|T > t\} = \frac{\sum_i I\{B_i(\hat{\theta}) > 0, X_i > t\} \hat{W}_i(t)}{\sum_i I(X_i > t) \hat{W}_i(t)} \tag{2}$$

where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^\top$, $\hat{\theta}_1 = (\hat{\Lambda}_{01}(t), \hat{\beta}_1^\top)^\top$, $\hat{\theta}_2 = (\hat{\Lambda}_{02}(t), \hat{\beta}_2^\top)^\top$, $\hat{W}_i(t) = I(X_i \leq t)\delta_i/\hat{H}(X_i) + I(X_i > t)/\hat{H}(t)$ and $\hat{H}(\cdot)$ is the KM estimator of $H(\cdot) = P(C > \cdot)$. Due to the equivalence between the KM estimator and the IPW estimator for marginal survival functions under independent censoring [21], the two estimators are likely to have very similar robustness and efficiency. Both estimators are consistent under an independent censoring assumption regardless of the adequacy of the two fitted models, $P(\theta_1)$ and $Q(\theta_2)$. This is particularly appealing for model comparisons.

One potential weakness of both estimators is that they can be biased if censoring is dependent on a subset of $\mathbf{Y}_{(2)}$. On the other hand, when model 2 is correctly specified, such covariate-dependent censoring can be incorporated based on the model since $C \perp T$ given $\beta_2^\top \mathbf{Y}_{(2)}$ or $Q(\theta_2)$. This motivates us to propose a more robust alternative to the [24] estimator by estimating censoring probabilities given $\mathbf{Y}_{(2)}$ via kernel smoothing over $Q(\theta_2)$. Let $H_q^1(t) = P(C > t | Q(\theta_2) = q, \Delta_i(\theta) = 1)$ and $H_q^\bullet(t) = P(C > t | Q(\theta_2) = q)$ where $\Delta_i(\theta) = I\{B_i(\theta) > 0\}$. To estimate $NRI(t)$, we propose to modify Eqs. 1 and 2 by considering the following more robust IPW censoring weights

$$\tilde{W}_i^{(t)}(t) = \frac{I(X_i \leq t)\delta_i}{\tilde{H}_{Q_i(\hat{\theta}_2)}^{(t)}(X_i)} + \frac{I(X_i > t)}{\tilde{H}_{Q_i(\hat{\theta}_2)}^{(t)}(t)} \quad \text{for } t = 1 \text{ and } \bullet,$$

where $\tilde{H}_q^{(t)}(t) = \exp\{-\hat{\Lambda}_q^{(t)}(t)\} = \exp\{-\int_0^t \hat{\pi}_q^{(t)}(s)^{-1} d\hat{N}_{Cq}^{(t)}(s)\}$,

$$\hat{N}_{Cq}^{(t)}(s) = n^{-1} \sum_{i:\Delta_i(\hat{\theta}) \in \mathcal{U}_1} K_h\{Q_i(\hat{\theta}_2) - q\} N_{Ci}(s), \quad \hat{\pi}_q^{(t)}(s) = n^{-1} \sum_{i:\Delta_i(\hat{\theta}) \in \mathcal{U}_1} K_h\{Q_i(\hat{\theta}_2) - q\} I(X_i \geq s),$$

$N_{Ci}(s) = I(X_i \leq s)(1 - \delta_i)$, $\mathcal{U}_1 = 1$ and $\mathcal{U}_\bullet = \{0, 1\}$, $K_h(\cdot) = \frac{1}{h}K\left\{\frac{q-Q_i(\theta_2)}{h}\right\}$, K is a symmetric kernel density function, with $h = h(n) \rightarrow 0$ as the bandwidth. Note that $\Delta_i(\hat{\theta}) \in \mathcal{U}_1$ is simply the subset of individuals with $B_i(\hat{\theta}) > 0$ and $\Delta_i(\hat{\theta}) \in \mathcal{U}_\bullet$ is the set of *all* individuals. Consequently we can then use these more robust kernel smoothing weights in the IPW estimator, to obtain the ‘Smooth-IPW (S-IPW) estimators’,

$$\widehat{\Pr}^{\text{S-IPW}}\{B(\theta) > 0|T \leq t\} = \frac{\sum_i \Delta_i(\hat{\theta}) \tilde{W}_i^{(1)}(t) I(X_i \leq t)}{\sum_i \tilde{W}_i^{(\bullet)}(t) I(X_i \leq t)} \text{ and} \tag{3}$$

$$\widehat{\Pr}^{\text{S-IPW}}\{B(\theta) > 0|T > t\} = \frac{\sum_i \Delta_i(\hat{\theta}) \tilde{W}_i^{(1)}(t) I(X_i > t)}{\sum_i \tilde{W}_i^{(\bullet)}(t) I(X_i > t)}. \tag{4}$$

This resulting estimator for $NRI(t)$ is

$$\widehat{NRI}(\hat{\theta}, t) = 2 \times \left[\widehat{\Pr}^{\text{S-IPW}}\{B(\hat{\theta}) > 0|T \leq t\} - \widehat{\Pr}^{\text{S-IPW}}\{B(\hat{\theta}) > 0|T > t\} \right].$$

The estimator can be shown to have the property of ‘double robustness’, i.e., it only requires that the risk model Q is correctly specified or that the independent censoring assumption holds.

Additionally, to improve upon the efficiency of the class of nonparametric estimators, we propose considering a semiparametric estimator. Note that

$$\Pr\{B(\theta) > 0|T > t\} = \frac{E[E\{I(B(\theta) > 0, T > t) | \mathbf{Y}_{(2)}\}]}{E[E\{I(T > t) | \mathbf{Y}_{(2)}\}]} = \frac{E\{I(B(\theta) > 0)P(T > t | \mathbf{Y}_{(2)})\}}{E\{P(T > t | \mathbf{Y}_{(2)})\}}.$$

Therefore $NRI(t)$ can be estimated semiparametrically as

$$\widehat{NRI}(\hat{\theta}, t) = 2 \times \left\{ \widehat{\Pr}^{\text{SEM}}(B(\theta) > 0|T \leq t) - \widehat{\Pr}^{\text{SEM}}(B(\theta) \leq 0|T > t) \right\},$$

with the ‘SEM’ estimators,

$$\widehat{\Pr}^{\text{SEM}}(B(\theta) > 0|T \leq t) = \frac{\sum_{i=1}^n \Delta_i(\hat{\theta}) Q_i(\hat{\theta}_2)}{\sum_{i=1}^n Q_i(\hat{\theta}_2)}, \tag{5}$$

$$\widehat{\Pr}^{\text{SEM}}(B(\theta) > 0|T > t) = \frac{\sum_{i=1}^n \Delta_i(\hat{\theta}) \{1 - Q_i(\hat{\theta}_2)\}}{\sum_{i=1}^n \{1 - Q_i(\hat{\theta}_2)\}}. \tag{6}$$

Under the correctly specified model $Q(\theta_2)$, the semiparametric estimator accommodates a covariate-dependent censoring situation and would be more efficient compared to the Smooth-IPW estimator. In practice, to estimate $NRI(t)$, if estimates from such a semiparametric method agree well with those of the nonparametric

methods, one may choose to report results based on the semiparametric method for additional gain in efficiency. To automatize the procedure, we suggest considering a combined estimator (hereafter referred as the ‘combined estimator’), which takes the form

$$\hat{p} \times \widehat{\text{NRI}}(\hat{\theta}, t) + (1 - \hat{p}) \times \widetilde{\text{NRI}}(\hat{\theta}, t),$$

with \hat{p} being a weight that is dependent on the aptness of the semiparametric model. For example, \hat{p} can be taken to be the p-value from a consistent test of the proportional hazards assumption for a Cox regression model fit. Such an estimator provides a simple procedure which is robust over a wide variety of situations. In numerical studies, we show that such a combined estimator can be more efficient compared with the nonparametric estimators, while maintaining the double robustness property.

We note that the proposed estimators can be easily generalized to NRI based on risk categories. Consider a situation where individuals are classified as low, intermediate or high risk: low risk if their risks are below r_1 , and high risk if their risks are above r_2 . The reclassification accuracy of risk models in such a setting can be quantified with a 3-category NRI of the form $\text{NRI}^{\text{category}}(\hat{\theta}, t) = P(\text{up}|T \leq t) - P(\text{down}|T \leq t) + P(\text{down}|T > t) - P(\text{up}|T > t)$. To estimate $P(\text{up}|T \leq t)$ and $P(\text{up}|T > t)$, we may simply replace $\Delta_i(\hat{\theta})$ with $\Omega_i^{\text{up}}(\hat{\theta}) = I(P_i(\theta_1) \leq r_1, Q_i(\theta_2) > r_1) + I(r_1 < P_i(\theta_1) \leq r_2, Q_i(\theta_2) > r_2)$ in Eqs. 3 and 4, respectively. Similarly, to estimate $P(\text{down}|T \leq t)$ and $P(\text{down}|T > t)$, one may replace $\Delta_i(\hat{\theta})$ with $\Omega_i^{\text{down}}(\hat{\theta}) = I(Q_i(\theta_1) \leq r_1, P_i(\theta_2) > r_1) + I(r_1 < Q_i(\theta_1) \leq r_2, P_i(\theta_2) > r_2)$ in Eqs. 3 and 4. Similarly, one may obtain a semiparametric estimator of $\text{NRI}^{\text{category}}(\hat{\theta}, t)$ by replacing $\Delta_i(\hat{\theta})$ with $\Omega_i^{\text{up}}(\hat{\theta})$, or $\Omega_i^{\text{down}}(\hat{\theta})$ in Eqs. 5 and 6.

Inference

To make inference about $\widetilde{\text{NRI}}(\hat{\theta}, t)$, we study the asymptotic properties of proposed estimators. In the Appendix, we show that $\widetilde{\text{NRI}}(\hat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$, where $\theta_0 = (\Lambda_{k0}(\cdot), \beta_{k0}^\top)^\top$ with β_{k0} being the unique maximizer of the expected value of the corresponding partial likelihood. Furthermore, we show that the process $\tilde{\mathcal{W}}(t) = \sqrt{n}\{\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)\}$ is asymptotically equivalent to a sum of i.i.d terms, $n^{-1/2} \sum_{i=1}^n \varepsilon_i(t)$ where $\varepsilon_i(t)$ is defined in the Appendix. By a functional central limit theorem of [20], the process $\tilde{\mathcal{W}}(t)$ converges weakly to a mean zero Gaussian process in t . We also show that $\widehat{\text{NRI}}(\hat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$, and that the process $\tilde{\mathcal{N}}(t) = \sqrt{n}[\widehat{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)]$ is asymptotically equivalent to a sum of i.i.d terms $n^{-1/2} \sum_{i=1}^n \zeta_i(t)$ where $\zeta_i(t)$ is defined in the Appendix. Again, by a functional central limit theorem, the process

$\mathcal{N}(t)$ converges weakly to a mean zero Gaussian process in t . With weak convergence of both $\widehat{\text{NRI}}(\hat{\theta}, t)$ and $\widetilde{\text{NRI}}(\hat{\theta}, t)$, it follows that the combined estimator converges to a zero-mean process. Due to the variation in $\hat{\rho}$, the combined estimators may not be a Gaussian process. We show in our simulation that to make inference, resampling procedures such as a bootstrap method can provide a valid approximation of the limit distribution. Specifically, at each of the b th bootstrap iterations, with $b = 1, \dots, B$, we conduct a random sampling with replacement of the original dataset, and fit our new and old risk models based on the sampled dataset, denoted as $P^b(\hat{\theta})$ and $Q^b(\hat{\theta})$. These estimates from the fitted models are then used to calculate $\widehat{\text{NRI}}^b(\hat{\theta}, t)$ and $\widetilde{\text{NRI}}(\hat{\theta}, t)$ based on the bootstrapped samples. This procedure will be repeated B times, and confidence intervals can be constructed either based on the percentile method, or a normal approximation where the standard error is calculated based on the empirical standard errors of $\{\widehat{\text{NRI}}^b(\hat{\theta}, t), b = 1, \dots, B\}$ and $\{\widetilde{\text{NRI}}(\hat{\theta}, t), b = 1, \dots, B\}$. The combined estimator can be inferred similarly by repeatedly calculate the weights based on each bootstrap sample in addition to $\widehat{\text{NRI}}^b(\hat{\theta}, t)$ and $\widetilde{\text{NRI}}(\hat{\theta}, t)$.

In the absence of an independent validating set, often in practice the same dataset is used for both fitting the model with several predictors and calculating a measure such as $\text{NRI}(t)$. Such an ‘apparent’ summary may potentially lead to the so-called ‘overfitting’ phenomenon, i.e. estimates of model performance will tend to be more optimistic compared with the corresponding estimates if the model were to applied to a new dataset. Several methods for correcting the bias from apparent estimates can be considered. The 0.632 Bootstrap method [9] has been shown to have better performance compared with a simple cross-validated approach. The estimator was derived in our simulation as follows: we first obtained a bootstrapped estimate $\widehat{\text{NRI}}^{bt}(t)$ by sampling the data with replacement to obtain the training set. The training set is used to estimate the model parameters $\{\hat{\theta}_k^{(\text{train})}, k = 1, 2\}$. The remaining subjects make up the validation set, and are used to calculate the various estimates of NRI using parameter values $\{\hat{\theta}_k^{(\text{train})}, k = 1, 2\}$. This is repeated B times and $\widehat{\text{NRI}}^{bt}(t)$ is the mean across the repetitions. The 0.632 bootstrap estimate is,

$$\widehat{\text{NRI}}^{0.632bt}(t) = 0.632\widehat{\text{NRI}}^{bt}(t) + (1 - 0.632)\widehat{\text{NRI}}^{\text{apparent}}(t),$$

where $\widehat{\text{NRI}}^{\text{apparent}}(t)$ is the estimate without using cross-validation. To construct a confidence interval based on $\widehat{\text{NRI}}^{0.632bt}(t)$, we follow the suggestions given in [22] and [23] by shifting the apparent error based confidence interval in the amount of bias estimated as $\widehat{\text{bias}} = \widehat{\text{NRI}}^{\text{apparent}}(t) - \widehat{\text{NRI}}^{0.632bt}(t)$. Specifically, if $[L, R]$ is the confidence interval calculated based on the procedure described above, then the bias corrected confidence interval is $[L - \widehat{\text{bias}}, R - \widehat{\text{bias}}]$.

Simulation Studies

To examine the performance of various $NRI(t)$ estimators, we conducted simulation studies under several different scenarios. Throughout we chose $n = 500$ and used 200 bootstrap samples to calculate standard errors. Results for each setting were produced from 1,000 simulations. We calculated $NRI(t)$, for $t = 3$ for comparing two risk models using the KM, IPW, Smooth-IPW, SEM and the combined estimators described in section “Estimation”. We fitted Cox regression models to calculate risks for both the new and existing models using corresponding predictors.

For the first setting presented in Table 1, two predictors Y_1 and Y_2 were simulated from a multivariate normal distribution with mean $(0, 0.5)$, $\sigma_{y_1} = \sigma_{y_2} = 1$ and a

Table 1 Simulation results under noninformative censoring and correctly specified new risk model with mean of bias (Mean(bias)) and standard deviation (Std. dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (Mean(std. error)), and coverage of the 95% bootstrap confidence interval based on the normal approximation. Note that KM = Kaplan-Meier estimator, IPW = Inverse Probability Weighted estimator, Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Method	$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i > t)$	$NRI(t)$
True values	0.592	0.358	0.468
KM			
Mean(bias)	0.003	0.001	0.002
Std. dev.	0.034	0.030	0.104
Mean(std. error)	0.034	0.030	0.103
95% bootstrap CI cov.	0.946	0.946	0.946
IPW			
Mean(bias)	0.002	0.002	-0.001
Std. dev.	0.034	0.030	0.105
Mean(std. error)	0.034	0.031	0.104
95% bootstrap CI cov.	0.943	0.95	0.951
Smooth IPW			
Mean(bias)	0.001	0.003	-0.003
Std. dev.	0.034	0.030	0.104
Mean(std. error)	0.034	0.030	0.103
95% bootstrap CI cov.	0.946	0.942	0.949
SEM			
Mean(bias)	0.001	0.003	-0.003
Std. dev.	0.024	0.029	0.082
Mean(std. error)	0.025	0.028	0.080
95% bootstrap CI cov.	0.952	0.942	0.937
Combined			
Mean(bias)	0.002	0.003	-0.002
Std. dev.	0.029	0.028	0.089
Mean(std. error)	0.031	0.029	0.095
95% bootstrap CI cov.	0.968	0.949	0.969

correlation ρ of 0.25. The relationship between survival time T and \mathbf{Y} followed a proportional hazards model with parameters $\beta_1 = \log(3)$ and β_2 equal to $\log(1.5)$. Censoring time was generated from a $U(0, a)$ distribution where a was chosen to produce approximately 40% censoring. Note that in this setting, model Q is correctly specified and the independent censoring assumption is correct. We took the baseline model to consist of Y_1 and the new model to include both predictors. As expected, all estimators shown in Table 1 provide unbiased estimates. The bootstrap-based variance estimators perform well with coverage percentage close to the 95% nominal level. Since the risk based on the new model is correctly specified, the semiparametric method is the most efficient. Improvement in efficiency over the nonparametric procedures is observed with our combined estimators.

Under this setting we also considered a null model where $\beta_2 = 0$ i.e. there is no incremental value of the new marker and $NRI(t) = 0$. We found that in this situation all estimators tend to slightly over estimate $NRI(t)$, and variance estimators based on the bootstrap estimators tend to be conservative (see Table 2). We do not recommend calculating $NRI(t)$ in the case when the new marker does not independently predict outcome in a model with conventional predictors. Note that all theoretical results in the Appendix are derived under the assumption that $\beta_2 \neq 0$ and thus our proposed procedures are only valid under this assumption. In practice, if the null setting is a likely possibility, estimation should be treated with care.

The second setting we considered was identical to the first setting, except that censoring time was dependent on marker values. Here, censoring time,

$$C = U \cdot B + \exp(X - 3Y_2) \cdot (1 - B),$$

where U was generated from a Uniform(0, a) distribution where with a chosen to yield about 40% censoring, X was generated from a $N(0, 1)$ distribution and B was generated from a $N(2 \cdot Y_1, 1)$ distribution. Note that in this setting, model Q is correctly specified but the independent censoring assumption is not correct. As seen in the results presented in Table 3, the KM estimator yields biased estimators for both $NRI(t)$ and its two key components. The IPW estimator is biased for both $\Pr(P > Q | T \leq t)$ and $NRI(t)$, whereas the smooth-IPW estimator substantially alleviates such biases. However, we observed large variation in nonparametric estimators of $NRI(t)$ as compared with the semiparametric and combined estimators (Table 3).

The third setting we investigated considers the case where survival time depends on four markers Y_i , for $i = 1, \dots, 4$, but we only have access to the first two. In particular, \mathbf{Y} comes from a multivariate normal distribution with mean 0, and $\sigma_{ij} = 1$ for $i = j$ and 0.25 otherwise. Survival time relates to the marker data through a model where the hazard function is specified as $\lambda(t|\mathbf{Y}) = 0.1 * \{3Y_1 + 1.5Y_2 + 2Y_3 + 2.5Y_4 + \exp(3Y_1)\}$. Note that in this setting, model Q is misspecified as depending only on Y_1 and Y_2 . Censoring time in this setting is generated the same as in the first setting, which does not depend on T or \mathbf{Y} . Since the SEM estimator misspecified the relationship between T and \mathbf{Y} as $\lambda(t|\mathbf{Y}) = \lambda_0 \exp(\beta_1 Y_1 + \beta_2 Y_2)$,

Table 2 Simulation results under noninformative censoring and correctly specified new risk model with mean of bias (Mean(bias)) and standard deviation (Std. dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (Mean(std. error)), and coverage of the 95% bootstrap confidence interval based on the normal approximation. Data is generated under the null model that $\beta_2 = 0$. Note that KM = Kaplan-Meier estimator, IPW = Inverse Probability Weighted estimator, Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Method	$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i > t)$	$\text{NRI}(t)$
Null model: $\beta_2 = 0$			
True values	0.5	0.5	0
KM			
Mean(bias)	0.01	-0.02	0.061
Std. dev.	0.034	0.026	0.091
Mean(std. error)	0.043	0.033	0.118
95% bootstrap CI cov.	0.996	0.971	0.98
IPW			
Mean(bias)	0.01	-0.019	0.058
Std. dev.	0.034	0.026	0.092
Mean(std. error)	0.044	0.033	0.119
95% bootstrap CI cov.	0.996	0.972	0.981
Smooth IPW			
Mean(bias)	0.009	-0.019	0.055
Std. dev.	0.034	0.026	0.092
Mean(std. error)	0.044	0.033	0.118
95% bootstrap CI cov.	0.996	0.972	0.981
SEM			
Mean(bias)	0.009	-0.019	0.057
Std. dev.	0.023	0.025	0.067
Mean(std. error)	0.029	0.031	0.081
95% bootstrap CI cov.	0.99	0.967	0.957
Combined			
Mean(bias)	0.008	-0.019	0.055
Std. dev.	0.029	0.025	0.077
Mean(std. error)	0.039	0.032	0.104
95% bootstrap CI cov.	0.997	0.971	0.977

it yields biased results. All other estimators are unbiased (Table 4). Throughout the three settings we considered, the combined estimator remained unbiased and more efficient than other nonparametric estimators.

To evaluate the procedures described above, we simulated 10 markers from a multivariate normal distribution with mean $\mathbf{0}$, $\sigma_{Y_i} = 1$ and pairwise correlations equal to 0.25. The number of parameters and sample size were chosen to mimic the setting of our data example described in section “Example”. We consider a Cox model for failure time, with hazard ratio parameters for 10 markers specified as $\beta = (\log(2), \log(.77), 0, \log(1.81), 0, 0, 0, \log(0.5), 0, \log(1.2))$. The baseline model

Table 3 Simulation results under covariate-dependent censoring and correctly specified new risk model with mean of bias (Mean(bias)) and standard deviation (Std. dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (Mean(std. error)), and coverage of the 95% bootstrap confidence interval based on the normal approximation. Note that KM = Kaplan-Meier estimator, IPW = Inverse Probability Weighted estimator, Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Method	$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i > t)$	$\text{NRI}(t)$
True values	0.611	0.45	0.322
KM			
Mean(bias)	0.067	-0.062	0.259
Std. dev.	0.040	0.040	0.126
Mean(std. error)	0.041	0.040	0.129
95% bootstrap CI cov.	0.615	0.659	0.483
IPW			
Mean(bias)	-0.024	0.005	-0.057
Std. dev.	0.034	0.045	0.131
Mean(std. error)	0.035	0.044	0.130
95% bootstrap CI cov.	0.897	0.944	0.918
Smooth IPW			
Mean(bias)	-0.013	0.007	-0.038
Std. dev.	0.041	0.041	0.133
Mean(std. error)	0.040	0.040	0.132
95% bootstrap CI cov.	0.937	0.939	0.941
SEM			
Mean(bias)	0	-0.001	0.002
Std. dev.	0.025	0.039	0.098
Mean(std. error)	0.026	0.037	0.095
95% bootstrap CI cov.	0.951	0.932	0.938
Combined			
Mean(bias)	-0.006	0.002	-0.016
Std. dev.	0.031	0.039	0.109
Mean(std. error)	0.035	0.039	0.117
95% bootstrap CI cov.	0.975	0.951	0.971

consists only of the first marker. To derive a new model based on the information on all 10 markers, for each simulation, we first fit a model with all ten markers. The expanded model consists of all markers that have non-zero β at an $\alpha = 0.05$ level. We found that in the case of estimating NRI, under our simulated scenario, the apparent summaries are quite close to the true values in many cases. Since the bias is at the rate of g/N , where g is the number of predictors under consideration for risk model building, overfitting may be of more concern when large numbers of genetic markers are involved with a relatively small sample size. In the situation there is a slight indication of overfitting, the 0.632 bootstrap procedure appears to be adequate in correcting the bias (see Table 5).

Table 4 Simulation results under noninformative censoring and misspecified new risk model with mean of bias (Mean(bias)) and standard deviation (Std. dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (Mean(std. error)), and coverage of the 95% bootstrap confidence interval based on the normal approximation. Note that KM = Kaplan-Meier estimator, IPW = Inverse Probability Weighted estimator, Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Method	$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i > t)$	$NRI(t)$
True values	0.646	0.395	0.504
KM			
Mean(bias)	0.007	-0.002	0.016
Std. dev.	0.072	0.023	0.160
Mean(std. error)	0.074	0.024	0.164
95% bootstrap CI cov.	0.94	0.945	0.947
IPW			
Mean(bias)	0.004	-0.001	0.008
Std. dev.	0.072	0.023	0.160
Mean(std. error)	0.074	0.024	0.165
95% bootstrap CI cov.	0.945	0.942	0.95
Smooth IPW			
Mean(bias)	0.003	-0.001	0.007
Std. dev.	0.072	0.023	0.160
Mean(std. error)	0.074	0.024	0.164
95% bootstrap CI cov.	0.943	0.946	0.95
SEM			
Mean(bias)	-0.046	0.003	-0.099
Std. dev.	0.022	0.022	0.068
Mean(std. error)	0.022	0.023	0.068
95% bootstrap CI cov.	0.448	0.943	0.682
Combined			
Mean(bias)	-0.009	0.000	-0.020
Std. dev.	0.057	0.022	0.128
Mean(std. error)	0.062	0.023	0.139
95% bootstrap CI cov.	0.970	0.947	0.976

Example

The Framingham risk model (FRM) has been used for population-wide CVD risk assessment. The model was developed based on several common clinical risk factors, including age, gender, total cholesterol level, high-density lipoprotein (HDL) cholesterol level, smoking, systolic blood pressure and high blood pressure treatment [25]. To improve the predictive capacity of the FRM, a new risk model has been developed recently using data from the Women’s Health Study [5], based on variables in the Framingham risk model and an inflammation marker, C-reactive protein (CRP). Prior to adapting the new model in routine practice, it is important

Table 5 Simulation results comparing apparent estimates and the 0.632 bootstrap for correcting overfitting. Note that Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Estimator		$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i > t)$	$\text{NRI}(t)$
Smooth IPW	True values	0.684	0.275	0.817
	Apparent			
	Mean(bias)	0.000	0.004	-0.007
	Std. dev.	0.036	0.028	0.108
	CI coverage	0.962	0.963	0.964
	0.632 bootstrap			
	Mean(bias)	-0.008	0.008	-0.032
	Std. dev.	0.034	0.027	0.102
	CI coverage	0.971	0.969	0.968
	Bootstrapped se			
Mean(std. error)	0.039	0.030	0.114	
SEM	Apparent			
	Mean(bias)	0.003	-0.001	0.009
	Std. dev.	0.023	0.025	0.072
	CI coverage	0.955	0.954	0.945
	0.632 bootstrap			
	Mean(bias)	0.005	-0.003	0.015
	Std. dev.	0.022	0.024	0.072
	CI coverage	0.953	0.962	0.937
Bootstrapped se				
Mean(std. error)	0.024	0.025	0.071	
Combined	Apparent			
	Mean(bias)	0.001	0.001	0.001
	Std. dev.	0.028	0.026	0.087
	CI coverage	0.982	0.969	0.975
	0.632 bootstrap			
	Mean(bias)	-0.002	0.003	-0.008
	Std. dev.	0.027	0.025	0.085
	CI coverage	0.989	0.975	0.983
Bootstrapped se				
Mean(std. error)	0.035	0.028	0.102	

to quantify its prediction performance, especially in comparison to that of FRM. We illustrate here how our proposed procedures can be used to evaluate and compare the clinical utility of the two risk models using an independent dataset from the Framingham Offspring Study [14].

The Framingham Offspring Study was established in 1971 with 5,124 participants who were monitored prospectively for epidemiological and genetic risk factors for CVD. We consider here 1,728 female participants who have CRP measurement and other clinical information at the second exam and are free of CVD at the time of examination. The average age of this subset was about 44 years (standard deviation = 10). The outcome we consider is the time from exam date to first

Table 6 NRI estimates for two risk models for predicting 10-year CVD risk among women in the Framingham offspring cohort. Note that KM = Kaplan-Meier estimator, IPW = Inverse Probability Weighted estimator, Smooth IPW = Smooth Inverse Probability Weighted estimator, SEM = Semiparametric estimator, Combined = Combined estimator, as defined in the text

Method		$\Pr(P_i - Q_i > 0 T_i \leq t)$	$\Pr(P_i - Q_i > 0 T_i < t)$	$\$NRI(t)$
KM				
	Est	0.483	0.508	-0.049
	SE	0.069	0.028	0.176
IPW				
	Est	0.478	0.508	-0.059
	SE	0.070	0.028	0.178
Smooth IPW				
	Est	0.480	0.508	-0.057
	SE	0.070	0.028	0.178
SEM				
	Est	0.587	0.503	0.167
	SE	0.015	0.026	0.067
Combined				
	Est	0.570	0.504	0.132
	SE	0.054	0.027	0.137

major CVD event including CVD-related death. During the followup period 269 participants were observed to encounter at least one CVD event and the 10-year event rate was about 4%. For illustration we chose $t = 10$ years as in [25]. For each individual, two risk scores were calculated: one based on the FRM (Model 1), combining information on age, systolic blood pressure, smoking status, high-density lipoprotein (HDL), total cholesterol, medication for hypertension; the other based on an algorithm developed in [5] (Model 2), with the addition of CRP concentration. We use Cox models to specify the relation between the time-to-CVD events and model scores (linear predictors from the models).

Both models are well calibrated based on calibration plots (not shown). For comparison, we first give AUC results and use the bootstrap to obtain confidence intervals. The AUC for an ROC curve at 10-years is 0.752 (95% CI: 0.721,0.783) for Model 1 and 0.758 (95% CI: 0.729,0.787) for Model 2. The difference between the two AUCs is not statistically significant: 0.006 (95% CI: -0.033, 0.046). We now investigate whether the new models reclassify patients in terms of their risks and CVD outcome at 10 years. We consider $NRI(10\text{-years})$ for such an evaluation using the methods described in section “Estimation”. Table 6 shows that estimates from the three nonparametric models are quite consistent, all indicating that the new model does not add significant improvement gauged by NRI. The semiparametric model, however, does indicate a significant incremental value with $NRI = 0.167$ ($SE = 0.067$), and the combined estimator indicates a similar magnitude of improvement, though not significant ($NRI = 0.132$, $SE = 0.137$).

Note that since we considered a continuous NRI with $u = v = 0$, the observed improvement at this magnitude may not be interpreted as clinically substantial. Since different conclusions could be reached depending on which estimation method is chosen, this analysis highlights the need to consider multiple robust approaches for calculating NRI.

Discussion

Net Reclassification Improvement provides an alternative tool for evaluating risk prediction models [18] beyond the traditional ROC curve framework. The concept has continued to gain popularity in the medical literature, yet its statistical properties have not been well studied to date in the statistical literature, and existing methods for calculating NRI under the failure time outcome setting are limited. In this manuscript, we provide a more thorough investigation of a variety of estimation procedures. Our proposed nonparametric and semiparametric estimators improve upon existing methods both in terms of robustness and efficiency under a variety of practical situations. Such improvement is quite important, since we observe that compared with other measures such as AUC, NRI estimates, in general, are not very stable with substantial variations in the estimators we have considered. The proposed procedures can be used for estimating both continuous NRI and NRI with pre-specified fixed categories. As illustrated in the example, the choice of estimation method can lead to different conclusions. In practice, the method chosen should depend on a number of important considerations including the likelihood that the model has been correctly specified and that the assumptions concerning censoring are correct. In addition, in situations where the new marker may be expensive or difficult to ascertain, an approach which considers both the risks and benefits of obtaining the marker should be considered in a decision-making process. We recommend such measures to be used in practice with caution. A thorough evaluation of a risk model should consider a wide spectrum of measures for assessing discrimination and calibration, and NRI may be better served as one of the summary measures to complement graphical displays of risk distributions [11]. All analyses were performed in R. Code for implementing the proposed procedures is available upon request.

Acknowledgements The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. The work is supported by grants U01-CA86368, P01-CA053996, R01- GM085047, R01-GM079330 awarded by the National Institutes of Health.

Appendix

Throughout, we assume that the joint density of (T, C, \mathbf{Y}) is twice continuously differentiable, \mathbf{Y} are bounded, and $1 > P(T > t) > 0, 1 > P(C > t) > 0$. The kernel function K is a symmetric probability density function with compact support and bounded second derivative. The bandwidth $h \rightarrow 0$ such that $nh^4 \rightarrow 0$. In addition, the estimator $\hat{\theta}_k$ converges to θ_{0k} for $k = 1, 2$ as $n \rightarrow \infty$ [13], where β_{k0} is the unique maximizer of the expected value of the corresponding partial likelihood and Λ_{k0} is the baseline cumulative hazard for $k = 1, 2$. We denote the parameter space for θ_k by Ω_k and assume that Ω_k is a compact set containing θ_{0k} . Furthermore, we assume that $\beta_2 \neq 0$ and note that $Q(\theta_2) = 1 - \exp\{\Lambda_{02}(t)e^{\beta_2^\top \mathbf{Y}(2)}\}$ and $P(\theta_1) = 1 - \exp\{\Lambda_{01}(t)e^{\beta_1^\top \mathbf{Y}(1)}\}$ are the respective limits of $Q(\hat{\theta}_2)$ and $P(\hat{\theta}_1)$, for any given $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(1)}$. The in-probability convergence of $Q(\hat{\theta}_2) \rightarrow Q(\theta_{02})$ and $P(\hat{\theta}_1)$ and $P(\theta_{01})$ are uniform in $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(1)}$ due to the convergence of $\hat{\theta} \rightarrow \theta_0 = (\theta_{01}^\top, \theta_{02}^\top)^\top$.

Asymptotic Properties of $\widetilde{\text{NRI}}(\hat{\theta}, t)$

From the same arguments as given in [3] and [7], it follows that we have the uniform consistency of $\hat{H}_q^{(t)}(t)$ to $H_q^{(t)}(t) = P\{C \geq t \mid Q(\theta_2) = q, \Delta(\theta) \in \mathcal{U}_t\}$, where $\mathcal{U}_1 = 1$ and $\mathcal{U}_\bullet = \{0, 1\}$, for $t = 1$ and \bullet . It follows, using the uniform law of large numbers [20], that

$$\sup_{\theta} |\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)| \rightarrow 0.$$

This along with the convergence of $\hat{\theta}$ to θ_0 implies that $\widetilde{\text{NRI}}(\hat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$.

Throughout, we will use the fact that $E\{\Delta_i(\theta)I(X_i \leq t)\delta_i\tilde{H}_{Q_i(\theta_2)}^{(1)}(X_i)^{-1} \mid Q_i(\theta_2) = q\} = P(\Delta_i(\theta) = 1, T_i \leq t \mid Q_i(\theta_2) = q)$ if either $C \perp T, \mathbf{Y}_{(2)}$ (model may be misspecified) or $Q(\theta_2) = \Pr(T \leq t \mid \mathbf{Y}_{(2)})$ i.e. the Cox model is correctly specified though censoring may be such that $C \perp T \mid \mathbf{Y}_{(2)}$ (double robustness). We first write the i.i.d representation of $\sqrt{n}[\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)]$ for any θ . Note that $\sqrt{n}\{\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)\} = 2\sqrt{n}\{\widetilde{\Pr}(\Delta(\theta) = 1 \mid T \leq t) - \Pr(\Delta(\theta) = 1 \mid T \leq t)\} - 2\sqrt{n}\{\widetilde{\Pr}(\Delta(\theta) = 1 \mid T > t) - \Pr(\Delta(\theta) = 1 \mid T > t)\}$. We first examine the initial component,

$$\widetilde{\Pr}(\Delta(\hat{\theta}) = 1 \mid T \leq t) = \frac{\sum_i \Delta(\hat{\theta})I(X_i \leq t)\delta_i/\tilde{H}_{Q_i(\hat{\theta}_2)}^{(1)}(X_i)}{\sum_i I(X_i \leq t)\delta_i/\tilde{H}_{Q_i(\hat{\theta}_2)}^{(\bullet)}(X_i)} \equiv \frac{\hat{N}(t, \hat{\theta}, \tilde{H})}{\hat{D}(t, \hat{\theta}, \tilde{H})}$$

where $\hat{N}(t, \theta, H) = n^{-1} \sum_i \Delta_i(\theta) I(X_i \leq t) \delta_i / H_{Q_i(\theta_2)}^{(1)}(X_i)$ and $\hat{D}(t, \theta, H) = n^{-1} \sum_i I(X_i \leq t) \delta_i / H_{Q_i(\theta_2)}^{(\bullet)}(X_i)$. Let $N(t, \theta) = \Pr(\Delta(\theta) = 1, T \leq t)$ and $D(t) = \Pr(T \leq t)$. Then by the uniform consistency of the IPW weights, we have

$$\sqrt{n}\{\tilde{\Pr}(\Delta(\theta) = 1|T \leq t) - \Pr(\Delta(\theta) = 1|T \leq t)\} \approx \sqrt{n}\{\hat{N}(t, \theta, \tilde{H})D(t) - N(t, \theta)\hat{D}(t, \theta, \tilde{H})\}/D(t)^2.$$

Examining the numerator, $\sqrt{n}\{\hat{N}(t, \theta, \tilde{H})D(t) - N(t, \theta)\hat{D}(t, \theta, \tilde{H})\} = \sqrt{n}\{(1) + (2) - (3)\}$ where (1) = $\hat{N}(t, \theta, H)D(t) - \hat{D}(t, \theta, H)N(t, \theta)$, (2) = $\hat{N}(t, \theta, \tilde{H})D(t) - \hat{N}(t, \theta, H)D(t)$, and (3) = $[N(t, \theta)\hat{D}(t, \theta, \tilde{H}) - \hat{D}(t, \theta, H)N(t, \theta)]$. Note that

$$(1) = \sqrt{n}\{\hat{N}(t, \theta, H)D(t) - \hat{D}(t, \theta, H)N(t, \theta)\} = n^{-\frac{1}{2}} \sum U_{1i}(t), \quad \text{where}$$

$$U_{1i}(t) = \frac{I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \Delta_i(\theta) D(t) - \frac{I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(\bullet)}(X_i)} N(t, \theta)$$

Using a Taylor series expansion, Lemma A.3 of [2] and the asymptotic expansion for $\hat{\Lambda}_q(t)$ given in [8],

$$(2) = D(t)\sqrt{n}\{\hat{N}(t, \theta, \tilde{H}) - \hat{N}(t, \theta, H)\}$$

$$= D(t)n^{-1/2} \sum_i \frac{\Delta_i(\theta) I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \left[\frac{H_{Q_i(\theta_2)}^{(1)}(X_i)}{\tilde{H}_{Q_i(\theta_2)}^{(1)}(X_i)} - 1 \right]$$

$$= D(t)n^{-1/2} \int \int_0^t \left[\frac{H_q^{(1)}(s)}{\tilde{H}_q^{(1)}(s)} - 1 \right] d \sum_i \frac{\Delta_i(\theta) \delta_i I(X_i \leq s, Q_i(\theta_2) \leq q)}{H_{Q_i(\theta_2)}^{(1)}(X_i)}$$

$$\approx D(t) \int \int_0^t \sqrt{n} \left[\hat{\Lambda}_q^{(1)}(s) - \Lambda_q^{(1)}(s) \right] d \left\{ \frac{1}{n} \sum_i \frac{\Delta_i(\theta) \delta_i I(X_i \leq s, Q_i(\theta_2) \leq q)}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \right\}$$

$$\approx D(t) \int \int_0^t \left[n^{-\frac{1}{2}} \sum K_h \{q - Q_i(\theta_2)\} M_{C_q}^{(1)}(s, X_i, \delta_i) \right] dP(\Delta(\theta)=1, T \leq t, Q(\theta_2) \leq q)$$

where

$$M_{C_q}^{(1)}(t, X_i, \delta_i) = \int_0^t \frac{dN_{C_i}(s) - I(X_i \geq s) d\Lambda_q^{(1)}(s)}{\pi_s^{(1)}(q)}.$$

Now by a change of variable, $\psi = \frac{q - Q_i(\theta_2)}{h}$ and $f(t, q) \equiv \partial^2 P(\Delta(\theta) = 1, T \leq t, Q(\theta_2) \leq q) / \partial t \partial q$,

$$\begin{aligned}
 (2) &\approx D(t) \int \int_0^t \sqrt{n} \left[\frac{1}{n} \sum K(\psi) M_{C(\psi h + Q_i(\theta_2))}(s, X_i, \delta_i) \right] f(t, \psi h + Q_i) ds d\psi \\
 &= D(t) n^{-1/2} \sum \int \int_0^t K(\psi) a\{s, h\psi + Q_i(\theta_2), X_i\} ds d\psi = n^{-1/2} \sum U_{2i}(t),
 \end{aligned}$$

where $U_{2i}(t) = D(t) \int_0^t a(s, q^*, X_i) ds$ and $a(t, q, X_i) = M_{Cq^*}(t, X_i, \delta_i) f(t, q^*)$. Similar arguments can be used to obtain an asymptotic expansion for (3) as $(3) \approx n^{-1/2} \sum U_{3i}(t)$ and therefore, the numerator, $\sqrt{n} [\hat{N}(t, \theta, \hat{H})D(t) - N(t, \theta)\hat{D}(t, \hat{H})] \approx n^{-1/2} \sum \{U_{1i}(t) + U_{2i}(t) + U_{3i}(t)\}$. The same arguments as given above can be used to obtain an asymptotic expansion for $\sqrt{n}\{\hat{\text{Pr}}(\Delta(\theta) = 1|T > t) - \text{Pr}(\Delta(\theta) = 1|T > t)\}$ as $n^{-1/2} \sum_{i=1}^n D(t)^{-2} \{U_{-1i}(t) + U_{-2i}(t) + U_{-3i}(t)\}$ where $D(t)^-, U_{-1i}(t), U_{-2i}(t)$, and $U_{-3i}(t)$ are defined similarly to $D(t), U_{1i}(t), U_{2i}(t)$, and $U_{3i}(t)$ with $T \leq t$ replaced with $T > t$. Therefore, $\sqrt{n}\{\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)\} \approx n^{-1/2} \sum_{i=1}^n 2\{D(t)^{-2} \{U_{1i}(t) + U_{2i}(t) + U_{3i}(t)\} - D(t)^{-2} \{U_{-1i}(t) + U_{-2i}(t) + U_{-3i}(t)\}\} = n^{-1/2} \sum_{i=1}^n \eta_i(t)$.

Note that regardless of correct model specification, $\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} \sum \psi_i + o_p(1)$ where ψ_i are i.i.d mean zero random variables by Lin and Wei [16] and Uno et al. [24]. Using a Taylor series approximation and the i.i.d representation of $\sqrt{n}[\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)]$ for any θ , we can write $\widetilde{\mathcal{W}}(t) = \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)]$ as a sum of i.i.d terms, $n^{-1/2} \sum_{i=1}^n \varepsilon_i(t)$ defined below.

$$\begin{aligned}
 &\sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)] \\
 &= \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\hat{\theta}, t) + \text{NRI}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)] \\
 &\approx \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\hat{\theta}, t) + \frac{\partial \text{NRI}(t)}{\partial \theta} \Big|_{\theta_0} (\hat{\theta} - \theta_0)] \\
 &= \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\hat{\theta}, t)] + \sqrt{n}(\hat{\theta} - \theta_0) \frac{\partial \text{NRI}(t)}{\partial \theta} \Big|_{\theta_0} \\
 &\approx \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\hat{\theta}, t)] + n^{-1/2} \sum \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} \Big|_{\theta_0} \\
 &\approx n^{-1/2} \sum_{i=1}^n \eta_i(t) + n^{-1/2} \sum \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} \Big|_{\theta_0} \\
 &= n^{-1/2} \sum_{i=1}^n \varepsilon_i(t)
 \end{aligned}$$

where $\varepsilon_i(u, v, t) = \eta_i(u, v, t) + \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} \Big|_{\theta_0}$. By a functional central limit theorem of [20], the process $\widetilde{\mathcal{W}}(t)$ converges weakly to a mean zero Gaussian process in t .

Asymptotic Properties of $\widehat{\text{NRI}}(\hat{\theta}, t)$

Recall that we assume the Cox model is correctly specified and thus, $Q(\theta_2) = Q(\theta_2, t, \mathbf{Y}_{(2)}) = \Pr(T \leq t | Y_{(2)}) = 1 - \exp\{\Lambda_{02}(t)e^{\beta_2^T Y_{(2)}}\}$ and $S_{Q_i(\theta_2)}(t) = \Pr(T > t | Y_{(2)}) = \exp\{\Lambda_{02}(t)e^{\beta_2^T Y_{(2)}}\}$. To derive asymptotic properties of $\widehat{\text{NRI}}(\hat{\theta}, t)$ we assume the same regularity conditions as in [1]. The uniform consistency of $Q(\hat{\theta}_2, t, \mathbf{Y}_{(2)})$ for $Q(\theta_2, t, \mathbf{Y}_{(2)})$ in t and $\mathbf{Y}_{(2)}$ follows directly from the uniform consistency of $\hat{\Lambda}_{02}(t)$ and $\hat{\beta}_2$. It follows from the uniform law of large numbers [20] that $\widehat{\text{NRI}}(\hat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$. Andersen and Gill [1] show that $\sqrt{n}(\hat{\beta}_2 - \beta_{02})$ is a normal random variable and $\sqrt{n}(\hat{\Lambda}_{02}(t) - \Lambda_{02}(t))$ converges to a Gaussian process. By the functional delta method it can be shown that $\sqrt{n}\{Q(\hat{\theta}_2, t, \mathbf{Y}_{(2)}) - Q(\theta_2, t, \mathbf{Y}_{(2)})\}$ converges to a zero mean Gaussian process in t and $\mathbf{Y}_{(2)}$. Similar to the derivation for $\widetilde{\text{NRI}}(\hat{\theta}, t)$, it can be shown that the process $\widetilde{\mathcal{N}}(t) = \sqrt{n}[\widetilde{\text{NRI}}(\hat{\theta}, t) - \text{NRI}(\theta_0, t)]$ is asymptotically equivalent to $n^{-1/2} \sum_{i=1}^n \zeta_i(u, v, t)$. In particular, for a fixed θ , $\sqrt{n}\{\widetilde{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)\} \approx n^{-1/2} \sum_{i=1}^n \eta_i^*(t)$ where $\eta_i^*(t) = 2[D(t)]^{-2}\{\Delta_i(\theta)Q_i(\theta_2) - \Pr(\Delta_i(\theta) = 1 | T_i \leq t)Q_i(\theta_2)\} - D(t)^{-2}\{\Delta_i(\theta)[1 - Q_i(\theta_2)] - \Pr(\Delta_i(\theta) = 1 | T_i > t)[1 - Q_i(\theta_2)]\}$. Thus, $\widetilde{\mathcal{N}}(t) \approx n^{-1/2} \sum_{i=1}^n \zeta_i(t)$ where $\zeta_i(u, v, t) = \eta_i^*(t) + \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0}$. Once again, using a functional central limit theorem, this implies that $\widetilde{\mathcal{N}}(t)$ converges to a Gaussian process with mean zero.

References

1. Andersen, P., Gill, R.: Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982)
2. Biliyas, Y., Gu, M., Ying, Z.: Towards a general asymptotic theory for cox model with staggered entry. *Ann. Stat.* **25**, 662–682 (1997)
3. Cai, T., Tian, L., Uno, H., Solomon, S., Wei, L.: Calibrating parametric subject-specific risk estimation. *Biometrika* **97**, 389–404 (2010)
4. Cook, N.: Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928 (2007)
5. Cook, N., Buring, J., Ridker, P.: The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Ann. Intern. Med.* **145**, 21 (2006)
6. Cui, J.: Overview of risk prediction models in cardiovascular disease research. *Ann. Epidemiol.* **19**, 711–717 (2009)
7. Dabrowska, D.: Smoothed cox regression. *Ann. Stat.* **25**, 1510–1540 (1997)
8. Du, Y., Akritas, M.: Uniform strong representation of the conditional kaplan-meier process. *Math. Methods Stat.* **11**, 152–182 (2002)
9. Efron, B., Tibshirani, R.: Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997)
10. Gail, M., Brinton, L., Byar, D., Corle, D., Green, S., Schairer, C., Mulvihill, J.: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI J. Natl. Cancer Inst.* **81**, 1879 (1989)

11. Gu, W., Pepe, M.: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **5**, 27 (2009)
12. Hemann, B., Bimson, W., Taylor, A.: The framingham risk score: an appraisal of its benefits and limitations. *Am. Heart Hosp. J.* **5**, 91–96 (2007)
13. Hjort, N.: On inference in parametric survival data models. *Int. Stat. Rev. (Revue Internationale de Statistique)* **60**, 355–387 (1992)
14. Kannel, W., Feinleib, M., McNamara, P., Garrison, R., Castelli, W.: An investigation of coronary heart disease in families. *Am. J. Epidemiol.* **110**, 281 (1979)
15. Khot, U., Khot, M., Bajzer, C., Sapp, S., Ohman, E., Brener, S., Ellis, S., Lincoff, A., Topol, E.: Prevalence of conventional risk factors in patients with coronary heart disease. *JAMA J. Am. Med. Assoc.* **290**, 898–904 (2003)
16. Lin, D., Wei, L.: The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.* **84**, 1074–1078 (1989)
17. Lloyd-Jones, D.: Cardiovascular risk prediction. *Circulation* **121**, 1768–1777 (2010)
18. Pencina, M., D'Agostino Sr, R., D'Agostino Jr, R., Vasan, R.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172 (2008)
19. Pencina, M., D'Agostino Sr, R., Steyerberg, E.: Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **30**, 11–21 (2011)
20. Pollard, D.: *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward (1990)
21. Satten, G., Datta, S.: The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *Am. Stat.* **55**, 207–210 (2001)
22. Tian, L., Cai, T., Goetghebeur, E., Wei, L.: Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**, 297–311 (2007)
23. Uno, H., Cai, T., Tian, L., Wei, L.: Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**, 527–537 (2007)
24. Uno, H., Tian, L., Cai, T., Kohane, I., Wei, L.: A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat. in Med.* **32**, 2430–2442 (2013)
25. Wilson, P., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H., Kannel, W.: Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837 (1998)

Erratum to: Risk Assessment and Evaluation of Predictions

Mei-Ling Ting Lee, Mitchell Gail, Ruth Pfeiffer, Glen Satten, Tianxi Cai and Axel Gandy

Erratum to:

Risk Assessment and Evaluation of Predictions

Mei-Ling Ting Lee, Mitchell Gail, Ruth Pfeiffer, Glen Satten, Tianxi Cai and Axel Gandy DOI 10.1007/978-1-4614-8981-8

The volume number for this book has been changed from 210 to 215.

The online version of the book can be found at:
[http://dx.doi.org/ 10.1007/978-1-4614-8981-8](http://dx.doi.org/10.1007/978-1-4614-8981-8)

M.T. Lee, M. Gail, R. Pfeiffer, G. Satten, T. Cai, A. Gandy (eds.), *Risk Assessment and Evaluation of Predictions*, DOI 10.1007/978-1-4614-8981-8_22,
© 2013 Springer Science+Business Media Dordrecht

E1