OR ESSENTIALS

OPERATIONAL RESEARCH FOR EMERGENCY PLANNING IN HEALTHCARE: VOLUME 1

EDITED BY
NAVONIL MUSTAFEE

Operational Research for Emergency Planning in Healthcare: Volume 1

The OR Essentials series
Series editor: **Simon JE Taylor**, Brunel University, UK

The *OR Essentials* series presents a unique cross-section of high quality research work fundamental to understanding contemporary issues and research across a range of Operational Research (OR) topics. It brings together some of the best research papers from the highly respected journals of the Operational Research Society, also published by Palgrave Macmillan.

OR deals with the use of advanced analytical methods to support better decision making. As a multidisciplinary field, it has strong links to management science, decision science, computer science and has practical applications in areas such as engineering, manufacturing, commerce, healthcare and defence.

OR has long-standing historical roots. However, as a modern discipline its origins lie in the years immediately before World War II when mathematical techniques were developed to address urgent defence problems. Now it is commonplace and a key discipline taught in universities across the world, at undergraduate and postgraduate levels. There are several international societies dedicated to the advancement of OR (e.g. the Operational Research Society and INFORMS – The Institute for Operations Research and the Management Sciences) and there are many high quality peer-reviewed journals dedicated to the topic.

The *OR Essentials* books are a vital reference tool for students, academics, and industry practitioners, providing easy access to top research papers on cutting-edge topics within the field of Operational Research.

Titles include:
Navonil Mustafee (*editor*)
OPERATIONAL RESEARCH FOR EMERGENCY PLANNING IN HEALTHCARE: VOLUME 1

Simon JE Taylor (*editor*)
AGENT-BASED MODELLING AND SIMULATION

Roger A Forder (*editor*)
OR, DEFENCE AND SECURITY

Mike Wright (*editor*)
OPERATIONAL RESEARCH APPLIED TO SPORTS

John S Edwards (*editor*)
THE ESSENTIALS OF KNOWLEDGE MANAGEMENT

Navonil Mustafee (*editor*)
OPERATIONAL RESEARCH FOR EMERGENCY PLANNING IN HEALTHCARE: VOLUME 2

The OR Essentials series
Series Standing Order ISBN 978–1–137–45360–0
(*outside North America only*)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

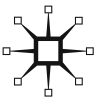
Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke RG21 6XS, UK

Operational Research for Emergency Planning in Healthcare: Volume 1

Edited by

Navonil Mustafee
Senior Lecturer in Operations Management,
University of Exeter Business School, UK

palgrave
macmillan



Selection, Chapter 1 and editorial matter © Navonil Mustafee 2016
Individual chapters © Operational Research Society 2016

Softcover reprint of the hardcover 1st edition 2016 978-1-137-53567-2

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No portion of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

Any person who does any unauthorized act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The authors have asserted their rights to be identified as the authors of this work in accordance with the Copyright, Designs and Patents Act 1988.

First published 2016 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978-1-349-56809-3 ISBN 978-1-137-53569-6 (eBook)
DOI 10.1057/9781137535696

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

Typeset by MPS Limited, Chennai, India.

Contents

List of Figures and Tables vii

1 A Synthesis of Operational Research for Emergency Planning in Healthcare through the Triple Lens of Technique-Domain-Context 1
N. Mustafee

Part I OR for Locating Emergency Services

2 Locating Emergency Services with Different Priorities: The Priority Queuing Covering Location Problem 15
F. Silva and D. Serra

3 Decision Support Tools for Ambulance Dispatch and Relocation 36
T. Andersson and P. Värbrand

4 A Study of Situationally Aware Routing for Emergency Responders 52
M. J. Henchey, R. Batta, A. Blatt, M. Flanigan and K. Majka

Part II OR for Operational Planning in Emergency Services

5 Multi-Criteria Approach Using Simulation-Based Balanced Scorecard for Supporting Decisions in Health-care Facilities: An Emergency Department Case Study 85
W. Abo-Hamad and A. Arisha

6 Combining Data Mining and Discrete Event Simulation for a Value-Added View of a Hospital Emergency Department 119
R. Ceglowski, L. Churilov and J. Wasserthiel

7 Uncovering Effective Process Improvement Strategies in an Emergency Department Using Discrete Event Simulation 139
O. H. Choon, Z. Dali, P. T. Beng and C. P. Y. Magdalene

8 Improving the Design and Operation of an Integrated Emergency Post via Simulation 164
N. J. Borgman, M. R. K. Mes, I. M. H. Vliegen and E. W. Hans

9 A Simulation Case Study to Improve Staffing Decisions at Mass Immunization Clinics for Pandemic Influenza 190
M. F. Beeler, D. M. Aleman and M. W. Carter

10 Modelling Treatment Effects in the HIV/AIDS Epidemic 224
J. D. Griffiths, Z. F. Lawson and J. E. Williams

Part III OR for Inventory Management in Emergency Services

11 Impact of the Influenza Season on a Hospital from a Pharmaceutical Inventory Management Perspective 249
A. R. Vila-Parrish, J. S. Ivy and B. He

12 Perishable Inventory Management System with a Minimum Volume Constraint 288
Z. Shen, M. Dessouky and F. Ordonez

13 A Bayesian Decision Model with Hurricane Forecast Updates for Emergency Supplies Inventory Management 330
S. Taskin and E. J. Lodree, Jr

14 Using Simulation to Improve the Blood Supply Chain 353
K. Katsaliaki and S. C. Brailsford

Index 373

List of Figures and Tables

Figures

2.1 GRASP pseudocode 23
2.2 Construction phase pseudocode 24
2.3 Local search phase pseudocode 26
2.4 Reactive GRASP pseudocode 27
2.5 Objective evaluation pseudocode 28
2.6 Locating the nine centres 30
2.7 Locating the nine centres 32
3.1 Illustrating the tree-search heuristic that solves the DYNAROC model 44
3.2 The ambulance operations simulator 46
3.3 Number of Prio 1 calls served (in %) when using a threshold value of 0, 0.277, and 0.923, respectively 49
4.1 Information fusion approach to ITS 54
4.2 EMS first responder locations and crash locations 56
4.3 (a) Unconstrained shortest path (b) and VCSP 66
4.4 Routes from dual optimization at $\mu = 0$ (a) and $\mu = \infty$ (b) 80
4.5 Optimal Route from $\mu^* = 0.43$ (c) and k-shortest route procedure (d) 81
5.1 An overview of the integrated framework 90
5.2 ED physical layout and main care areas 94
5.3 Detailed patient flow through the ED 96
5.4 Mapping main ED processes 98
5.5 The ED Balanced Scorecard 99
5.6 Validation of simulation results against actual data 103
5.7 Alternatives value tree 105
5.8 ED KPIs 108
5.9 Full AHP weighted value tree 112

5.10 The ED performance for all the scenarios against the current ED 113
5.11 The change in ED performance with average LOS for all scenarios 113
5.12 ED performance showing staff utilisation and burnout levels 114
6.1 Screenshot of the SOM treatment clusters in Viscovery SOMine 126
6.2 Schematic view of the simulation 131
6.3 Comparison of patient bed times 132
7.1 Historical ED attendance and proportions of four patient acuity categories (see Table 7.1 for their definitions) 140
7.2 Simplified schematic diagram of patient flow at ED 143
7.3 Daily load of ED 146
7.4 Arrival patterns of ED patients 149
7.5 TFC times of non-fever P2 and P3 patients 150
7.6 TTs of non-fever P2 and P3 patients 151
7.7 Baseline hourly profiles of P2 and P3 patients and doctors 153
7.8 Baseline and new hourly profiles of P2 and P3 patients and doctors 154
7.9 Hourly profiles of P3 patients and doctors in the second new manpower schedules 155
8.1 Patient flow and processes at the IEP Almelo (Mes and Bruens, 2012) 171
8.2 Significant ED and GPP main and two-way interaction effect confidence intervals ($\alpha = 0.05$) (A × B denotes the interaction effect between interventions A and B) 179
8.3 Average LOS outcome confidence intervals (95%) per intervention set 182
9.1 Flowchart of patient movement through MIC in Niagara Region, ON 195
9.2 Interarrival times in minutes for patient batches in two scenarios 197
9.3 Histogram showing positive skew for injection times recorded in electronic system, 26–28 October 2009 202

9.4 Registration staff productivity by time of day 204

9.5 Comparison of hourly productivity, 26–28 October and November 205

9.6 Vaccinations and average time in system *versus* patient interarrival time adjustment 211

9.7 Vaccinations and average time in system *versus* patient wait time tolerance adjustment 212

10.1 Cardiff HIV/AIDS epidemic model 227

10.2 Sequence of parameter estimation 233

10.3 Fit between observed and expected AIDS incidence, 1979–2002 236

10.4 Fit between observed and expected HIV incidence, 1995–2002 236

10.5 Comparison between observed and expected deaths from AIDS, 1985–2002 238

10.6 HIV prevalence investigation and comparison, 1996–2002 239

10.7 Extrapolated HIV, AIDS and deaths from AIDS up to 2008 241

10.8 Extrapolated prevalence in sub-populations up to 2008 241

10.9 Extrapolated prevalence in sub-populations, given a change in behaviour, up to 2012 243

11.1 State transition diagram that describes a patient's condition over time, given that the patient is admitted to the hospital in state z_1 or z_2, and leaves the hospital from state H or D 255

11.2 Effect of GAR on total hospitalization and total death 262

11.3 Flow chart of the hospital pharmacy inventory system model 263

11.4 Timeline of events with corresponding variable definitions 265

11.5 The SIR model in Arena 281

11.6 The process of generating inpatients 281

11.7 The pharmacy decision-making process 283

11.8 The medication allocation process 284

11.9	The signal module to trigger the health status update of the patients at the end of each period	285
11.10	The status update module	286
11.11	Details of sub-model 1, which updates the health status of patients in age-group 1	287
12.1	Perishable inventory system with small $I_{\min}$	296
12.2	Illustration for the modified EMQ model	297
12.3	All 5 cases of possible scenarios	299
12.4	Graph illustration for an extreme boundary case	300
12.5	Graph illustration for case 1	304
12.6	Plot of non-continuous, non-differentiable total cost function with minimum inventory	308
12.7	Graph illustration for the global optimum proof	309
12.8	Profit v.s. Y with $Q_{\max} = 1600$ at different $I_{\min}$ level	315
12.9	Profit v.s. Y which allows salvage at most 300 million pills	316
12.10	Profit v.s. Y with $Q_{\max} = 1600$ at different unit government-paid price	318
12.11	Illustration of the different slopes in the profit v.s. Y plot	319
12.12	Graph illustration for case 2	323
12.13	Graph illustration for case 4	323
12.14	Graph illustration for case 5	324
12.15	Graph illustration for case 3	325
12.16	Graph illustration for the boundary value of case 3	326
13.1	NHC's tropical cyclone wind speed probability product	343
13.2	Wind speed probability map at $t = 0$ h	346
13.3	Wind speed probability map at $t = 120$ h	347
14.1	Flow chart of blood units and orders in the blood supply chain	355
14.2	Snapshot of simulation model	360
14.3	Phases of a simulation run with no initial contents. (a) Average age of all units in stock in the Hospital Blood Bank (HBB). (b) Number of units in stock in the HBB	364

Tables

1.1	An outline of volume 1 presented through the triple lens of technique-domain-context	5
1.2	An outline of volume 2 presented through the triple lens of technique-domain-context	7
2.1	Computational results—deviations	29
2.2	Computational results—average computing times	30
2.3	Heuristic's results for one server per centre and two priorities	31
2.4	Heuristic's results for one server per centre and one priority	31
2.5	Heuristic's results for one server per centre and two priorities with closest assignments	33
3.1	The ambulance dispatch algorithm	41
3.2	The tree search algorithm that finds solutions to DYNAROC	43
3.3	The average number of relocated ambulances per day	48
4.1	Conditions for event-based congestion testing	62
4.2	Summary of experimental scenario results	68
4.3	"Insensitive" EMS in high congestion experimental results	74
4.4	"Sensitive" EMS in high congestion experimental results	74
4.5	Results of three distinct regions of weather	75
4.6	Event based congestion results of the two test configurations	76
4.7	VCSP Solution Example Summary	82
5.1	Summary of the analysis of patient's records	101
5.2	Analysis of patient allocation within the ED	102
5.3	ED senior managers' rating of alternatives for each criterion	106
5.4	The relative importance of the evaluation criteria	107
5.5	Aggregated weights and values for 'percentage of patients treated'	107
5.6	Final scores and alternative rankings using SMART procedure	107

5.7	Simulation variables for baseline system and scenarios	109
5.8	Simulation results of Scenarios 1, 2, and 3	110
5.9	AHP comparison matrix for main KPIs in ED performance criteria	110
5.10	The comparison matrix for the KPIs of the patient throughput criterion	111
5.11	Weighted results for all scenarios against the baseline scenario	113
6.1	Selected treatment clusters showing proportion of patients (1.00 = 100%) in each cluster that undergo each of the medical procedures (Ceglowski *et al*, 2004c)	127
6.2	Number of patient types existing in the data	130
6.3	Patient types with the highest cumulative weighted impact on the ED	134
7.1	Patient acuity categorization at the ED	145
7.2	List of data and their sources for DES model development	148
7.3	Numbers of P2 and P3 doctors in existing shifts in the second new doctor manpower schedule	157
7.4	Projected reduction in TT of patients under selected scenarios	158
8.1	Identified and categorized interventions	176
8.2	Experiments per intervention category	177
8.3	Intervention sets	181
8.4	Absolute outcomes (seconds) per intervention with significant outcomes (α=0.05) made bold	184
9.1	Percentage of patients per age group, by vaccine priority restriction scenario (Statistics Canada, 2006; Gentry, 2010)	198
9.2	Percentage of patients per family batch size, by vaccine priority restriction scenario (Gentry, 2010)	199
9.3	Effects of patient age on service time (minutes), relative to service times for the 0–4 age group	203
9.4	Effects of family group arrival size on registration time per-patient	203
9.5	MIC operating costs (CAD)	206
9.6	Pandemic influenza transmission hazard rates (λ) per minute (Haber *et al*, 2007)	207

9.7	Baseline scenario results, 21 replications	209
9.8	Factors examined in experiment	213
9.9	Factors affecting number vaccinated, ATS, expected infections, direct costs per vaccination, and total societal costs per vaccination	214
11.1	Expressions for the total number of susceptible at times t and $t+\lambda$, and number of new cases during time length of λ	253
11.2	Input parameters for the SIR model	253
11.3	Variables used to define a patient's age-group, risk group, and final health outcome	255
11.4	Assumed GAR for different age-groups (uniformly distributed) (Meltzer *et al*, 1999)	256
11.5	U.S. population and total number of clinical cases by age-groups	257
11.6	Rates of health outcomes by age and risk group (Meltzer *et al*, 1999)	257
11.7	Proportions of hospitalized but finally recovered patients by age and risk groups	258
11.8	Percentage of non-high-risk inpatients (y_u), inpatients admitted with non-high/high-risk condition and finally recovered (p_u/q_u), and death rate (d_u)	259
11.9	Results of transition probability matrix for the three age-groups	260
11.10	Comparison between simulation results and other literature	261
11.11	Sensitivity analysis of GAR	261
11.12	Notation	264
11.13	Cost coefficients	271
11.14	Patients' demand rate μ_{ij}	272
11.15	Cost structure of Policies I and II in base scenario for 200 periods	272
11.16	Performance of proposed policies under different GARs	273
11.17	Comparison of total cost of the proposed policies with respect to input costs, raw holding, expired finished good, raw and finished good stockout, ordering and production costs	274

12.1 Estimated numerical value of the parameters used in the experiment 313
12.2 Model comparison of decomposed cost (mil $/year) 314
12.3 Model comparison of idle machine capacity per year 314
13.1 Demand scenario probabilities 339
13.2 Simulated hurricane track based on the CLIPER model 346
13.3 Demand information 348
13.4 Example results 348
13.5 Example 1 calculations based on the insurance policy framework 349
13.6 Impact of ρ and VaDR on production quantities 350
14.1 Mismatching donors against patients' group 357
14.2 National Blood Service, Price List 2004/05 358
14.3 Validated model results for wastage figures according to the length of run and number of runs 365
14.4 Percentage change in the performance measurements achieved with different policies. The model runs for 26 weeks for a hospital with about 630 orders of RBC units per month 367
14.5 The baseline scenario compared with the improved policy 369
14.6 Percentage improvements in the performance indices by applying the improved policy, compared with the baseline 369

1

A Synthesis of Operational Research for Emergency Planning in Healthcare through the Triple Lens of Technique-Domain-Context

N. Mustafee

1.1 Introduction

Operational Research (OR) is the discipline that applies analytical methods, most of which are mathematical, statistical and computational in nature, to arrive at optimal or near-optimal solutions to complex decision-making problems. OR has strong ties with disciplines such as Mathematics, Business and Management, Industrial Engineering and Computer Science. The growth of the OR field has meant greater specialization and division into subfields (Miser, 2000) which sometime lack distinct boundaries. A topical example here is the emerging research area of Big Data, Business Intelligence and Analytics. They enable introspection of large corpus of data and provide both data-driven and predictive insights for organizational decision-making; this again is a fertile area of enquiry for OR research, as also for Computer Science (e.g., technologies associated with Big Data storage and retrieval), Information Systems (e.g., business intelligence dashboards) and Mathematics and Statistics (e.g., algorithms for data mining, predictive analytics). The fuzzy boundaries of the discipline are further evidenced by the fact that the terms 'OR' and 'MS' (Management Science) have not been absorbed as one but both exist conjoint and make the well-known OR/MS discipline. The lack of boundaries is a reflection of the interdisciplinary nature of the discipline.

1.2 The triple lens

How does one describe the sheer enormity of OR content and how can it be logically structured? One possible approach is through the

identification of broad categories of *OR techniques*, their *domain of application (application area)* and *application context*. In most application areas OR can be applied to a multitude of problems with the objective of enabling the stakeholders to make better and more informed decisions; the context of application identifies the specific purpose of using OR techniques. Categorization of OR literature is important because it is arguable that the readers must not only build an understanding of the analytical methods that are discussed in this book, they should in addition possess domain knowledge which is specific to their area of work or study, and further, they should carefully consider the context of application since the choice of particular OR techniques can be largely dictated by specific problems that are to be solved! It therefore follows that this triple lens of technique-application-context will help readers to better appreciate the practical application of OR, and the 26 studies that have been presented in Volumes 1 and 2 of this book have been structured keeping this in mind (see Tables 1.1 and 1.2).

A few examples of the broad categories of technique, application and context are presented next. It is based on research that focuses on the development of a classification scheme for OR/MS (Mustafee and Katsaliaki, n.d). As the subject matter of this book is healthcare some of the ensuing narrative will be specific to the application of OR in health.

1.2.1 OR technique

At the most basic level the modeling techniques can be classified as either, (a) Qualitative or 'Soft OR', and (b) Quantitative or 'Hard OR'; (a) Soft OR includes approaches from systems thinking (e.g., *Critical Systems Thinking* and *Theory of Constraints*) and problem-structuring methods (e.g., *Cognitive Mapping*, *Collaborative Planning*, *Qualitative System Dynamics*, *Soft Systems Methodology*, *Strategic Choice Approach* and *SODA*), (b) Quantitative modeling techniques can be grouped under a far wider array of sub-categories, for example, *Algorithms*, *Artificial Intelligence*, *Combinatorial Analysis*, *Complexity Theory, Data Envelopment Analysis, Forecasting, Fuzzy Logic, Inventory Theory, Multicriteria decision-making, Optimization, Probability, Simulation, Statistics*. Specific OR techniques such as *Constraint Programming, Bilevel Programming, Integer Programming, Approximate Dynamic Programming, Linear Programming, Simulated Annealing, Tabu Search, Minimum Spanning Trees* can all be grouped under the sub-category *Optimiz*ation. Similarly, *Logistic Regression, Structural Equation Modeling, Ranking and Selection, Importance Sampling, Bayesian Statistics, Bootstrapping, Importance Sampling,* etc. are specific methods that are related to *Statistics*. Only two sub-categories (*Optimization* and

Statistics) have been discussed and there are already more than a dozen techniques! The book mostly focuses on the application of 'Hard OR' techniques like *Mathematical Modeling, Dynamic Programming, Data Mining, Discrete-event Simulation* and *Quantitative System Dynamics*, with only a couple of studies reporting the use of 'Soft OR' techniques like *Soft Systems Methodology* and *Qualitative System Dynamics*.

1.2.2 Application context

It is important to gain an understanding of specific problem scenarios to which OR techniques are being frequently applied. Following the categorization approach presented above, Mustafee and Katsaliaki (n.d) have identified the common application scenarios of OR to include *Cost Analysis, Decision-Making, Emergency Response, Maintenance, Marketing, Operations Management, Quality Management, Reliability, Scheduling, Strategy, Facility Planning and Design, Inventory Management, Policy Analysis, Supply Chain Management, etc*. These application categories can be further sub-divided into a hierarchy of sub-categories, with the subsequent levels representing a shift from the more general to the more specific! For example, *Decision-Making* could include the application of OR techniques for *Decision Analysis, Cost-Benefit Analysis, Decision Support, Group Decision and Negotiations, Conflict Analysis*, etc. The application of OR in the general area *Economics* could take the form of OR for *Econometric Analysis, Actions/Bidding, Experimental Economics,* etc. Similarly, OR techniques have been specifically applied for *Disruption Management* and *Evacuation Planning* (under *Emergency Response*) and in the context of *Accidents, Epidemiology and Health Screening* (under *Healthcare Management*). The book covers the application of OR for Emergency Planning and includes chapters that have covered specific aspects related to location and routing of emergency services, inventory management, healthcare supply chain management, multi-agency planning, healthcare policy assessment, public health education, etc.

1.2.3 Application area

OR has been widely applied in application domains/industries like *Agriculture, Audit, Defence, Finance, Education, e-Commence, Manufacturing, Public Service, Sports and Recreation, Service Industries, Transporting, Utilities, Health Services*, etc. Examples of specific industries that have benefited from OR include, for example, under the category *Transportation – Airlines, Container Terminals, Railways, Shipping, Freight Transportations, etc.*; under *Utilities – Electricity, Power, Energy; under Finance – Bank/Banking, Financial Institutions, Insurance*; under *Health Services* mainly *Hospitals*. The book

consists of chapters that have applied OR techniques in relation to A&E departments, first responder and ambulance services, mass immunization clinics and emergency response pertaining to public health.

1.3 Structure of the book: a triple lens approach for the study of OR for emergency planning in healthcare

The application of OR methods in healthcare has been widely reported in literature (Brailsford et al., 2009; Jun et al., 1999; Katsaliaki and Mustafee, 2011; Mustafee et al., 2010, 2013; Mustafee and Katsaliaki, 2015; Rais and Viana, 2011). Several studies have used these methods for healthcare emergency planning. The book presents a collection of studies that have applied analytical methods for achieving heightened preparedness, better planning and faster response to A&E and public health emergencies. The book covers four broad themes that are relevant to healthcare emergency planning. These are A&E, ancillary services, outbreak of epidemic and public health emergency response. A&E is core to emergency planning, and the availability of adequate resources is crucial to saving lives. The second theme is that of ancillary services like ambulance service and healthcare supply chain. They support A&E and the wider healthcare operations. Planning for epidemics (the third theme) enables public health institutions to minimize the effect of outbreaks among the population. The final theme is on public health emergency response and it acknowledges the need for multi-agency planning using a qualitative approach to problem-solving. These themes intersect through the various parts and chapters of the book. The structure of Volumes 1 and 2 is discussed next.

1.3.1 Volume 1: OR for emergency planning at an operational level

Volume 1 consists of 13 studies which are on operational level planning in the context of emergency preparedness and response. These chapters are presented in three well-defined parts of the book, namely, *OR for Locating Emergency Services* (Part I; Chapters 2–4), *OR for Operational Planning in Emergency Services* (Part II; Chapters 5–10) and *OR for Inventory Management in Emergency Services* (Part III; Chapters 11–14). The chapters mainly focus on 'Hard' OR techniques. Table 1.1 presents an outline of Volume 1 using the triple lens of technique-domain-context.

1.3.2 Volume 2: OR for emergency planning at a strategic level

Volume 2 of the book is mainly on strategic level planning made possible through the application of OR. It consists of 13 studies that are organized

Table 1.1 An outline of volume 1 presented through the triple lens of technique-domain-context

Volume 1 Chapter Number	The Triple Lens of Technique-Domain-Context		
	OR Technique	**OR Application Area (General / Specific)**	**OR Application Context**
Chapter 2	Location Models; Queuing Theory	Emergency Services / **Emergency Centres**	Location of emergency centres
Chapter 3	Mathematical Modelling; Simulation	Emergency Services / **Ambulance Logistics**	Ambulance dispatch and relocation
Chapter 4	Mathematical Modelling; Simulation	Emergency Services / **First responder**	Routing of emergency responders
Chapter 5	Balance Scorecard; Multi-Criteria Decision Analysis (MCDA); Discrete-event Simulation	Emergency Services / **A&E**	Improving planning and efficiency of healthcare processes using a framework-based approach
Chapter 6	Data Mining; Discrete-event Simulation	Emergency Services / **A&E**	Identification of bottlenecks in emergency departments
Chapter 7	Discrete-event Simulation	Emergency Services / **A&E**	Improving processes to reduce time to first consult and turnaround time of patients in A&E
Chapter 8	Discrete-event Simulation	Emergency Services / **Integrated Emergency Post (A&E + GP)**	Delivery of acute care outside office hours
Chapter 9	Discrete-event Simulation	Public Health/ **Control of Epidemic**	Workforce management
Chapter 10	Mathematical Modelling; Maximum Likelihood Estimation (MLE)	Public Health/ **Control of Epidemic**	Modelling treatment effects of HIV/AIDS; investigation of possible future scenarios
Chapter 11	Dynamic Programming; Simulation	Public Health/ **Control of Epidemic**	Inventory management for pharmaceutical products
Chapter 12	Mathematical Modelling; Economic Manufacturing Quantity (EMQ)	Public Health/ **Emergency Response**	Inventory management for pharmaceutical products
Chapter 13	Bayesian Decision Model	Public Health/ **Emergency Response**	Inventory management for emergency supplies
Chapter 14	Discrete-event Simulation	**Healthcare Supply Chain**	Supply chain management related to blood; inventory management for blood

under *OR for Assessment and Review of Emergency Services* (Part IV; three chapters), *OR for Policy Formulation in Emergency Services* (Part V, two chapters), *OR for Broader Engagement in Planning for Emergency Services* (Part VI, four chapters) and *Application of OR within the Wider Healthcare Context* (Part VII, four chapters). The chapters included in Volume 2 are outlined in Table 1.2; for every chapter (except for Part VII which includes chapters on literature review) the table lists the OR technique, domain of application and context of use.

1.4 Overview of volume 1: a focus on operational level planning

Following this introductory chapter, the book consists of 13 chapters that are organized into four parts based mainly on the context of application. Part I focuses on the use of ***OR for Locating Emergency Services***. It focuses on location of resources, movement of resources from one place to the other (resource relocation) and resource routing; the specific studies relate to planning for emergency centres, dispatch of ambulances and first responder service. In **Chapter 2**, Silva and Serra present work on prioritization of calls relating to emergency services. Their objective is to locate emergency centres and to allocate users to these centres in order to maximize the covered population while also considering the varying waiting time limit associated with the different priorities of healthcare services. The authors approach this problem by using location models that incorporate queuing behaviour (queuing maximum coverage location allocation model) and connect this to priority queuing theory. Results from their study show that strategically located emergency centres can avoid congestion of higher priority services. **Chapter 3** focuses on ambulance relocation; the objective is to better manage the demand for ambulances, which varies with time, by planning relocation of the existing fleet. Authors Andersson and Värbrand present new algorithms for the ambulance dispatch and the dynamic ambulance relocation problems. Computational tests using a simulation model show that the algorithms are beneficial in reducing the waiting periods for the patients. **Chapter 4** presents the work of Henchey et al. which compares the performance of conventional shortest path model and a variance-constrained shortest path model for emergency responders. Comparisons are made under varying conditions of traffic and weather in a 'smart' environment that simulates an intelligent transportation system with real-time traffic feeds. The routing methodologies and data captured techniques that are presented in this chapter

Table 1.2 An outline of volume 2 presented through the triple lens of technique-domain-context

Volume 2 Chapter Number	The Triple Lens of Technique-Domain-Context		
	OR Technique	**OR Application Area (General / Specific)**	**OR Application Context**
Chapter 2	System Dynamics; Whole-systems Approach	Emergency Services / **Emergency and On-demand Health**	Whole-system review
Chapter 3	Process Mapping; Discrete-event Simulation	Emergency Services / **A&E Department**	Assessing the impact of systems modeling in care redesign
Chapter 4	Monte Carlo Simulation	Emergency Services / **First Responder**	Assessment of cardiac first responder schemes
Chapter 5	Discrete-event Simulation	Emergency Services / **A&E Department**	Informing national policy development in A&E
Chapter 6	System Dynamics	Emergency Services / **A&E Department**	Policy formulation in A&E
Chapter 7	Qualitative System Dynamics	Emergency Services / **Acute Hospital**	Mapping of acute patient flows within the NHS
Chapter 8	Critical Systems Thinking; Soft Systems Methodology	Public Health/ **Emergency Response**	Planning for a multi-agency counselling service that could be activated in the event of an emergency
Chapter 9	Network-based Simulation	Public Health/ **Control of Epidemic**	Simulations for epidemiology as a pedagogical tool for public health education
Chapter 10	Knowledge Management	Emergency Services / **Hospital and Ambulance Service**	Importance of taking a systems view of knowledge management
Chapter 11	Literature review in the area of healthcare OR and modelling & simulation		
Chapter 12	Literature Review in the area of healthcare modelling & simulation		
Chapter 13	Literature Review focussing on System Dynamics applications to European healthcare issues		
Chapter 14	Viewpoint on the contribution of OR in health from the context of UK NHS		

can help planning for the better and will consequently decrease the response time of the first responder service.

Part II of the book is on ***OR for Operational Planning in Emergency Services*** and it consists of six chapters on planning of A&E departments, integrated emergency posts (combining GP surgeries with A&E) and public health immunization clinics. **Chapter 5** is set in the context of building a more sustainable and efficient healthcare system that achieves the defined strategic as well as tactical and operational objectives. Authors Abo-Hamad and Arisha examine how operations management practices can be translated to support decision-making in healthcare. They present an integrated framework that uses simulation modeling, Balanced Scorecard and Multi-Criteria Decision Analysis (MCDA) and applies them to a real-world case study of an A&E department in a hospital based in Dublin. In **Chapter 6** authors Ceglowski, Churilov and Wasserthiel make the case for the combined application of Data Mining and Discrete-event Simulation (DES) for modeling emergency departments. This is an example of hybrid simulation (Powell and Mustafee, 2014) in which individual techniques are leveraged and the combination of techniques allow for better representation of the system of enquiry. The authors use data-driven patient grouping and incorporate them into a DES, thereby providing insights into the complex relationship between, for example, patient urgency and queues for treatment. **Chapter 7** presents a DES study by Choon et al. that was conducted in a hospital in Singapore. The DES modelled the various processes in an emergency department, for example, registration, triage, consultation and x-ray, and was used to evaluate the impact of operational strategies on performance measures related to, (a) the *time to first consult*, and (b) *turnaround time* of patients in A&E. The results of the study enabled the management team to gain practical insights into the relative cost-effectiveness of various process improvement strategies and their consequent impact on the aforementioned performance measures. **Chapter 8** by Borgman et al. focuses on a problem that is increasingly being experienced in the UK NHS: the overcrowding of emergency departments and consequently the significant number of breaches in the government's four-hour waiting target for A&E. The study is based in the Netherlands where one source of overcrowding is attributed to an increasing number of patients who choose to visit emergency departments for acute care outside office hours. The authors seek to transform emergency care by designing an optimal process which merges individual services to provide a single point of access (Integrated Emergency Post). Towards this they develop a DES, the results of which show a sustainable solution to

emergency department overcrowding, resulting in cost savings for the hospital, increased quality of care for the patients and better usage of resources. **Chapter 9** presents a case study on improving staffing decisions at mass immunization clinics for pandemic influenza. Towards this, authors Beeler, Aleman and Carter use a DES that is based on existing data from Canadian clinics; they show that the marginal benefit of adding staff is greatly underestimated if indirect waiting costs and intra-facility infections are not considered. The aim of the study presented in **Chapter 10** (authors Griffiths, Lawson and Williams) is to model the effect that treatments and therapies have had on the HIV/AIDS epidemic, particularly since the introduction of a combination of antiretroviral drugs. Towards this the authors propose a compartmental model to analyse the effects of various treatment regimes. The model is also used to project levels of incidence and prevalence of HIV/AIDS in the future and to investigate several possible scenarios, for example, extrapolated prevalence in sub-populations if prevention and awareness campaigns were introduced successfully.

Part III of the book comprises of four chapters on the application of ***OR for Inventory Management in Emergency Services***. The first two chapters focus on the stockpiles of medical supplies for emergency public health response while the third chapter is on the efficient management of the blood supply chain and includes inventory management. In **Chapter 11**, Vila-Parrish, Ivy and He present an original study of a hospital pharmacy inventory system where medication demand is driven by an influenza outbreak. They formulate a Dynamic Programming (DP) model related to disease outbreak that uses historic influenza data, and use it as an input to the inventory simulation model. Alternate policies are compared through use of the simulation; results provide insights on the management of perishable medication inventory during an outbreak of an infectious. In **Chapter 12**, authors Shen, Dessouky and Ordonez present a study which extends the traditional Economic Manufacturing Quantity (EMQ) lot sizing model for perishable inventory with a minimum volume constraint, which is applicable to managing the inventory of medicines for the Strategic National Stockpile (SNS). SNS is a reserve of medical supplies that is maintained by the federal government to protect the American public in case of a public health emergency. The authors perform numerical experiments and sensitivity analysis to illustrate how the government can obtain lower costs or a larger stockpile at the same cost by allowing more freedom in the management of the stocks. **Chapter 13** is also on inventory management. In this chapter authors Taskin and Lodree focus

on inventory management problem that is encountered by manufacturers who prepare emergency supplies for humanitarian relief efforts in response to an observed hurricane. The specific application addressed is a single-supplier, multi-retailer supply chain system in which demand at each retailer location is affected by the trajectory of an observed hurricane. A sequential Bayesian decision model that incorporates information updates associated with the National Hurricane Center's (a division of the United States' National Weather Service) wind speed probability forecasts is introduced to address the problem from the manufacturer's perspective. The effectiveness of the proposed approach is explored based on an existing framework for evaluating the expected costs and benefits associated with proactive preparation for potential disaster relief activities within the context of inventory control. Authors Katsaliaki and Brailsford are the authors of the final chapter of Volume 1 of this book. **Chapter 14** is on blood supply chain. Their case study is concerned with analysing policies for managing the blood inventory system in a typical UK hospital that is supplied by a regional blood centre. The objective of the study is to improve procedures and outcomes by modeling the entire supply chain, to include donors, regional blood centres (UK National Blood Service), hospitals, doctors and recipients of blood. They use DES to determine ordering policies that may lead to reductions in shortages and wastage of blood. A distributed version of the blood supply chain model has also been developed (Katsaliaki et al., 2009; Mustafee et al., 2009).

1.5 Chapter summary

This book (Volume 1) provides an exploration of quantitative techniques that are frequently used for decision-making in healthcare. The chapters have covered OR techniques such as, Mathematical Modelling and Optimization, Multiple-Criteria Decision Analysis, Discrete-event Simulation and Bayesian Decision Models. These techniques have been applied for better management of entities engaged in emergency care, for example, first responders, ambulance services, A&E departments and mass immunization centres.

The overarching theme of Volume 2 is on the use of OR for emergency planning at a strategic level. It consists of a further 13 chapters on the application of OR for policy formulation, assessment and review of emergency services, OR approaches that facilitate broader engagement in planning for emergency services, and chapters that focus on the use of OR within the wider healthcare context. It includes studies

that have applied quantitative as well as qualitative OR techniques. It is expected that both the volumes will serve as important reference source for policymakers, NHS trusts, managers, clinicians, researchers and OR practitioners working in this area.

References

Brailsford, S. C., Harper, P. R., Patel, B. and Pitt, M. (2009). 'An Analysis of the Academic Literature on Simulation and Modelling in Healthcare'. *Journal of Simulation,* 3(3): 130–140.

Jun, J. B., Jacobson, S. H. and Swisher, J. R. (1999). 'Application of Discrete-Event Simulation in Health Care Clinics: A Survey'. *Journal of the Operational Research Society,* 50(2): 109–123.

Katsaliaki, K. and Mustafee, N. (2011). 'Applications of Simulation Research within the Healthcare Context'. *Journal of the Operational Research Society*, 62(8): 1431–1451.

Katsaliaki, K., Mustafee, N., Taylor, S. J. E. and Brailsford, S. (2009). 'Comparing Conventional and Distributed Approaches to Simulation in Complex Supply-Chain Health Systems'. *Journal of the Operational Research Society*, 60(1): 43–51.

Miser, H. J. (2000). 'The Easy Chair: What OR/MS Workers Should Know about the Early Formative Years of their Profession'. *Interfaces*, 30: 99–111.

Mustafee, N. and Katsaliaki, K. (n.d.). *A Keyword Classification Scheme for OR/MS*. Submitted.

Mustafee, N. and Katsaliaki, K. (2015). 'Simulation for Sustainable Healthcare' – Guest Editorial for the Special Issue on Modelling and Simulation for Sustainable Healthcare. *Journal of Simulation*, 9(2): 83–85. Palgrave Macmillan.

Mustafee, N., Katsaliaki, K. and Taylor, S. J. E. (2010). 'Profiling Literature in Healthcare Simulation'. *SIMULATION: Transactions of the Society of Modelling and Simulation International*, 86(8–9): 543–558.

Mustafee, N., Katsaliaki, K., Williams, M. D. and Gunasekaran, A. (2013). 'Healthcare Operations Management through Use of Simulation: An Introduction to the Special Issue'. *Journal of Enterprise Information Management*, 26(1 and 2): 5–7.

Mustafee, N., Taylor, S. J. E., Katsaliaki, K. and Brailsford, S. (2009). 'Facilitating the Analysis of a UK National Blood Service Supply Chain Using Distributed Simulation'. *SIMULATION: Transactions of the Society of Modelling and Simulation International*, 85(2): 113–128.

Powell, J. and Mustafee, N. (2014). 'Soft OR Approaches in Problem Formulation Stage of a Hybrid M&S Study'. In *Proceedings of the 2014 Winter Simulation Conference*, 7–10 December 2014, Savannah, GA, pp. 1664–1675. IEEE Press Piscataway, NJ, USA.

Rais, A. and Viana, A. (2011). 'Operations Research in Healthcare: A Survey'. *International Transactions in Operational Research*, 18(1): 1–31.

Part I
OR for Locating Emergency Services

2

Locating Emergency Services with Different Priorities: The Priority Queuing Covering Location Problem

F. Silva[1,2] *and D. Serra*[1]
[1]*Universitat Pompeu Fabra, Barcelona, Spain; and* [2]*Universidade dos Açores, Ponta Delgada, Portugal*

Previous covering models for emergency service consider all the calls to be of the same importance and impose the same waiting time constraints independently of the service's priority. This type of constraint is clearly inappropriate in many contexts. For example, in urban medical emergency services, calls that involve danger to human life deserve higher priority over calls for more routine incidents. A realistic model in such a context should allow prioritizing the calls for service. In this paper, a covering model which considers different priority levels is formulated and solved. The model heritages its formulation from previous research on Maximum Coverage Models and incorporates results from Queuing Theory, in particular Priority Queuing. The additional complexity incorporated in the model justifies the use of a heuristic procedure.

2.1 Introduction

Questions related to emergency services have been studied by researchers over the last 25 years. They refer to medical systems, police operations, firefighting systems, emergency repair systems, and others. Researchers agree that the outcome of service is in great part defined by the time that the customer waits.

Emergency services planners must solve the strategic problem of where to locate emergency service centres and the tactical problem of allocating

Reprinted from *Journal of the Operational Research Society*, 59(9): 1229–1238, 2008, 'Locating Emergency Services with Different Priorities: The Priority Queuing Covering Location Problem', by F. Silva and D. Serra. With kind permission from Operational Research Society Ltd. All rights reserved.

demand to those centres. The performance of an emergency centre may be judged by the number of persons in queue or by the length of time that a person must wait after he or she arrives at the centre. These indicators are strongly correlated with the number of centres available and with their locations. Not all cases have the same dependence on time: rush jobs are taken ahead of other jobs, and important customers may be given precedence over others. This is clearly the case in hospital emergency rooms, where patients are roughly divided into three categories: critical cases, where prompt treatment is vital for survival; serious cases; and stable cases, where treatment can be delayed without adverse medical consequences.

Reducing waiting lists is a struggle for many healthcare organizations, especially given that the cost of the resources demands a high utilization rate. A lengthy patient wait in the healthcare industry shows more adverse consequences than in most other services. This generates stress and dissatisfaction, increases the cost of medical care, and can constitute a barrier to effective treatment. An optimized location for all facilities and equal allocation of patients to those facilities is a vital factor in improving time performance. This paper addresses emergency healthcare management, but the research can easily be expanded to other areas such as distribution centres or repair systems.

Marianov and Serra (1998) introduce the queuing maximal covering location allocation model, which locates p centres and allocates users to these centres in order to maximize the covered population. Coverage is defined as (i) when patients are allocated to a centre within a standard time or distance from home location, and (ii) when patients are served within time τ of arrival at the centre, with a probability of at least α. This paper presents a natural expansion of this model that considers different time standards for different priorities of healthcare services.

Decisions that do not account for different priorities may lead to less effective locations. This paper considers different priority levels. The formulation of this model derives from previous research on maximum coverage models and incorporates results from queuing theory, in particular priority queuing. The additional complexity incorporated in the model justifies the use of a heuristic procedure. We develop an evaluation based on numerical examples.

2.2 Related literature

Location models that incorporate queuing effects appear in the literature of the early 1980s. Berman *et al* (1985) present a work that is considered

as the 'beginning in a potentially fecund marriage between location and queuing theories.' They extend Hakimi's (1964) one-median problem by embedding it in a general queuing context. The formulation explicitly includes dependence on service times, travel times, and queuing delays on the location of the service facility. Their work was in part motivated by the pioneer hypercube queuing model developed by Larson (1974).

Batta *et al* (1988) note that queuing disciplines frequently used in decision models (such as first-come first-served, last-come first-served, and service in random order) are clearly inappropriate in many contexts. They point to urban emergency services and police patrols as examples of when the risk of life or the violence of the crime will factor into the service order. A formulation for the single server queuing location model is provided (*k* priority queuing location) as well as some solution techniques that allow calls to be selected from an arbitrary number of priority classes. The optimal K-PQL model location usually differs from the one obtained by grouping calls from all priorities into a single category and using the single queue length model.

Batta (1989) considers the problem of locating a single server on a network operating as an M/G/1 queue, in which queued calls are serviced by a class of queuing disciplines that depend solely on expected service time information. The model is analysed as an M/G/1 non-preemptive priority queuing model, with location-dependent priorities. Motivated by the problem of locating fire trucks in a geographical area, which requires multiple trucks to be located within an acceptable distance standard to achieve coverage, Batta and Mannur (1990) examine the set covering problem and the maximal covering location problem in the context of multiple units being required by some demands.

For congested service systems, Brandeau and Chiu (1992) present the stochastic queue centre location model, which seeks to minimize the maximum expected response time to any customer. Expected response time comprises expected waiting time until the server becomes free and expected travel time. A more recent work in the same line of research was developed by Jamil *et al* (1999). The stochastic queue centre problem considers the objective of locating a single facility operating as an M/G/1 queue in steady state so as to minimize a weighted linear combination of the square of the average response time and the variance of the response time. Berman and Vasudeva (2000) also consider the problem of locating a general number of service units, which return to their home locations only if no calls are waiting for service.

Branas and Revelle (2001) developed the trauma resource allocation model for ambulances and hospitals as a guide for healthcare planners.

The model combines a mixed-integer linear program with a new heuristic and considers two resource, trauma centres and aeromedical depots, in a two-level hierarchy. The objective is to maximize coverage, which is defined as when at least one trauma centre is sited within a ground standard time or when an aeromedical depot/trauma centre pair is sited in such a way that the sum of the flying time from the aeromedical depot to *i* and the flying time from *i* to the trauma centre is within the same time standard.

Ball and Lin (1993) propose a reliability model for emergency service vehicle location. Based on the probability of system failure, they derive a 0–1 integer programming optimization model that is solved using a branch-and-bound procedure. The computational results show that the processing technique is highly effective. Mandell (1998) formulates a covering type model for two-tiered emergency medical services that maximizes the expected number of calls for service and takes server availability into account through a two-dimensional queuing model. Considering a redeployment problem for a fleet of ambulances, Gendreau *et al* (2001) propose a dynamic model and a parallel tabu search heuristic.

Harewood (2002) offers a multi-objective version of the maximum availability location problem in a real application by solving the problem of emergency ambulance deployment in Barbados. The first objective seeks to maximize the population covered within a given distance standard and with a given level of reliability, while the second objective chooses locations that minimize the cost of covering the population. Verter and Lapierre (2002) address the problem of locating preventive healthcare facilities. They assume that distance is a major determinant of participation and that people will go to the closest facility for preventive care. Goldberg (2004) offers a review of the development and current state of operations research for deployment and planning analysis pertaining to emergency services and fire departments.

2.3 Formulation

The formulation of the model presented here closely follows the methodology proposed by Marianov and Revelle (1994). The authors relax the assumption of dependence of server availability and model the behaviour in each region as an M/M/s-loss queuing system obtaining a probabilistic formulation of the location set covering problem. The reliability constraints are formally incorporated using queuing theory to model the arrival–departure process within the location model itself. The authors refer to fixed facilities, in contrast with earlier research that

considers a mobile server. Marianov and Serra (1998) present several probabilistic, maximal covering, and location–allocation models with constrained waiting times for queue length. In this section, we propose a model that connects the queuing maximal covering location allocation model with priority queuing theory. We describe the results from priority queuing that will be included in the model and explain each of the equations in this new model.

2.3.1 Results from queuing theory

Priority queuing has been analysed in many research works. In this paper, we follow Kleinrock's (1975) textbook notation. In a priority queuing system, we assume that an arriving customer belongs to priority class k ($k = 1, 2, ..., K$). The smaller the priority index, the higher the priority of the class. Let us consider non-preemptive priorities, that is when a customer in the process is not liable to be ejected from service and returned to the queue when a higher priority customer appears. Customers from priority k arrive in a Poisson stream at rate $\lambda^{[k]}$ for time unit. Each customer from this group has his or her service time S selected independently from the distribution $B_k(S)$ with mean $\overline{S}^{[k]}$. Let us also consider a Head of Line discipline within each priority level. The average waiting time for priority k services is given by the following expression:

$$\begin{aligned} W^{[k]} &= \frac{W_0}{(1-\sigma_k)(1-\sigma_{k-1})} \quad \textit{if } 1-\sigma_k > 0 \\ &= +\infty \quad \text{otherwise} \end{aligned} \tag{1}$$

where

$$\sigma_k = \sum_{i=1}^{k} \rho^{[i]} \quad \text{with } \sigma_0 \equiv 0$$

and

$$\rho^{[i]} = \frac{\lambda^{[i]}}{\mu^{[i]}} = S^{[i]}\lambda^{[i]}$$

$\mu^{[i]}$ is the service rate for priority class i.

The interpretation of ρ here is the fraction of time that the server is busy (as long as $\rho < 1$).

W_0 corresponds to the average delay that the tagged customer experiences due to another customer found in service and can be computed using the following formula:

$$W_0 = \sum_{i=1}^{R} \frac{\lambda^{[i]} \overline{(S^{[i]})^2}}{2}$$

where $\overline{(S^{[i]})^2}$ is the second moment of service time distribution.

2.3.2 Problem definition

The structure of the problem comprises a discrete space in which customers, representing demand, are positioned and in which facilities are to be positioned by a decision maker. Customers are assumed to be located at the nodes of the network (demand nodes). We assume that the service call process for each node *i* is a Poisson process with rate λ_i. The problem will locate a given number of facilities providing the service. We assume that a discrete set of potential facility locations eligible to provide coverage to the demand nodes has been identified.

For the purposes of this model, utilization factors are defined by the product of the average service time and the arrival rate as defined by the summation of the call rates of all demand points allocated to a centre at *j*

$$\rho_j^{[k]} = \overline{S_j^{[k]}} \lambda_j^{[k]} = \overline{S_j^{[k]}} \sum_i f_i^{[k]} X_{ij}^{[k]}$$

where $X_{ij}^{[k]}$ is a zero one variable defining allocation of demand point *i* to a centre at *j* for priority *k* services. $f_i^{[k]}$ is the call rate for services of priority class *k* at demand point *i*. For example, $f_i^{[k]} = 0.08$ indicates that at demand point *i*, the number of calls per unit of time for services of priority *k* equals 8% of the total population on average.

2.3.3 Priority queuing covering location problem

The priority queuing covering location problem (PQCLP) assumes static assignments of customers to service facilities. This is typical in the case of fixed server locations, where customers travel to facilities to obtain service (see Berman and Krass, 2002). We offer two versions of the PQCLP: the first assumes a directed choice environment in which a central authority dictates the assignment of a customer (patient) to a centre, while the second assumes a user choice environment in which

the assignment is made by each individual customer. In our case, they will always go to the closest centre.

2.3.3.1 Directed choice environment

This version of the PQCLP will define separate allocations for the different priorities that may or may not coincide, that is a demand point may be allocated to a centre at j for one priority and to centre $k \neq j$ for other priorities. The objective is to maximize the population covered in all priorities. Different time constraints are imposed for different priorities. The model's formulation is as follows:

$$\text{Max} \quad Z = \sum_k \sum_i \sum_j a_i X_{ij}^{[k]}$$

s.t.

$$X_{ij}^{[k]} \leq Y_j \quad \forall i \in I, \forall j \in N_i, \forall k \tag{2}$$

$$\sum_{j \in N_i} X_{ij}^{[k]} \leq 1 \quad \forall i \in I, \forall k \tag{3}$$

$$\sum_j Y_j = p \tag{4}$$

$$W_j^{[k]} \leqslant \tau^{[k]} \quad \forall j, \forall k \tag{5}$$

$$X_{ij}^{[k]}; Y_j \in \{0,1\} \quad \forall i \in I, \forall j \in N_i \tag{6}$$

where $X_{ij}^{[k]} = 1$ if demand point i is allocated to a centre at j for priority k's urgencies and zero in all other cases. $Y_j = 1$ if a centre is located at j and zero in all other cases. $W_j^{[k]}$ is the average waiting time for priority class k at centre j (see Equation 1). Parameter $\tau^{[k]}$ represents the limits imposed on the waiting times for priority k; p is the number of centres to be sited; a_i is the population at demand point i; I is the set of all demand points; and N_i is the set of centres located at a distance smaller or equal to d from demand point i.

Constraint (2) states that if population i is allocated to a centre at j for priority k, then there must be a centre located at j; (3) forces each demand point to be allocated to only one centre; (4) defines the number of centres to be located; and (5) ensures that the average waiting time

is less than the given standard. Additionally, j must be within set N_i so that a demand point requires a centre located within distance d in order to be covered.

2.3.3.2 User choice environment

The above model can be converted to a user choice model by assuming that customers will always choose the closest centre and by adding the following group of constraints to enforce this assumption:

$$X_{ij}^{[k]} \geq Y_j - \sum_{l \in C_{ij}} Y_l$$

$$C_{ij} = \{l \mid d_{il} < d_{ij}\} \quad \forall i \in I, \forall j \in N_i \tag{7}$$

Originally introduced by Rojeski and ReVelle (1970) in the context of the budget constrained median problem, this equation establishes that if j is an open centre and no closer centre is open, then demand i must be assigned to j. If j is open but a closer centre is also open, then this relation does not constrain assignment in any way (for a more detailed discussion of closest assignment constraints, see Gerrard and Church, 1996).

2.4 Heuristic solution procedure

A greedy randomized adaptive search procedure (GRASP) is a metaheuristic that has been applied to a wide range of operations research and industrial optimization problems. These include problems of scheduling, routing, logic, partitioning, location, layout, graph theory, assignment, manufacturing, transportation, telecommunications, automatic drawing, electrical power systems, and VLSI design (see Resende, 1998). Festa and Resende (2001) present an extensive annotated bibliography of GRASP literature. The following notations are used in this section:

- j index for potential locations
- i index for demand nodes
- k index for priority level
- D_j list of potential facility location ordered in accordance with total population
- S solution set
- $\overline{S}$ complement of the solution set
- C set of candidate sites

p	number of facilities to locate
n	number of demand nodes
inc_k_j	incoming call rate at potential facility location j
inc_j	total incoming call rate at potential facility location j
ρ_k_j	utilization factor at potential facility location j for priority level k
D_{ij}	list of demand points within the standard distance from potential facility location j
R	total priority levels
w_k_j	waiting time for priority level k at potential facility location j
f_k_i	frequency of demand for priority k services at demand node i
GammaSet	set containing all possible values for γ
$prob_\gamma$	probability of choosing a specific value for γ
$ut(\gamma)$	utility value of a specific γ value

GRASP is an interactive process with a feasible solution constructed at each independent iteration. One GRASP iteration consists of two phases: construction phase and local search phase. Figure 2.1 describes a pseudocode for the GRASP algorithm.

The procedure for the construction phase, which will return an initial solution at each of the iterations, is denoted as *Greedy_Randomized_Construction*(*Seed,* γ) and is a function of the seed in the random numbers generation and of a parameter gamma defining which solutions are included in the restricted candidate list (RCL), which contains the best greedy solutions. The development of this procedure appears in Figure 2.2.

The algorithm starts by sorting the candidate nodes in accordance with their populations. In our example, every demand point is considered as a potential facility location. Another possibility would be to consider only a subset of the demand points in the list D_j. Beginning

```
procedure GRASP (Max_iterations, Seed)

1       For k=1 to Max_iterations do
2               S ← Greedy_Randomized_Construction(Seed, γ);
3               S ← Local_Search(Solution);
4               Update_Solution(Solution, Best_Solution);
5       enddo
6       end GRASP
```

Figure 2.1 GRASP pseudocode

procedure *Greedy Randomized Construction* (Seed,γ)

{sort candidate sites by decreasing order of population}
1 $D_j \leftarrow$ *Sort_Candidate_Sites (population);*
{initialize solution set}
2 $S := \{\}$;
3 $\overline{S} := C$;
{while solution is not a complete solution}
4 **while** $|S| \neq p$ **do**
{loop over all candidate sites not in the solution list}
5 **For** j=1 to $|\overline{S}|$ **do**
{initialize parameters}
6 $inc_k_j := 0$; $inc_j := 0$; $W_k_j := 0$; $\rho_k_j := 0$;
{restrict demand points list to the standard covering distance to site j}
7 $D_{ij} \leftarrow \{i \in D, d_{ij} \leq d\}$
{sort demand points by increasing distance to site j}
8 $D_{ij} \leftarrow$ *Sort_Demand_Points(distance);*
{loop over priorities}
9 **For** k=1 to R **do**
{loop over demand points in set Dij}
10 **For** i=1 to $|D_{ij}|$ **do**
{sum frequencies at each demand point if waiting time limit is not reached}
11 **If** ($W_k_j < \tau^{[k]}$ and $\rho_k_j<1$)**do**
12 $inc_k_j := inc_k_j + f_k_i$;
13 *actualize w_k_j;*
14 *actualize* ρ_k_j;
15 **endif**
16 **enddo**
{compute incoming call rate at the potential center location}
17 $inc_j := inc_j + inc_k_j$;
18 **enddo**
{construct the restricted candidate list}
19 $c^{max} := max\{inc_j\}$;
20 $RCL \leftarrow \{j \in \overline{S}, inc_j \geq \gamma\, c^{max}\}$;
{select randomly one site from the RCL}
21 $j^* \leftarrow$ *Random_Select(RCL);*
22 $S := S \cup \{j^*\}$;
23 $\overline{S} := \overline{S} \setminus \{j^*\}$;
{take the demand points allocated to j* out of the demand point's list}
24 **For** i=1 to $|D_{ij^*}|$ **do**
25 $D := D \setminus \{i \in D_{ij^*}\}$;
26 **enddo**
27 **enddo**
28 **end** *Greedy Randomized Construction*

Figure 2.2 Construction phase pseudocode

with the first candidate on the list, we allocate the closest demand points to each node until the coverage limit is reached. This limit can be reached by the utilization coefficient or by the limit imposed on the waiting time at each priority level.

The total demand assigned to each of the potential locations *j* is notated as the incoming call rate. Since there are different waiting time limits corresponding to the different priority levels, there are also different incoming call rates. The total incoming call rate is defined as the sum of the incoming call rates for all priorities. The incoming call rates are the greedy function of the algorithm and can be defined as the weight of as yet uncovered demand points that would become covered if centre location *j* were to be chosen.

We include in the RCL the candidate nodes with total incoming call rates greater or equal than gamma percent of the incoming call rate corresponding to the potential facility location with the higher value. In the GRASP algorithm, parameter gamma is defined *a priori* (eg for a gamma value equal to 0.8, we include all of the potential locations with total incoming call rate greater than 80% of the highest incoming call rate). At each of the iterations, we randomly choose the facility locations within the candidate sites with the highest incoming call rates. Note that in a pure greedy heuristic, like the one suggested by Marianov and Serra (1998), the choice would always be to locate a centre in the node with the highest sum of incoming call rates ($\gamma = 1$).

In the local search phase, we de-allocate the demands that were allocated to each centre and move the centre to all possible unused potential facility sites, repeating steps 9–20 of the greedy randomized construction procedure. If some locations give a better objective, the centre should remain at that location; otherwise, the facility should return to its initial location (see Figure 2.3). This procedure continues until no further improvements can be made.

To avoid fixing parameter γ in an arbitrary manner, we have implemented a reactive procedure in which γ is updated at each one of the iterations. Reactive GRASP, proposed by Prais and Ribeiro (2000), is a procedure in which the parameter is self-adjusted according to the quality of the solutions previously found. Instead of fixing the value of γ to determine which elements will be placed in the RCL, R-GRASP randomly selects this parameter value from a discrete set $\{\gamma_1, \ldots, \gamma_m\}$. The probability distribution used in the γ selection will be updated after the execution of each block of iterations considering the quality of the solutions obtained by each of the γ_i (the utility of the γ_i). Figure 2.4 describes the R-GRASP algorithm.

```
procedure Local_Search (Solution, Best_Solution)

1       obj_best := obj(S);
        {loop over sites in the solution}
2       for all j ∈ S do
3               S := S \ {j};
        {loop over sites not in the solution}
4               for all j* ∈ S̄ do
5                       evaluate obj (S ∪ {j*});
6                       if obj_best < obj (S∪{j*}) do
7                               S : = S ∪ {j*}:
8                               obj_best : = obj (S ∪ {j*});
9                       else
10                              S : = S ∪ {j}:
11                      endif
12              enddo
13      enddo
14      end Local_Search
```

Figure 2.3 Local search phase pseudocode

In a user choice environment, the heuristic algorithm must be adapted in the construction and local search phases to enforce closest assignment. In this new version, a demand point will always be assigned to the closest centre location, which may lead to infeasible solutions such that the limits imposed for the average waiting times will not be verified. The proposed algorithm penalizes the objective when an infeasible solution is obtained. When obtaining a feasible solution, this set of locations is considered as a potential centre location. Otherwise, this set is considered only as an initial solution and not as a potential location penalizing the objective with a large negative value *M*. This corresponds to the following objective procedure in Figure 2.5.

In the local search phase, a demand point is always allocated to the closest potential centre location, and the feasibility of waiting time limits

```
procedure reactive_GRASP
        {initialize gammas' probabilities}
1       For all γ∈ gammaSet do
2               prob_γ ← 1/|gammaSet|, ∀γ ∈ gammaSet;
3       enddo
4       For block=1 to nblocks do
        {implement GRASP algorithm}
5               For iter=1 to niter do
6                       γ ← Random_Select(gammaSet, prob);
7                       S ← Greedy_Randomized_Construction(γ);
8                       S ← Local_Search(Solution);
9                       Update_Solution(Solution,Best_Solution)
10              enddo
        {actualize gammas' probabilities after each block of
        iterations}
11              prob_γ ← ut(γ)/(Σ_{allγ} ut(γ)), ∀γ ∈ gammaSet;
12      enddo
13      end reactive GRASP
```

Figure 2.4 Reactive GRASP pseudocode

is checked. If the solution is not feasible, then the objective is penalized with *M*. When new locations result in a better objective, the centre is placed at that location; otherwise, it remains at the initial location. This procedure is repeated until no further improvements can be made.

2.5 Computational experience

2.5.1 Evaluation of heuristic

2.5.1.1 Deviation from the optimal solution

To observe the difference between the results from the heuristic and the optimal solution, a simple experiment is implemented that consists of randomly generating problem instances and comparing results from the heuristic with those from complete enumeration. The coordinates of

```
procedure evaluate_objective (S)

1     Allocate each demand point to its closest center location;
2     Evaluate W_k_j and ρ_k_j;
3     obj(S) : =0;
4     If(W_k_j < τ[k] and ρ_k_j<1) do
5             For k=1 to R do
6                     For j=1 to p do
7                             For i=1 to n do
8                                     If (i is allocated to j) do
9                                             obj(S):=obj(S) + f_k_i;
10                                    endif;
11                            enddo;
12                    enddo;
13            enddo;
14    else
15            obj(S):=M;
16    end evaluate_objective;
```

Figure 2.5 Objective evaluation pseudocode

the demand points and the potential facility sites are randomly generated from a uniform distribution between 0 and 6. The populations for each demand point are generated from a uniform distribution between 60 and 800. In all examples, only two priorities are considered. Call rates for first and second priorities are 0.5 and 1%, respectively, of the populations. A waiting time limit of 5 min is imposed for first priority and 20 min for second priority.

The objective values obtained using the first heuristic are compared with those obtained through a complete enumeration procedure in which allocations of demand points to potential facility sites are defined by repeating steps 1–5 of the heuristic algorithm described in Section 5.1. The objective values obtained from the second heuristic are compared with those from a complete enumeration procedure that always allocates a demand point to its closest centre. For Heuristics 1, we consider $\gamma = 1$ (ie the pure greedy algorithm). For Heuristic 2, the pure greedy algorithm does not behave well, and it was necessary to implement GRASP.

In these examples the number of blocks (*nblocks*) was 4, each with a number of iterations (*niter*) equal to 50. The average deviations and maximum deviations are computed for the 50 instances generated. The main results for the different combinations of the number of demand points and the number of centres appear in Table 2.1.

Table 2.1 Computational results—deviations

	Average deviation (% of complete enumeration objective)		Maximum deviation (% of complete enumeration objective)	
	Heuristic 1	**Heuristic 2**	**Heuristic 1**	**Heuristic 2**
20 nodes				
2 centres	0.07	0.29	2.48	11
3 centres	0.41	0.30	5.04	12
4 centres	0.46	1.64	9.95	21
30 nodes				
2 centres	0.03	0.84	1.71	10
3 centres	0.39	0.57	4	14

2.5.1.2 Computing time

To evaluate the heuristic's behaviour, we measure processing times for larger problem instances. Table 2.2 shows the average computing times for simulated instances with 50 and 100 demand points (in a Pentium(r) III processor with 128MB of RAM).

2.5.2 Illustrative example

For simplicity, we consider only two priorities: critical and stable. Additionally, we assume that the first and second moments of service time are the same for both priorities and are independent from the demand's location; we also assume that customers must travel to the centre to obtain service and that service time will not vary among locations. In the computational experience, the 30-node problem instance (Marianov and Serra, 1998) is used. We authors assume that the servers are physicians, that each demand point is also a potential centre location, and that the distances are Euclidean.

Figure 2.6 represents each demand point as a square which area signals the relative population at that demand point. Demand points are numbered from 1 to 30 in decreasing order of population. For our purposes, demand is separated into the daily call rate for first priority cases (0.005 times the population) and the daily call rate for second priority cases (0.01 times the population). The instances are adequate for the problem in analysis. The more populated demand points are grouped in the centre; this eliminates the location of all centres in the nodes with higher frequencies. The instances seem to be quite representative of many urban areas, where most of the population is concentrated in the

Table 2.2 Computational results—average computing times

		Average computing time	
Number of demand points	**Number of centres**	**Heuristic 1**	**Heuristic 2**
50	5	15.44	67.07
	10	24.17	321.81
100	5	143.13	580.95
	10	207.07	1706.92

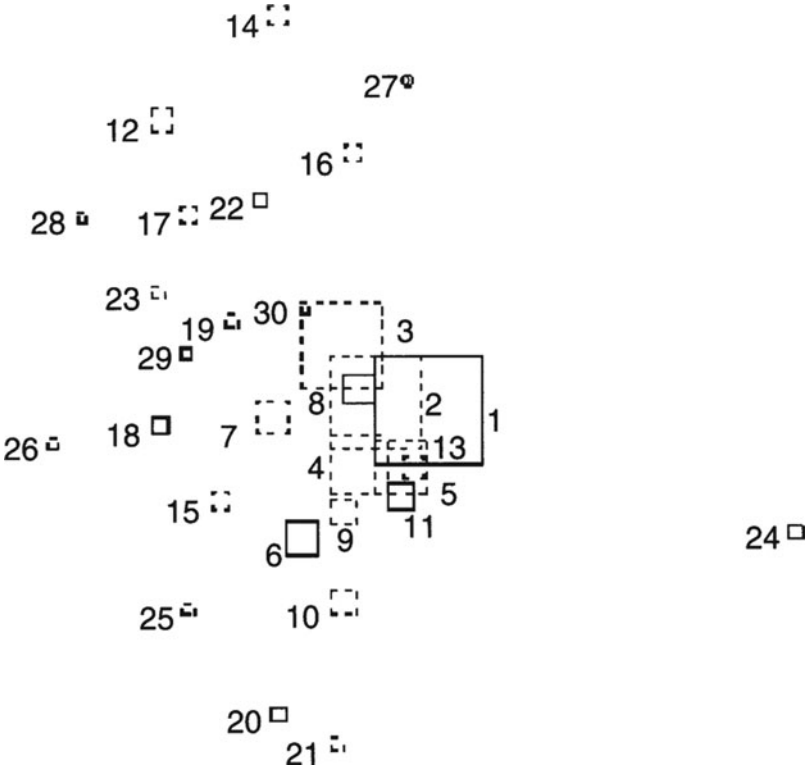


Figure 2.6 Locating the nine centres

centre with a variety of smaller populated areas around the centre. The results of imposing a waiting time limit of 5.5 min on the first priority and 20 min on the second priority appear in Table 2.3.

Table 2.3 Heuristic's results for one server per centre and two priorities

			% pop covered	
Number of centres	**Objective**	**Locations**	**1st priority**	**2nd priority**
10	10940	1; 6; 8; 11; 14; 18; 20; 22; 24; 29	100	100
9	10880	1; 6; 8; 11; 18; 20; 22; 24; 29	98.9	100
8	10810	1; 6; 8; 11; 20; 22; 24; 29	92.1	96.9
7	10650	1; 6; 8; 11; 20; 22; 29	90.7	95.4
6	10350	1; 6; 8; 20; 22; 29	86.3	95.4
5	10160	1; 2; 20; 22; 29	85.2	95.4

Table 2.4 Heuristic's results for one server per centre and one priority

Number of centres	**Objective**	**Locations**	**% pop covered**
10	5470	1; 3; 5; 6; 7; 14; 17; 18; 20; 24	100
9	5470	3; 5; 6; 7; 14; 17; 18; 20; 24	100
8	5400	3; 5; 6; 7; 14; 17; 20; 24	98.7
7	5320	3; 5; 6; 7; 14; 17; 20	97.3
6	5140	3; 5; 6; 7; 17; 20	94.0
5	4950	5; 6; 7; 17; 20	90.0

Table 2.4 shows the results of considering one priority with an average time limit of 12.75 min. The average service time in this example is 10 min. The values on Table 2.4 were computed using the heuristic procedure proposed by Marianov and Serra's (1998). Although nine centres would cover the entire population, considering only one priority, Table 2.3 shows that 1.1% of the population would not receive first priority service within the time limit. This percentage increases drastically with a decrease in the number of centres: with only five centres, 14.8% of the population would not receive first priority service within the time standard. Location decisions will vary considerably depending on whether or not different priorities are taken into account. For example, in the 10-centre case, only six coincide in both scenarios.

Figure 2.6 illustrates the nine-centre case; nodes with a centre are represented by non-dashed squares. Centres are not located in the demand points with the highest population but are strategically distributed.

Considering separate allocation variables for different priorities, it is possible for a demand point to be allocated to more than one centre. At the same time, some centres may receive many calls for some priorities and no calls for others. A centre that offers first priority service will be located at j if (and only if) $\exists i \in N_j : X_{ij}^{[1]} \neq 0$. A centre that offers second priority service will be located at j if (and only if) $\exists i \in N_j : X_{ij}^{[2]} \neq 0$ where $N_j = \{i | d_{ij} \leqslant d\}$.

Figure 2.7 presents the results of the nine-centre case. Only six centres (represented by the double circle) are needed to cover both priorities if three centres (represented by the single circle) that offer service only to

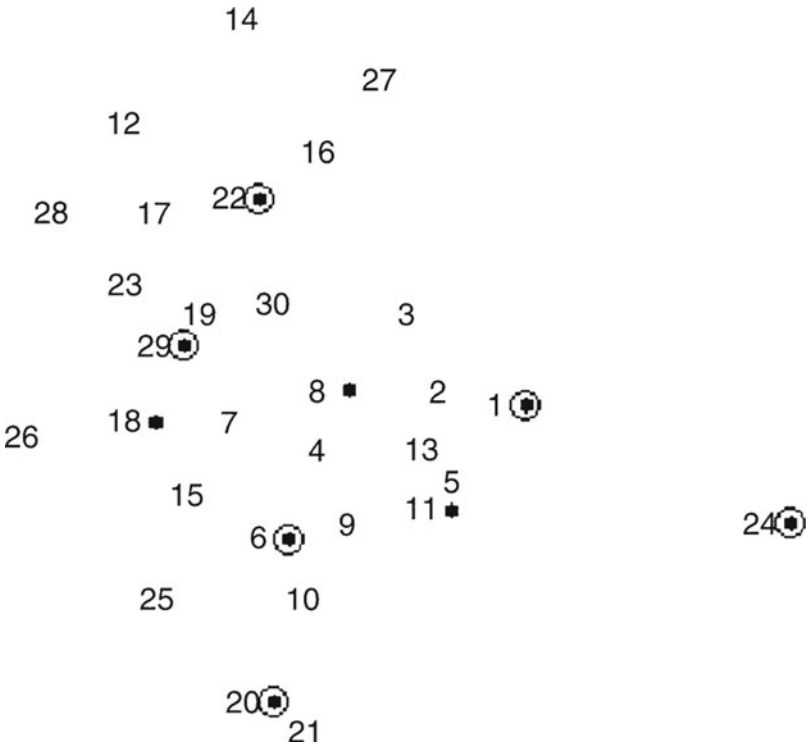


Figure 2.7 Locating the nine centres

Table 2.5 Heuristic's results for one server per centre and two priorities with closest assignments

		% pop covered	
Number of centres	**Objective**	**1st priority**	**2nd priority**
10	10 940	100	100
9	10 940	100	100
8	10 940	100	100
7	10 800	98.7	98.7
6	10 780	98.5	98.5
5	6200	56.6	56.6

first priority cases are located. Note that these centres are located closer to the demand points with the highest population and are needed in order to avoid congestion among first priority services.

In this model, the need for service dictates the number of centres and the type of services offered by each centre. This can lead to important resource savings. Table 2.5 illustrates the results obtained by applying Heuristic 2 to the previous example. The waiting time limits are 5.5 min for first priority cases and 20 min for second priority cases. Demand points are always assigned to the closest facility independently from the priority level; therefore, the percentage covered on first and second priorities coincide. In this new situation, eight centres are needed to cover the entire population.

2.6 Conclusions and future research

When waits are common, utilization rates are high, and resources are expensive, considering more than one priority helps to create a more efficient geographic distribution of service centres. Some centres accumulate more customers of a low priority; emergency centres, when strategically located, can avoid congestion of higher priority services. The PQCLP links population needs with resource allocation. These results should be applied to a directed choice environment in which there is some control over the assignment of demand points to the located centres.

The proposed heuristic procedures provide results that are coherent with the rationality of the model. An extensive evaluation of these procedures shows that the solutions are quite close to those obtained by complete enumeration. The heuristic procedure shows important advantages in terms of computational speed. It allows us to find

solutions for problems that may otherwise require a prohibitive amount of computer time. Priority queuing is a useful theory for many types of services and should be extended to other location-allocation models.

Acknowledgements

This research has been possible thanks to the grant SFRH/BD/2916/2000 from the Ministério da Ciência e da Tecnologia, Fundação para a Ciência e a Tecnologia of the Portuguese government, and grant BEC2000-1027 from the Ministerio de Ciencia y Tecnologia, of the Spanish government.

References

Ball MO and Lin FL (1993). A reliability model applied to emergency service vehicle location. *Opns Res* **41**: 18–36.

Batta R (1989). A queuing-location model with expected service time dependent queuing disciplines. *Eur J Opl Res* **39**: 192–205.

Batta R and Mannur NR (1990). Covering-location models for emergency situations that require multiple response units. *Mngt Sci* **36**: 16–23.

Batta R, Larson RC and Odoni AR (1988). A single-server priority queuing-location model. *Networks* **8**: 87–103.

Berman O and Krass D (2002). Facility location problems with stochastic demands and congestion. *Facility Location: Applications and Theory.* Springer Verlag: Berlin.

Berman O and Vasudeva S (2000). *Approximating performance measures for public services*. Working Paper, Joseph L. Rotman School of Management, University of Toronto.

Berman O, Larson RC and Chiu SS (1985). Optimal server location on a network operating as an M/G/l queue. *Opns Res* **12**: 746–771.

Branas CC and Revelle CS (2001). An iterative switching heuristic to locate hospitals and helicopters. *Socio-Econ Plann Sci* **35**: 11–30.

Brandeau ML and Chiu SS (1992). A center location problem with congestion. *Ann Opns Res* **40**: 17–32.

Festa P and Resende MGC (2001). GRASP an annotated bibliography. In: Hansen P and Ribeiro CC (eds). *Essays and Surveys on Metaheuristics*. Kluwer Academic Publishers: Norwell, MA.

Gendreau M, Laporte G and Semet F (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Comput* **27**: 1641–1653.

Gerrard RA and Church RL (1996). Closest assignment constraints and location models: Properties and structure. *Loc Sci* **4**: 4.

Goldberg JB (2004). Operations research models for the deployment of emergency services vehicles. *EMS Mngt J* **1**: 1.

Hakimi S (1964). Optimal location of switching centers and the absolute centers and medians of a graph. *Opns Res* **12**: 450–459.

Harewood S (2002). Emergency ambulance depolyment in Barbados: A multi-object approach. *J Opl Res Soc* **53**: 185–192.

Jamil M, Baveja A and Batta R (1999). The stochastic queue center problem. *Comput Opns Res* **26**: 1423–1436.
Kleinrock L (1975). *Queuing Systems. Volume I: Theory.* John Wiley and Sons: New York.
Larson RC (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Comp Opl Res* **1**: 67–95.
Mandell MB (1998). Covering models for two-tiered emergency medical services systems. *Loc Sci* **6**: 355–368.
Marianov V and Revelle C (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Econ Plann Sci* **28**: 167–178.
Marianov V and Serra D (1998). Probabilistic, maximal covering location–allocation models for congested systems. *J Region Sci* **38**: 401–424.
Prais M and Ribeiro C (2000). Reactive GRASP: An application to a matrix decomposition problem in TDMA traffic assignment. *INFORMS J Comput* **12**: 164–176.
Resende MGC (1998). Computing approximate solutions of the maximum covering problem with GRASP. *J Heuristics* **4**: 161–177.
Rojeski G and ReVelle C (1970). Central facilities location under an investment constraint. *Geograph Anal* **2**: 343–360.
Verter V and Lapierre SD (2002). Location of preventive health care facilities. *Ann Opns Res* **110**: 123–132.

3

Decision Support Tools for Ambulance Dispatch and Relocation

T. Andersson and P. Värbrand
Linköping University, ITN, Norrköping, Sweden

In this paper, the development of decision support tools for dynamic ambulance relocation and automatic ambulance dispatching is described. The ambulance dispatch problem is to choose which ambulance to send to a patient. The dynamic ambulance relocation problem occurs in the operational control of ambulances. The objective is to find new locations for some of the ambulances, to increase the preparedness in the area of responsibility. Preparedness is a way of evaluating the ability to serve potential patients with ambulances now and in the future. Computational tests using a simulation model show that the tools are beneficial in reducing the waiting periods for the patients.

3.1 Introduction

The two most important ambulance logistics services are the medical treatment of patients and the transportation of patients. An ambulance call typically starts with a triage, where medically qualified personnel determine the urgency of the call. This is followed by an ambulance assignment, where an ambulance dispatcher decides which ambulance to send to the call site (where the patient is located). The time from when the call has been received until the ambulance personnel have reached the patient is called the *waiting period*. The waiting period is more commonly referred to as the response time, but as it has been shown that many different definitions of response time are used

Reprinted from *Journal of the Operational Research Society*, 58: 195–201, 2007, 'Decision Support Tools for Ambulance Dispatch and Relocation', by T. Andersson and P. Värbrand. With kind permission from Operational Research Society Ltd. All rights reserved.

(Moeller, 2004), waiting period is used here instead. The waiting period is the time from when the call has reached the emergency operator until the ambulance personnel have reached the patient. After the medical treatment, if this is necessary, the ambulance will transport the patient to a health care facility. Not all ambulance calls are urgent; non-urgent transportations can be ordered several days in advance, making it possible to perform some sort of transportation planning. The core of ambulance logistics is however to plan and control the emergency medical services, that is, take care of the urgent calls. An ambulance dispatcher is commonly found in an emergency centre, where distress calls are received and from where the resources are controlled. The main tasks for the dispatcher are to assign ambulances to incoming calls, and to ensure that there are units available to serve future calls. The dispatchers are also expected to support the ambulance personnel with route guidance and possibly medical advice.

In Sweden, the foremost ambulance logistics provider is called SOS Alarm AB. They are responsible for receiving all calls to the national emergency number, 112, and also for controlling all ambulance movements. The operations are run from a SOS centre, of which there is one in each county (administrative region) in Sweden. This paper describes a number of applications that can support the ambulance dispatchers in the SOS centres. The development of these decision support tools has been done as a part of OPAL—Optimized Ambulance Logistics, which is a joint project between SOS Alarm and Linköping University.

From an operations research perspective, the contributions in the ambulance logistics area have mainly focused on reducing the waiting periods by trying to find optimal locations for ambulance stations. The first models dealing with the emergency station location problem surfaced in the 1970s (eg Toregas *et al*, 1971; Church and Revelle, 1974), and new models and algorithms keep appearing, as do surveys of the area (see eg Brotcorne *et al*, 2003; Goldberg, 2004).

In order to evaluate the value of a set of ambulance station locations, it is possible to use simulation. One early simulation model was used for evaluating possible improvements in ambulance service is described by Savas (1969), and a more recent simulation study is described by Henderson and Mason (2004). Furthermore, the hypercube model (Larson, 1974), and later extensions of this work, can be_used to evaluate a solution from a location model. The contributions in operational ambulance control are much more sparse. It is however possible to identify two problems that have received some attention: the ambulance dispatch problem and the ambulance relocation problem.

The most common and natural dispatch rule is to send the closest unit, since a general objective is to minimize the response times. However, this rule is not always optimal (Carter *et al*, 1972). Consider a case where two units, A and B, have equally large areas of responsibility, but A's area has a significantly higher call frequency. In this case, the mean response time will decrease if B is allowed to respond to some of the calls for which A is the closest unit. This result can be generalized for cases involving more than two units (Cunningham-Green and Harries, 1988), and it may be better to send unit C to take A's call when A already is busy, than to send the closer unit B (Repede and Bernardo, 1994). This is done if the call frequency in B's primary district is higher than in C's. In Weintraub *et al* (1999) a dispatch system for vehicles servicing the electrical system in Santiago de Chile is described. Vehicles travel from call site to call site, and the dispatcher tries to maintain an adequate preparedness for quickly servicing high-priority calls when deciding which unit should be assigned to each call.

The demand for ambulances commonly varies with time, and some efforts to compensate for this by matching the amount of resources to the demand have been made. Even without changing the number of ambulances, it is possible to plan relocations of the existing fleet to better match the demand, if it changes in the area of responsibility during the day for example. Carson and Batta (1990) present a way of locating just one ambulance at a campus, where its position is changed several times during a 24-h period, to compensate for population changes. To find pre-planned relocations, a location problem can be solved for each interesting time period. For a specific time period, one set of location points is used, and in the shift to a new period, ambulances have to be relocated to a new set of location points (Repede and Bernardo, 1994).

In practice, it is maybe more common to dynamically relocate ambulances to cover for busy units. A dynamic relocation algorithm for fire companies is developed by Kolesar and Walker (1974), and a call for relocation is triggered when some part of the city is not covered by any unit. Gendreau *et al* (2001) describe a tabu search heuristic for the dynamic relocation of ambulances, and a similar model for physician cars is presented by Gendreau *et al* (2006). Both models maximize the coverage of the area.

In this paper, new algorithms for the ambulance dispatch and the dynamic ambulance relocation problems are presented. The assumptions made in the dispatch algorithm are similar to the work in Weintraub *et al* (1999) in that the closest unit is not always sent to a new call, but adapted for the pick-up and delivery nature of ambulance calls rather

than the repair problem that is studied in Weintraub *et al* (1999). The relocation algorithm is dynamic, that is, the problem is solved when there is a lack of ambulances somewhere in the area. This is a similar problem to the one studied by Kolesar and Walker (1974) and Gendreau *et al* (2001). The major difference is that it is possible to relocate an ambulance to any zone in the area, not just to vacant stations, as shown in Kolesar and Walker (1974). It is also possible to run the algorithm on a common PC, that is, no expensive hardware is needed, and still obtain solutions within a few seconds. Another novel aspect of the algorithms is that both of them utilize a new quantitative preparedness measure.

3.2 Preparedness

In ambulance logistics, preparedness has been used as a qualitative measure for a long time, but two people do not always mean the same thing when using the word. Also, two ambulance controllers may have different opinions on what can be considered good or bad preparedness, depending on their experience, their risk aversion and their personality. For example, one controller may think that less than 20 available ambulances in the county means that the preparedness is low, while another controller thinks that 20 ambulances are more than enough for an adequate preparedness.

In order to find a quantifiable measure for preparedness, we first divide the area of consideration into a number of zones. To each zone j, a weight c_j is associated, which mirrors the demand for ambulances in the zone. The weight can for example be proportional to the number of calls served in the zone during a specific time period, or to the number of people currently resident in the zone. The preparedness in zone j can then be calculated as

$$p_j = \frac{1}{c_j} \sum_{l=1}^{L_j} \frac{\gamma^l}{t_j^l} \tag{1}$$

where L_j is the number of ambulances that contribute to the preparedness in zone j, t_j^l is the travel time for ambulance l to zone j, γ^l is the contribution factor for ambulance l and the following properties hold:

$$t_j^1 \leq t_j^2 \leq \ldots \leq t_j^{L_j} \tag{2}$$

$$\gamma^1 > \gamma^2 > \ldots \gamma^{L_j} \tag{3}$$

Thus, the preparedness is calculated by letting the L_j closest ambulances to zone j contribute to the preparedness with an impact that is decreasing as the travel time to the zone increases. One basic quality of (1) is that the preparedness in a zone increases if an ambulance moves closer to the zone, that is, some t_j^l decreases. Furthermore, if the call frequency in a zone, that is, c_j, increases, the preparedness decreases (Andersson, 2005).

3.3 Dispatching support

In Sweden, the prioritization of an ambulance call results in one of three degrees, Prio 1, 2 or 3. Prio 1 calls are the most urgent, life-threatening calls, while Prio 2 are urgent but not life threatening and Prio 3 are non-urgent calls.

Sometimes it is trivial to decide which ambulance to assign to a new call, for example, for a Prio 1 call that requires only one ambulance, the ambulance with the shortest expected travel time to the call site is always dispatched. If the call is not as urgent, an ambulance dispatcher may choose to assign an ambulance with a longer travel time, if this assignment means that the drop in preparedness will be less significant. The dispatcher may also reassign an ambulance already on its way to a call site, if the new call is more urgent.

The implementation of the preparedness measure (1) includes a list of the closest ambulances for each zone, sorted according to the expected travel time. Thus, it is easy to find the closest ambulance to a certain zone. To check which ambulance to dispatch to a Prio 2 or 3 call, an algorithm (see Table 3.1) has been developed that checks all available ambulances within a certain travel time from the zone, and picks the one whose unavailability causes the least drop in the preparedness as calculated by (1).

The algorithm in Table 3.1 starts, after the initializations, by checking the priority of the call. If it is a Prio 1 call, that is, PRIO(j) = = 1 in Step 2, the closest ambulance is dispatched to the call. If the call is a Prio 2 or 3 call, the algorithm starts by checking if the closest ambulance, that is, the one first in the list, can reach the zone within T_2 (or T_3) minutes. If not, the algorithm will stop and the closest ambulance will be dispatched. If it can, the ambulance will be set as unavailable, and new levels of preparedness will be calculated for all zones in NC^k, which is the set of zones that will be affected by the assignment (Step 6). The lowest level for any of the zones, p_{min}, is saved, and used as a measure on how the preparedness is affected by the dispatch. When the first ambulance in the list has been processed, the algorithm checks if ambulance number two can reach the zone within T_2 (or T_3) minutes. If it can, a

Table 3.1 The ambulance dispatch algorithm

1. Let j be the zone to where an ambulance needs to be dispatched, and $l = 1, ..., L_j$ an ordered list of the ambulances that contribute to the preparedness in j. Let $A = \varnothing$ be the ambulance that is dispatched and let $p_{min} = 0$.
2. IF PRIO(j) = = 1
3. Set $A = 1$ and dispatch A, that is, dispatch the ambulance that is closest to zone j and therefore first in the list.
4. IF PRIO(j) = = 2 OR 3
5. Check the ambulances in the list, beginning with the closest:
   FOR $l = 1, ..., L_j$
   IF $t_j^l <$ T2 (or T_3 if PRIO(j) = = 3)
6. Remove ambulance l from the list of ambulances contributing to the preparedness, and recalculate the preparedness, p_i, in all zones that are affected by this action.
   IF $\min_{i \in N}\{p_i\} > p_{\min}$
7. $p_{\min} = \min_{i \in N}\{p_i\}, A = l$
8. Dispatch A

new minimum preparedness level is calculated and compared to the current p_{min}. This is continued until an ambulance is too far away from the zone or until there are no ambulances left in the list.

By letting ambulances on their way to a Prio 2 or 3 call still contribute to the preparedness, it is also possible to assign these to calls that are more urgent, for example, an ambulance on its way to a Prio 3 call, can be assigned to a new Prio 2 or Prio 1 call. In this case it is necessary to check, in Step 5, if the ambulance is already on its way to serve a call.

To ensure that the waiting periods for the less urgent calls do not grow beyond what is practically feasible, *pseudo* priorities are used when ambulances are reassigned. The *pseudo* priority for a call changes if the call has not been served within a certain time, for example, a Prio 3 call that has not been reached by an ambulance in T_3 minutes changes *pseudo* priority from 3 to 2. This means that an ambulance that is on its way to serve this Prio 3 (pseudo Prio 2) call cannot be reassigned to a Prio 2 call, but still to a new Prio 1 call. It may be noted that also the real priority of a call, and thus not only the pseudo priority, may change if a patient has to wait for medical care.

3.4 Dynamic ambulance relocation

By colour coding the measure (1) in a geographical information system, an ambulance dispatcher can manually check where the preparedness

is low, and thus where to send ambulances. What is even more useful is a tool that automatically checks the preparedness in the zones, and suggests ways to relocate ambulances in order to maintain a sufficient level. The measure (1) can be used as a base for this kind of tool. First however, it is necessary to calibrate the measure (ie deciding parameter values for c and γ) and finding a lowest level of preparedness, P_{min}, that should be kept in all zones. In Andersson (2005), the measure (1) is calibrated for use in the county of Stockholm in Sweden, and a level P_{min} of 0.923 is identified.

The ambulance relocation problem occurs when one or more zones have a preparedness level less than P_{min}. The objective is then to reach the P_{min} level in all zones as quickly as possible. The preparedness is increased by relocating one or more ambulances closer to the zones that suffer from a low level of preparedness. A model, DYNAROC, that solves the dynamic ambulance relocation problem follows:

$$\min z \tag{4}$$

$$\text{s.t.} \quad z \geq \sum_{j \in N^k} \tau_j^k x_j^k, \quad k = 1, \ldots, A \tag{5}$$

$$\sum_{j \in N^k} x_j^k \leq 1, \quad k = 1, \ldots, A \tag{6}$$

$$\sum_{k=1}^{A} \sum_{j \in N^k} x_j^k \leq M \tag{7}$$

$$\frac{1}{c_j} \sum_{l=1}^{L_j} \frac{\gamma^l}{t_j^l(\boldsymbol{x})} \geq P_{min}, \quad j = 1, \ldots, N \tag{8}$$

$$\boldsymbol{x} \in \{0, 1\} \tag{9}$$

The objective (4) is to minimize the variable z, which is the maximum travel time for any of the relocated ambulances, that is, the time it will take until the preparedness is at least P_{min} in all zones, which is required in (8). Constraint (5) states that z has to be greater than or equal to any of the travel times τ_j^k, which is the time required for ambulance k to reach zone j. The variable x_j^k equals 1 if ambulance k is relocated to zone j. Each of the ambulances can be relocated to at

most one zone in the set N^k (6), which is the set of zones that can be reached by ambulance k in less than R minutes. By setting R low, the set N^k and the set of feasible solutions will be smaller. This will however also decrease the number of feasible solutions to the model, with the risk that there will not exist any solutions in some instances. R is an upper bound on the objective function variable z, which means that if R is set, for example, to 20 min, no ambulance will have a relocation travel time longer than 20 min. Constraint (7) ensures that not more than M ambulances are relocated. $t_j^l(\boldsymbol{x})$ in constraint (8) is a function of the variable $\boldsymbol{x}$, which is the vector form of x_j^k. Naturally, the travel time for the l'th closest ambulance to zone j, that is, t_j^l, depends on where the ambulances are located, which is decided by the values on the variable $\boldsymbol{x}$.

Since a short computation time is of uttermost importance, a tree-search heuristic is used to solve DYNAROC. It is schematically described in Table 3.2 and illustrated in Figure 3.1. Starting with the current situation, the heuristic iteratively tries to raise the preparedness in the zone with the lowest preparedness. This is done by moving ambulances closer to this zone. Only ambulances that can be moved close enough to raise the preparedness in the zone to at least P_{min} are evaluated. Ambulances that can raise the preparedness enough are compared to each

Table 3.2 The tree search algorithm that finds solutions to DYNAROC

1. Let the current (infeasible) solution, that is, $x_j^k = 0\ \forall j, k$, be the root of the tree.
2. Let j = the zone with the lowest preparedness.
3. REPEAT
4. Find the n ambulances, with the minimum travel times, that can be relocated in a way that ensures that $p_j \geqslant P_{min}$. Save a maximum of m zones for each of these ambulances that satisfy the conditions above. The ambulances must not have been relocated once already (earlier in the tree) and not more than $M - 1$ ambulances must have been moved.
5. Each of the moves in Step 4 gives a potential solution. FOR all new solutions
6. Check if $p_i \geqslant P_{min}\ \forall i = 1, \ldots, N$ and check the longest travel time against the best solution found so far if this is true. If the new solution is not feasible, create a new node and connect it to its parent solution.
7. Pick a new node and let j = the zone with the lowest preparedness in the new solution.
8. UNTIL there are no nodes left to examine, or some other stop criterion triggers

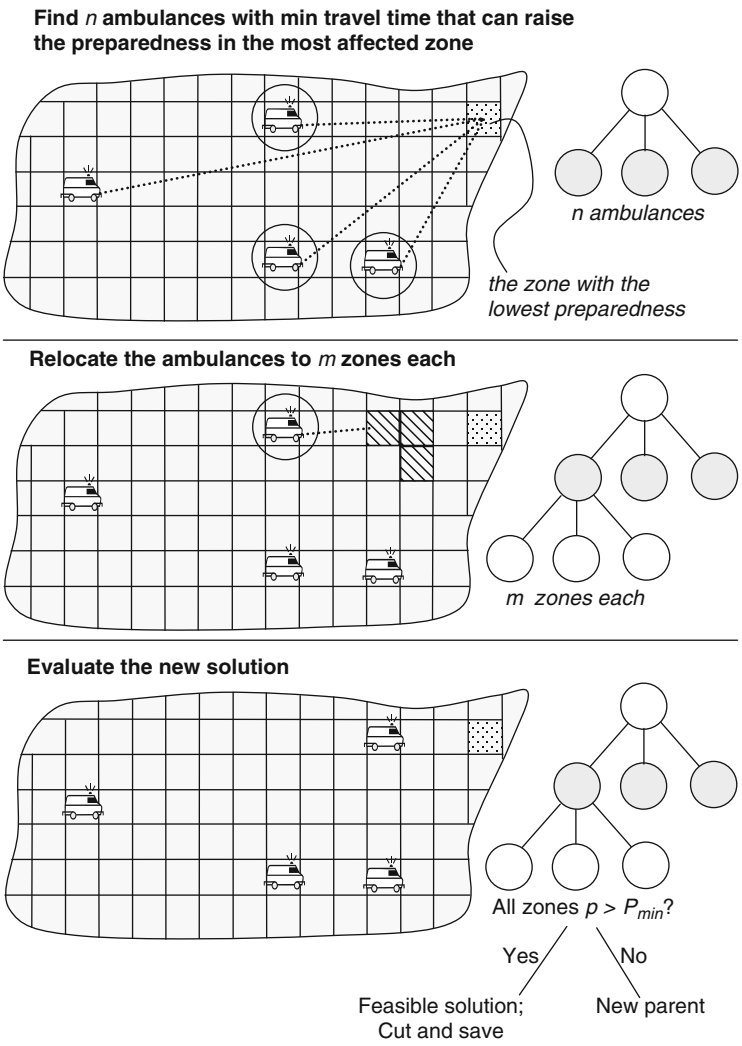


Figure 3.1 Illustrating the tree-search heuristic that solves the DYNAROC model

other concerning the required travel time for the relocation (Step 4). If possible, n ambulances are saved, and each of these ambulances is relocated to a maximum of m zones. The reason for not just saving a set of the best possible relocations is that these may be performed by a single ambulance, and a certain divergence in the search for solutions is desired.

Every move that has raised the preparedness is checked for feasibility in Step 6. If some zone has a preparedness below P_{min}, the move has given a new infeasible solution that is saved in the search tree. In the new solutions, another ambulance is relocated to raise the preparedness. In Figure 3.1, n and m are both set to 3, that is, three ambulances are marked as potential relocation candidates, and for each of these the three best relocation zones are identified. Best in this aspect means the zones to which the travel times are shortest, but where a relocation would raise the preparedness to above the threshold value. The relocation of an ambulance gives rise to a new potential solution, which is feasible if the preparedness in all zones is above the threshold value. If the solution is not feasible, it is possible that further relocations are needed. Therefore, a new node is created and the infeasible solution is used as a start solution (a parent node) further down in the search tree.

The first level of the search tree contains (infeasible) solutions where only one ambulance has been relocated, and the second level where two ambulances have been moved. Thus, the tree will never grow more than M levels and each node can have at most nm children. If no feasible solution can be found, the search tree will have searched a total of $1 + (nm)^M$ nodes. With a feasible solution, it is however possible to cut extensively during the search process. It is never beneficial to evaluate a solution where one of the relocation travel times is longer than the objective function for the currently best feasible solution. It is therefore possible that all nodes will be examined and the algorithm terminated because of this, but an alternative stop criterion, for example, number of iterations or elapsed computation time should be used as well.

3.5 An ambulance operations simulator

Simulation of the ambulance operations can be used for evaluating strategic decisions, such as where to locate ambulance stations or how large the ambulance fleet should be. It can also be a valuable tool for education and training of the ambulance dispatchers. Moreover, it can be used as a visualization tool for information purposes towards customers, decision-makers, and the public.

The two decision support tools described earlier, which give suggestions on ambulance assignments and ambulance relocations, are necessary for it to be possible to simulate the ambulance operations faster than real time. When simulating in real time, an ambulance dispatcher can decide which ambulance to send to the incoming calls, and if any relocations should be made. This would however be very

time consuming if the simulation tool is to be used for evaluating strategic decisions, when it is necessary to simulate days, weeks or months of operations in order to get significant results. Thus, the ambulance dispatcher, or more accurately the decisions made by him or her, has to be simulated as well. This includes foremost dispatching decisions and relocation decisions.

The developed simulation model is schematically described in Figure 3.2. The *call generator* simulates incoming calls to the emergency centre; in the current implementation, these are stochastically generated, but in the future, it should be possible to use historical sequences of calls, or to use sequences that are constructed to serve a certain, for example, educational, purpose. The *resource simulator* handles the incoming calls and the ambulances that are used to serve them. The simulation is time based, and a time step of one minute is used, so each minute there is a certain possibility, depending on the population, that a call will be generated in a zone. When a new call is generated, the *automatic dispatch module* finds an ambulance to assign to the call. If the ambulance was already on its way to serve a call, the dispatch model finds a new ambulance to serve the old, less urgent call. If there are no available, or soon to be available, ambulances to assign to a call, it will be put in a queue. Each iteration starts with the processing of the

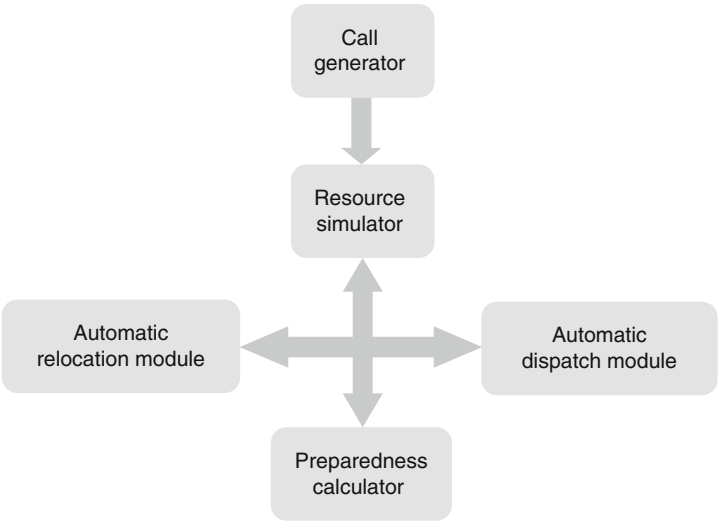


Figure 3.2 The ambulance operations simulator

calls on queue, to see if any ambulances have become available and are able to serve them.

The assignment of ambulance to a call may affect the preparedness level in some zones, which is why the *preparedness calculator* is used to check if the level has dropped below a certain threshold value, P_{min}. If it has, the *automatic relocation module* tries to find a relocation of one or more ambulances that will raise the preparedness in the affected zones.

3.6 Computational results

The algorithms are tested using data for the county of Stockholm in Sweden. It offers the most complex ambulance control situation in Sweden, with at most 58 ambulances and about 400 ambulance calls a day. The data that is used consists of travel times in minutes between the 1240 zones into which the county is divided. The travel times were originally collected for fire engines, but were deemed directly translatable to ambulances. Population data for each of the zones is used to calculate the weights, c_j. The population in the zones varies between 1 and 17985 and these values are divided by 20 000, in order to get more convenient values to work with, which means that l/c_j varies between 0.9 and 20 000. γ^l is set to $1/2^{l-1}$ for $l = 1,2, \ldots, 7$, that is, $\gamma^1 = 1$, $\gamma^2 = 0.5$, $\gamma^3 = 0.25$, $\gamma^4 = 0.125$, etc. A maximum of seven ambulances are used to calculate the preparedness for a zone. Furthermore, a maximum relocation travel time (R) of 40 min is used in DYNAROC, meaning that N^k is the set of zones that can be reached by ambulance k within 40 min. M is set to 3, as it is reasoned that an ambulance dispatcher would be reluctant to relocate too many ambulances. n and m in the DYNAROC algorithm in Table 3.2 is set to 5 and 3, respectively, and a maximum of 50 iterations is performed by the DYNAROC algorithm. On a PC computer with a 1000 MHz Pentium III processor, it takes about 6s to perform 50 iterations. However, the heuristic most often finishes before 50 iterations are performed. In a quick test where DYNROC is solved 2275 times, the mean running time for the algorithm is 2.24 s, and the maximum running time 5.89 s. About 53% of the calls generated in the simulations are Prio 1 calls, 33% Prio 2 and 14 % Prio 3 calls, in accordance with historical data for the county of Stockholm. The probability that a call will appear is proportional to the population in the zone.

One proposition made in Andersson (2005), is that if the minimum level of preparedness is kept above 0.923 in the county of Stockholm, the waiting period targets should be satisfied. The targets differ depending on the priority of the call. For Prio 1 calls, the target is that 75% of

all calls should be served within 10 min, 95% within 15 min and 99% within 20 min. For Prio 2 and 3 calls, these times are allowed to be longer. To test this, simulations of 1000 hours of operations with P_{min} = 0.923, are performed and compared to simulations with P_{min} = 0.277 and 0, 0.277 is the lowest level of preparedness in the county when all ambulances are available at their stations, and when P_{min} is zero, no relocations are performed. As can be seen in Figure 3.3, the number of calls served within the specified times increases as P_{min} is increased. Especially the waiting period target that 99% of all Prio 1 calls should be served within 20 min, benefits from the relocations. It should be noted that instant relocations are used. This means that the relocation travel time (τ_j^k in DYNAROC) for each relocated ambulance is set to a very small value, and thus the ambulances are instantly transferred to their new zones. The reason for this is that the simulation model has no way of keeping track of the ambulances while they are relocating. Thus, when the actual relocation travel times are used and an ambulance that is being relocated is assigned to call, the travel time to the new call will be overestimated.

Even though instant relocations are used, the minimum level of preparedness does not stay above the threshold value at all times; in fact, as the number of calls is increased, the time p_{min} is above 0.923 drops from 96.1% at 200 calls per day to 0.22% at 800 calls a day. This is because the heuristic increasingly fails to find relocations that will improve the situation (at 800 calls a day the heuristic fails to find a solution for P_{min} = 0.923 in 38.4% of the cases). Here it should be noted that 800 calls a day is almost double to the normal amount, and the ambulance fleet is not large enough to handle that kind of call frequency for a longer period. Therefore it is not surprising that improving relocations are hard to find. Still, it seems reasonable to state that the waiting period targets probably will be satisfied if a minimum preparedness level of 0.923 is kept at all times.

However, trying to maintain a high minimum preparedness level requires an unreasonable amount of relocations, as indicated by Table 3.3. Looking at the results when P_{min} = 0.923, in average more than one

Table 3.3 The average number of relocated ambulances per day

No.of calls/day	**200**	**250**	**300**	**350**	**400**	**450**	**500**	**550**	**600**	**650**	**700**	**750**	**800**
P_{min} = 0.277	4.4	6.7	9.3	10.2	14.7	19.6	25.1	33.9	51.6	65.1	101.5	146.2	223.6
P_{min} = 0.923	105.5	148.7	194.0	253.4	331.4	412.5	500.4	578.5	662.2	700.9	752.3	778.4	819.6

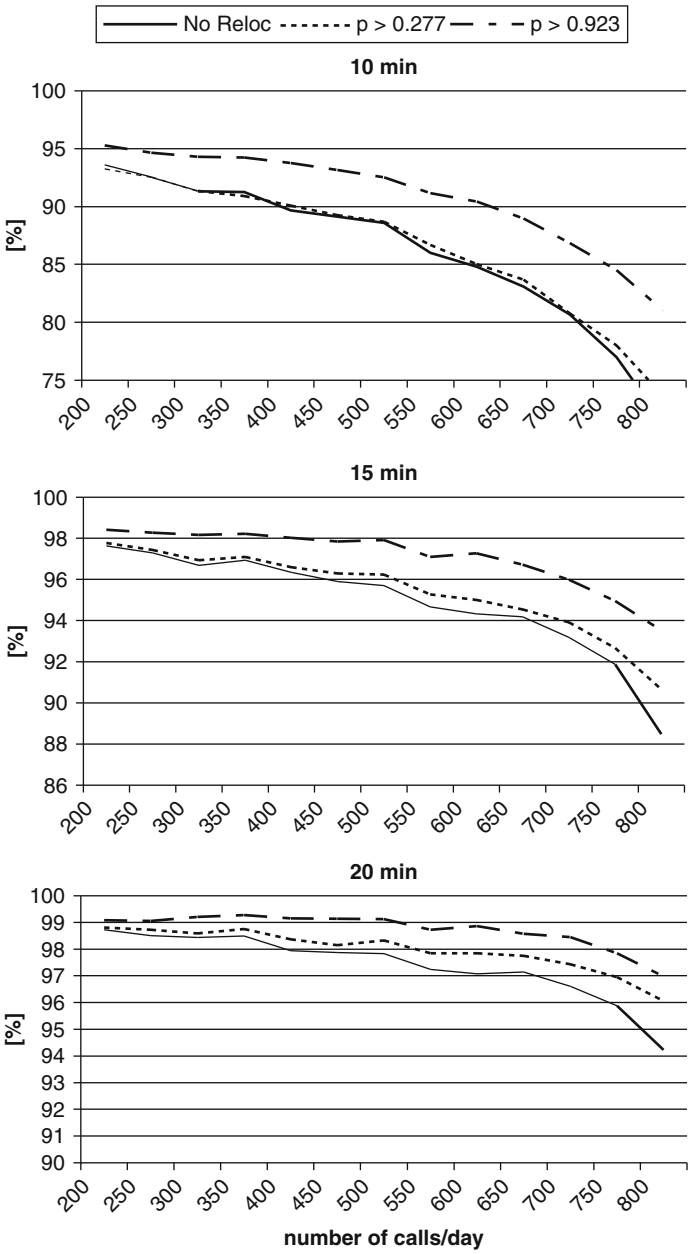


Figure 3.3 Number of Prio 1 calls served (in %) when using a threshold value of 0, 0.277, and 0.923, respectively

ambulance is relocated for each call when the call volumes are high, even though the heuristic often fails in finding a solution.

How often an ambulance may be relocated has to be negotiated between the ambulance personnel, or the company providing the vehicles, and the people responsible for the ambulance health care. Relocations will most probably increase the total travel distance for the ambulances, and they cause the ambulance personnel to spend more time on the road and less at the station. Therefore, it is natural that the ambulance personnel want to be compensated if the number of relocations increases.

3.7 Conclusions

In this paper, decision support tools for dynamic ambulance relocation and automatic ambulance dispatching are presented. The tools utilize a measure for preparedness, which is a way of evaluating the ability to serve current and future calls anywhere in the area. Simulations show that maintaining a high level of preparedness as calculated by the developed measure, is helpful in reaching the waiting period targets that are set by the county councils in Sweden. During the simulations, the preparedness is improved by dynamically relocating ambulances and a large amount of relocations is necessary to get significant results. The purpose of the relocations is to keep the preparedness in the area high. Using a different set of ambulance station locations is another possible way of raising the minimum preparedness level, at least initially. By changing these locations as the demand changes in the area, that is, using pre-planned relocations, it will be easier to maintain this level, which may decrease the need for dynamic relocations. Dynamic relocations can then be used as they are in practice now, to cover for a temporary and serious lack of ambulances somewhere in the area.

Acknowledgements

This research was partly sponsored by SOS Alarm. We thank SOS Alarm for the privilege of working with them and for the constant support and encouragement.

References

Andersson T (2005). *Decision support for dynamic fleet management*. PhD thesis, Department of Science and Technology, Linköping University, Sweden.

Brotcorne L, Laporte G and Semet F (2003). Ambulance location and relocation models. *Eur J Opl Res* **147**: 451–463.

Carson Y and Batta R (1990). Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces* **20**(5): 43–49.

Carter G, Chaiken J and Ignall E (1972). Response areas for two emergency units. *Opns Res* **20**(3): 571–594.

Church R and ReVelle C (1974). The maximal covering location problem. *Papers in Regional Science* **32**(1): 101–120.

Cunningham-Green R and Harries G (1988). Nearest-neighbour rules for emergency services. *ZOR—Zeitschrift for Opns Res* **32**: 299–306.

Gendreau M, Laporte G and Semet F (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing* **27**: 1641–1653.

Gendreau M, Laporte G, Semet F (2006). The maximal expected coverage relocation problem for emergency vehicles. *J Opl Res Soc* **57**: 22–28.

Goldberg J (2004). Operations research models for the deployment of emergency service vehicles. *EMS Mngt J* **1**(1): 20–39.

Henderson S and Mason A (2004). Ambulance service planning: Simulation and data visualization. In: Brandeau M, Sainfort F and Pierskalla W (eds). *Operations Research and Health Care*. Kluwer Academic Publishers, Boston, pp 77–102.

Kolesar P and Walker W (1974). An algorithm for the dynamic relocation of fire companies. *Opns Res* **22**(2): 249–274.

Larson R (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput Opns Res* **1**(1): 67–95.

Moeller B (2004). Obstacles to measuring emergency medical services performance. *EMS Mngt J* **1**(2): 8–15.

Repede J and Bernardo J (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur J Opl Res* **75**: 567–581.

Savas E (1969). Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Mngt Sci* **15**(12): B608–B627.

Toregas C, Swain R, ReVelle C and Bergman L (1971). The location of emergency service facilities. *Opns Res* **19**(6): 1363–1373.

Weintraub A *et al* (1999). An emergency vehicle dispatching system for an electric utility in Chile. *J Opl Res Soc* **50**: 690–696.

4

A Study of Situationally Aware Routing for Emergency Responders

M. J. Henchey[1,2], *R. Batta*[2,3], *A. Blatt*[3], *M. Flanigan*[3] *and K. Majka*[3]

[1]*Herren Associates, Washington, DC, USA;* [2]*University at Buffalo, State University of New York, Buffalo, USA; and* [3]*Center for Transportation Injury Research, CUBRC, Buffalo, USA*

This paper presents tests conducted on routes determined from a Dijkstra-based shortest path problem and a Variance-Constrained Shortest Path problem under varying conditions of traffic and weather in a simulated 'smart environment'. Utilizing envisioned future advanced transportation systems' real-time information on traffic parameters allows data fusion techniques to provide situation awareness to its users. Taking advantage of this real-time data, the routing methodologies and data capture techniques studied in this paper provides Emergency Medical Services with better routes when responding to a vehicular crash. Comparing the performance of both routing methodologies in terms of both their ability to provide better routes as well as computation times demonstrates two alternatives for aiding in future emergency response.

4.1 Introduction

The development of advanced Intelligent Transportation Systems (ITS) will create a 'smart environment' in which real-time data capture will provide information on traffic conditions, CO_2 emissions, weather and even dangers a driver may experience if they do not take action (RITA, 2011a, b). The primary objective of this paper is to compare the performance of a conventional shortest path model from Dijkstra (1959) and a Variance-Constrained Shortest Path (VCSP) model from Sivakumar and Batta (1994) for Emergency Medical Service (EMS) responders in a

Reprinted from *Journal of the Operational Research Society*, 66(4): 570–578, 2015, 'A Study of Situationally Aware Routing for Emergency Responders', by M. J. Henchey, R. Batta, A. Blatt, M. Flanigan and K. Majka. With kind permission from Operational Research Society Ltd. All rights reserved.

nondeterministic 'smart environment'. The simulation model utilizes traffic-monitoring sensors to aid in routing decisions determined by both the shortest path models, providing information on road segment congestion. Each are tested under varying conditions of traffic and weather, and the performance results are presented along with computation times required to obtain the solutions.

Section 2 summarizes the motivation behind this paper and related work in simulating future transportation systems and emergency response. Section 3 provides a brief literature review of the state-of-the-art advanced technologies used in ITS, traffic simulations and emergency vehicle-routing methods. Section 4 presents the simulation model, laying out the study area and the assumptions that were required about traffic flow, EMS travel and sensor technology and data capture. Section 5 demonstrates the shortest time path methodology, including the data requirements to solve for the optimal route, the design of experiments to test the travel time estimation technique and results of the experiments. Section 6 presents a brief overview of the VCSP problem, its performance in providing routes to EMS vehicles responding to a vehicular crash within the study area. Finally, Section 7 presents concluding remarks for this study.

4.2 Motivation

As part of an ongoing project with the University at Buffalo and CUBRC's Center for Transportation Injury Research, the Automated Situation-awareness Assessment Platform (ASAP) is a research tool being developed to simulate these sensor-rich environments, enable fusion of the data acquired and assess potential improvements in responder awareness. An earlier paper focused its attention on detailing the development of a simulation methodology in order to 'create' an ITS environment for use in studying problems the ASAP hopes to address (Henchey *et al*, 2013). More specifically, it presented the 'Simulation in ARENA' module, which is part of the approach shown in Figure 4.1.

The components shown in bold under the 'Sensed Data', 'Data Fusion', 'Information' and 'Decisions' modules in Figure 4.1 are the foci of this paper. The development of ITS technologies will make acquiring real-time data feasible; therefore, the problem of interest is the determination of a method for using available real-time data on congestion for the traffic network. To achieve better estimates for travel times, data fusion techniques are used that incorporate the real-time congestion data as well as weather data. These travel time estimates then lead to the

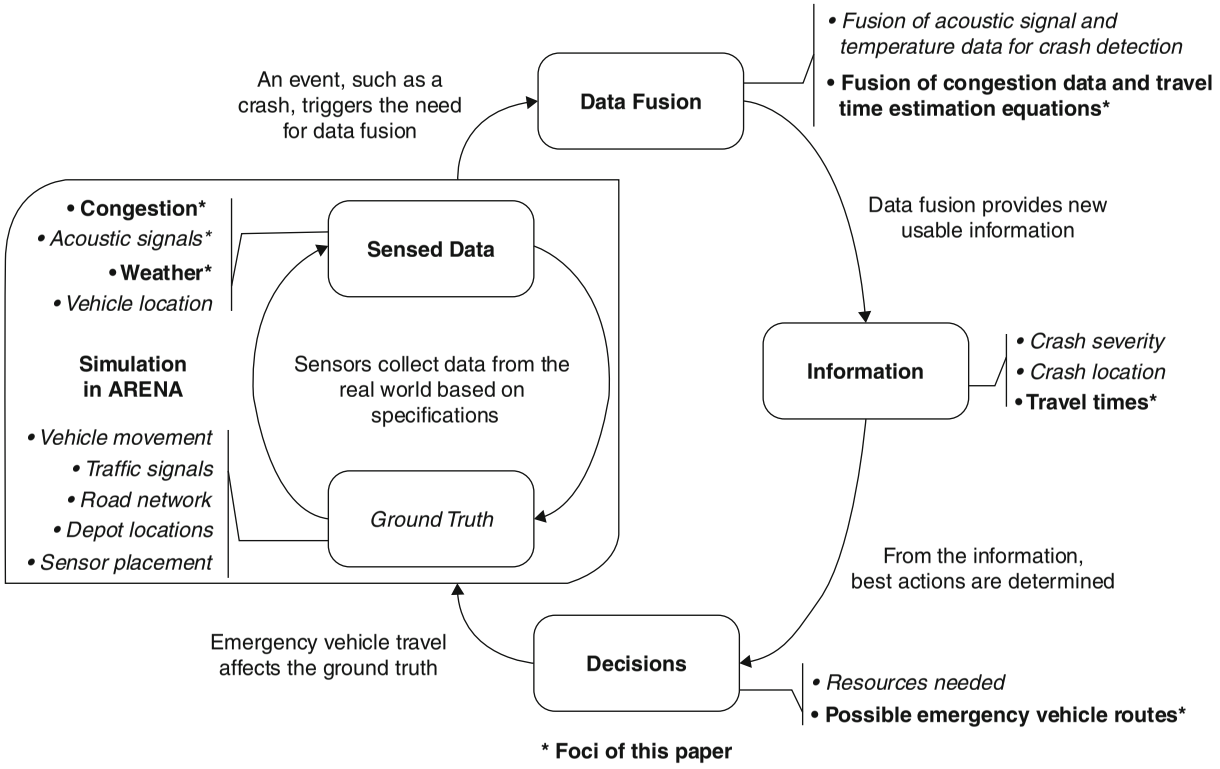


Figure 4.1 Information fusion approach to ITS

identification of efficient travel routes for the emergency vehicles. This paper compares the performance of two possible routing methods that use these data fusion techniques, a shortest path model from Dijkstra (1959) and a VCSP model from Sivakumar and Batta (1994).

4.3 Literature review

The accuracy and performance of roadside sensors as well as user-based sensors (smartphones, Global Positioning System navigation units, etc) continues to improve. The Cooperative Intersection Collision Avoidance System (CICAS) program uses multiple sensors and Dedicated Short Range Communication technologies to communicate potential dangers to drivers (RITA, 2011a). Gorjestani *et al* (2010) demonstrate the configuration and accuracy requirements in a CICAS-Stop Sign Assist (CICAS-SSA) intersection, defining the expectations of future sensor (radar and lidar) capabilities and deployment requirements of ± 0.5 miles per hour and a greater than 99.99% detection rate. Additional research shows inductive loop sensors and active infrared systems capable of 95% accuracy (OHPI, 2004; ITS UW, 2012). Google's services demonstrate how real-time data can benefit travel time estimation, using a combination of historical and real-time data to provide drivers with awareness of traffic conditions Dillow (2012). The Niagara International Transportation Technology Coalition (NITTEC, 2012) uses roadway sensors that report travel speeds to Traffic Operations Centers, which are updated every 60 s (approximately).

These different types of sensors are used as a basis for our work in estimating travel times for use in shortest time path routing, such as Dijkstra's algorithm (1959). The algorithm begins from an origin node and determines the shortest path to each node in the network by iteratively 'visiting' the next node with the shortest path from the origin node until all nodes have been 'visited'. Ultimately, Dijkstra's algorithm was chosen due to its near real-time solution capabilities in the study network. The real-time data provided by user-volunteered information or roadway sensors also supply the required travel costs (time) that can be used in Sivakumar and Batta's (1994) VCSP problem. By observing travel speeds over time, each link in the transportation network can be assigned an expected travel time along with a variance and covariances between links necessary in solving the VCSP problem.

Current research from Nie *et al* (2012) develops a 'reliable *a priori* shortest path' using historical data captured from loop detectors and toll plazas around Chicago, showing that reliable routes could potentially save

up to 5–15% of travel time. Huang (2012) studied adaptive routing with 'online information' of travel times (assuming no errors in this information) at different levels: all information up to the current time (perfect online information), all information up to a specified time before the current time (a lag in information), all information up to the departure time (pre-trip information) and information on a subset of links without lag (radio information). The results from Huang show that the pre-trip information performs nearly as well as the perfect online information using their routing methodology. Employing the simulation model created previously for the ASAP and presented in Henchey *et al* (2013), it is possible to study both the shortest time path and VCSP techniques under varying conditions. The remainder of this paper shows the steps taken to simulate a test environment, the experiments conducted using the simulation, the results and outcomes of those experiments.

4.4 Simulation overview

The transportation network simulation model, developed in ARENA from Rockwell Automation and described in previous work, provides a test bed for studying the Dijkstra-based shortest time path and VCSP techniques (Henchey *et al,* 2013). The simplified study area, presented in Figure 4.2, was modelled to create ground truth (using vehicle arrival

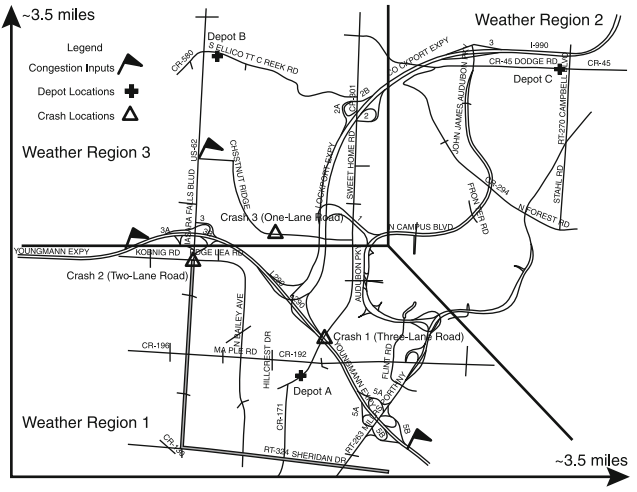


Figure 4.2 EMS first responder locations and crash locations

rates by time of day, signal timing, free-flow travel speeds, etc). Together, these determined the movement of traffic throughout the simulation model and made it possible for sensors to be 'placed' by simulated data-capture of traffic statistics (congestion or occupancy rates, travel speeds, etc) from the road segments within the model.

4.4.1 Regular and EMS vehicle travel

Equations (1) and (2) show the formulation used to simulate the travel speed for regular ($\mathbf{S}_{\mathbf{c}}^{\mathbf{regular}}$) and EMS vehicles ($\mathbf{S}_{\mathbf{c}}^{\mathbf{EMS}}$), respectively, on the road segment *r*. The effect that emergency vehicles have on regular traffic is the **e** term, and the weather effect is the **w** term. The emergency vehicle effect is not well studied as it relates directly to driver behaviour. Because the emergency vehicle routing is modelled at a microscopic level and regular vehicle traffic is modelled at a macroscopic level, the **e** term is necessary to approximate the effect the EMS has on the regular traffic. For **w**, the penalty intensity determination uses data collected by the Federal Highway Administration (FHWA) from studies such as Hranac *et al* (2006), who conducted an empirical study on the effects of weather on different traffic parameters. The term **u** refers to the congestion or utilization of a road segment, which also has an effect on the travel speed.

$$\mathbf{S}_{\mathbf{r}}^{\mathbf{regular}} = \frac{S_r^f \mathbf{w}}{1 + \alpha(\mathbf{ue})^{\beta}} \tag{1}$$

$$\mathbf{S}_{\mathbf{r}}^{\mathbf{EMS}} = \frac{S_r^f \mathbf{w}}{1 + \alpha \mathbf{u}^{\beta}} \tag{2}$$

The values of α and β are coefficients used in the Bureau of Public Road (BPR) curve that approximates the travel speed of vehicles based on a ratio of traffic volume to traffic capacity (the **u** factor in Equations (1) and (2)), which was based on a specified level of service that users of the road experienced in the BPR curve. Two sets of α and β values are studied for EMS vehicles based on views of how an emergency vehicle might operate under different road conditions. The first set of values (0.83 and 5.5 for α and β, respectively) are based directly on the original BPR formulation, and represent an 'insensitive' EMS vehicle that does not begin to slow down due to congestion until the utilization of the road is greater than 50% (when the volume to capacity ratio is about one). The second set of values are the α and β values used for a regular vehicle in the simulation model, 30 and 5,

respectively, which have been scaled to employ ARENA's utilization parameter of the road segments specified by the term **u**. These values represent a 'sensitive' EMS vehicle that begins to slow down much sooner, following the BPR curve.

Free-flow travel speeds (S_r^f) are set at the road speed limit for regular vehicles and at 10 miles per hour over the posted speed limit for EMS vehicles. This assumption comes from the FHWA's (2011b) 'Appendix N' of the *Traffic Analysis Toolbox Volume IV: Guidelines for Applying CORSIM Microsimulation Modeling Software*. The International Association of Fire Chiefs (IAFC) policies and procedures support this assumption. The IAFC (2011) also sets policies on travelling in adverse weather and specific procedures as to how an emergency vehicle must travel through an intersection with either a traffic signal or stop sign. Expected travel times through intersections are based on the signal timings and the IAFC policies. These have been adopted into the simulation model by requiring emergency vehicles to slow down to the speed limit in adverse weather and come to a complete stop when they approach an intersection with either a red traffic signal or stop sign. The simulation assumes no signal preemption is used by EMS responders, realistic for the geographical area modelled.

4.4.2 'Real-time' data capture

The transportation network simulation provides a test bed that can be used to study a variety of problems in future advanced transportation systems. Both the Dijkstra-based shortest path and VCSP approaches use expected travel time estimates, with the VCSP approach also using variance and covariance of travel times.

The expected cost values (l_{ij}) are obtained through the simulation model using real-time data capture from simulated roadside sensors deployed in the network. Traffic congestion rates and weather slowdown effects are captured at 1-min intervals, using Equation (2) to compute estimated travel times associated with each road segment. These sensors are 'placed' in the simulation at intersections as well as major entrances or exits to highways assuming 100% accuracy. Huang (2012) also uses this assumption, which follows the requirement for traffic detection rates in CICAS to be greater than 99.99% (Gorjestani *et al*, 2010).

Data from multiple time periods are used to allow for calculation of expected travel times (l_{ij}) for each link in the transportation network (from i to $j \; \forall \; (i,j)$ in the network). Determining the variance (v_{ij}) of link (i,j) and covariance (c_{ijkl}) between links (i,j) and (k, l) for use in the VCSP method is done by computing a Variance–Covariance Matrix where

entry (i,j), (k, l) is the covariance between (i,j) and (k, l) (when $(i,j) \neq (k, l)$). If $(i,j) = (k, l)$, then this entry is the variance of link (i,j).

Signalized intersections are a special case, as the variance is determined in the same manner as the expected travel times are determined. The IAFC policies and signal timing are used to determine the variance of travel time, based on the travel time for the EMS vehicle arriving at a green or red signal. The number of data sets collected for computation of expected travel times, variance and covariance is described with each experiment.

4.5 Dijkstra-based shortest time path

A variation of Dijkstra's algorithm is used to solve the shortest time path problem. Modifying Dijkstra's algorithm, it is possible to find the shortest time path from an origin node (such as a fire department) to a specific destination node (such as a crash) rather than determining the shortest path to all nodes in the network. Supplementary Appendix B presents the pseudocode used to create a Java program capable of determining the shortest time path. This section will present the results comparing the 'Insensitive' and 'Sensitive' EMS vehicles, the effects of distinct weather patterns in the area, and finally the effects of event-based congestion (eg a crash).

To measure the performance of the travel time estimation, a variable $R = A_e/A_s$ is assigned as the estimated travel time over the average simulated travel time, at the different factor levels. A_e is the estimated travel time from the shortest time path methodology and A_s is the average simulated travel time across the five replications for the given combination of factors. A value of R close to one indicates that the travel time estimation method (when compared to results from multiple simulation runs) is sufficiently robust to use in testing various emergency response scenarios.

The travel time estimates are computed using the average congestion over the past 5 min before EMS departure at 1-min intervals. First, the simulation runs to the 45-min mark, allowing warm-up time to disperse traffic throughout the network. At the 45-min mark, the crash occurs and the EMS responders are notified 1 min later about this crash. Next, a stochastically generated chute time (time between notification and EMS vehicle departure) simulates the time at which the EMS actually leaves from the fire department. Analyzing the 2008 Public-Release Research Data Set available from the National Emergency Medical Services Information Systems (NEMSIS, 2008) project provided a normal

distribution with a mean of 2.21min for the chute time of EMS vehicles. Finally, the EMS arrives at the scene of the crash by following the path provided by the routing algorithm.

4.5.1 'Insensitive' versus 'sensitive' EMS vehicle

For congestion, using different arrival rates based on time of day provide the network with different levels of congestion, with 'Low' being in the evening and generally around 5–10% congestion (measured on highways), and 'Medium' being at rush hour with 15–30% congestion (measured on highways). It is important to note that this simulates the highways with no bottlenecks that cause major congestion, which brought about the need for the 'High' level of congestion.

While experiments were run under the 'Low' and 'Medium' congestion scenarios, presented in the Supplementary Appendix, the 'High' congestion scenario yielded the most interesting results. The 'insensitive' EMS and 'sensitive' EMS vehicle were tested under the 'High' congestion scenario across all three weather conditions to look at the best and worst-case scenarios of the EMS. Supplementary Appendices C and D present the results of this experiment, with Figure 4.2 showing the origin locations (Depots A, B and C) and crash locations (Crashes 1, 2 and 3 at different road types).

Under 'High' congestion conditions, the 'sensitive' EMS experiences significant increases in average travel time over all weather conditions. Average travel times were approximately three times higher under 'Clear' and 'Rain' conditions, and three-and-a-half times higher in 'Snow'. There are also higher variances of travel time associated with the 'sensitive' EMS when compared with the 'insensitive' EMS vehicle. While some routes experience much higher travel times, other routes experience only small increases in travel times if those routes were already the 'best' routes. This means these routes were not affected by the increased levels of congestion achieved by altering the free-flow speeds at the flags in Figure 4.2 to create a higher level of congestion.

The Dijkstra-based technique used estimated travel times based on an 'insensitive' EMS assumption to determine these routes. Assuming the EMS was more 'sensitive' to congestion (α and β were 30 and 5, respectively), the solution to the Dijkstra-based technique would have been a better route. As an example, consider the route from Depot A to Crash 1. The average travel time from the simulation using the 'sensitive' EMS was 10.15 min using the route optimized for the 'insensitive' EMS. Using a 'sensitive' EMS model to generate estimated travel times for each road segment, the shortest time path would have resulted in an

alternate path with an estimated travel time of 5.66 min. This is true for all of the routes that experienced a much higher 'sensitive' EMS average travel time compared with the 'insensitive' EMS travel time.

If the estimation model correctly represents the effects of congestion on an EMS vehicle's travel time, then it is possible to estimate accurate travel times based on real-time data on congestion and weather. However, as shown in this section, if the estimation model is significantly different from the actual capabilities of the EMS vehicle (ie manoeuvring through traffic, using shoulders, etc), then the performance of the estimation technique can be very poor and unable to provide the shortest time route. To study EMS vehicle travel in congested roadways, recording both the EMS vehicle travel and the amount of congestion on its route will aid in creating better estimated travel time models in the future.

4.5.2 Distinct weather regions

Supplementary Appendix E presents the results of the first set of additional testing: splitting up the area into three distinct weather regions and testing the response to Crash '1' from the different EMS starting locations. This experiment uses the 'insensitive' EMS model, only looking for impacts of weather in 'Medium' congestion. The goal of this testing was to determine if the shortest time path methodology using the estimated travel times could still perform well under varying weather conditions across the region. Crash '1' was chosen as it was near the middle of the regions (shown in Figure 4.2), causing the EMS routes to cross into different regions when responding. Overall, the best choice of which EMS should be sent is consistently correct when comparing the estimated travel times to the average simulated travel times.

The shortest time path methodology demonstrates its ability to determine different routes and starting depot location that may be better given the weather conditions in the three regions. This occurs when sending the EMS from the Depot B, with 'Weather Region 1' having the weather that affects the travel speed the greatest (the 'Rain, Clear, Clear' and 'Snow, Rain, Rain' conditions). The shortest time path attempts to 'avoid' the bad weather condition by sending it on an alternate path. The results of this experiment show the importance of knowing different weather patterns in the area as they may affect the best route for an EMS vehicle to take to an emergency scene. In future advanced transportation systems, data on roads that have already been cleared could also be of great use in routing emergency vehicles, providing responders with safer routes to the crash location.

4.5.3 Scenario testing: Event-based congestion

Event-based congestion, such as from a crash, demonstrates this method's ability to: (1) provide EMS responders with alternate routes given current levels of event-based congestion and (2) provide dispatchers with the best depot location that the EMS should be sent from when additional responders are necessary. The 'sensitive' EMS is used in order to better study the possible impacts of event-based congestion on emergency response routing. The Crash 1 site in Figure 4.2 provides a three-lane highway to test EMS response at different departure times after the accident has occurred and under two different crash severities. Depots A and B were used in this experiment since the shortest path never resulted from Depot C under the test conditions.

The first test is more minor than the second, allowing lanes to clear more quickly. The 2- and 5-min departure times were used to simulate the first responders under different possible notification and chute times and the 10-min was used to simulate a second responder travelling to the crash. This may occur after receiving more information about the severity of the accident and victims involved. Table 4.1 outlines the conditions for the number of lanes closed and the duration of closure for two test cases.

The results of these experiments are presented in Supplementary Appendix F. For Test 1 and EMS Depot A, the estimated travel times increase by 12.5% from the 2- to 5-min departure times and by 13% from the 2- to 10-min departure times. When running this in the simulation, however, the average travel times increase by 10% between the 2- and 5-min departures and actually decrease by 2% between the 2- and 10-min departure times. This is caused by the 5- and 10-min departure times allowing some of the congestion to clear by the time the EMS responder actually gets on to the highway as well as the variance in travel time caused by traffic lights. The 10-min departure time may have been better for the EMS as it allowed it to avoid more red traffic signals along the route.

Table 4.1 Conditions for event-based congestion testing

Conditions	**Test 1 (min)**	**Test 2 (min)**
Three lanes closed	1	2
Two lanes closed	2	5
One lane closed	30	30

For EMS Depot B, there was a 10.8% increase in estimated travel times from the 2- to 5-min departures and a 12.8% increase from the 2- to 10-min departures. The average travel times from simulation showed only a 6% increase between the 2- and 5-min departures and an 8% increase between the 2- and 10-min departures. Departing from Depot B did not experience the same dip in average travel time after the 10-min mark as the traffic lights played a larger role in causing longer travel times. For the maximum travel time values between the 5- and 10-min departures there is a 2.5% decrease, which may be a better indication that less congestion was present after the 10-min mark on the highway.

For Test 2, there is a more dramatic increase in estimated travel times, with a 13.5% increase between the 2- and 5-min departures and a 36% increase between the 2- and 10-min departures for Depot A. The simulation results do not reflect the increase between the 2- and 5-min departures, showing a small decrease in average travel times. Between the 2- and 10-min departures, the average travel time from the simulation increased by 13.6%. For Depot B, the route of the EMS actually changes between the 2- min departure and the 5- and 10-min departure. The increase in estimated travel times were 13.2% between the 2- and 5-min departures and 17.2% between the 2-and 10-min departures.

Consider the case when the EMS departs 10 min after the accident occurs. In addition to the route being different to the situation where the EMS is sent from Depot B, it also shows the point at which a dispatcher would send the EMS from Depot B rather than Depot A. This demonstrates the two abilities this technique has in reacting to event-based congestion, namely route determination and dispatch response determination. The results of these experiments also bring about some important issues for future research, as the values of R from Supplementary Appendix F are worse than those found in previous testings. This methodology uses the average congestion from the 5 minutes prior to the EMS departing, and thus there is no way to capture the increase or decrease in congestion that may begin to occur as it travels to the crash. This may result in further back up due to lane closures or dispersion of traffic due to the opening of a lane. Additionally, the traffic signals play a big role in how much the EMS vehicles are delayed while travelling on their route.

4.6 Variance-Constrained Shortest Path problem

As an extension to the shortest time route methodology, the VCSP methodology routes EMS vehicles around the traffic on a risk-averse path in the shortest time possible. This means the EMS vehicle will

be less likely to encounter possible significant delays on its route. This will ensure that the EMS vehicle is not sent on a route in which a large amount of time variability may cause the EMS to arrive much later than anticipated. The VCSP problem takes the form of a binary integer programme that minimizes the sum of costs (ie travel time) of a path while constraining the sum of variance v and covariance c to a provided value λ.

Supplementary Appendix G contains the method of solving the VCSP problem. This includes the solution of a Constrained Shortest Path (CSP) problem, additional steps beyond the CSP solution method required to solve the VCSP, and an example using our modelled network. These have been programmed in Java using the Eclipse Software Development Kit. Explanations described by Sivakumar and Batta (1994) as well as Handler and Zang (1980) are used throughout this section.

Measures were taken to limit or reduce the number of links in the transportation network to ensure that the VCSP problem would quickly be able to determine a solution. Due to the nature of the solution methodology, as the number of links in a solution increases, the number of iterations within Yen's (1971) *k*-shortest-path to solve Dijkstra's algorithm (1959) also increases. This resulted in an additional loss of fidelity as travel times and variances associated with intersections are tied to the following link, whereas in the simulation the intersections include both a small approach and departure links and the intersection itself.

This section presents the testing and results used to study the VCSP solution methodology and performance. This includes a description of the experiment, the different parameters (such as λ) as well as different conditions of the transportation network tested. The findings and observations formed from the results of this testing are then presented with discussion on the VCSP. The envisioned best use for this technique is in the future as countries move towards advanced transportation systems.

4.6.1 Experiment design

To test the performance of the VCSP problem solution, a 'risk-averse' value of λ, the constraint on the sum of variance and covariance, is obtained by first finding the unconstrained shortest time path and determining its level of variance and covariance. Tightening this constraint by lowering λ will then allow conclusions to be drawn on the trade-off of higher expected travel time *versus* risk-aversion. This is accomplished by running the different routes, resulting from different λ values, through the simulation to determine maximum travel time experienced by the unconstrained shortest path and the minimum

travel time experienced by the 'risk-averse' path across 10 replications. These are shown in the results portion of this section.

Similar to the earlier routing methodology described in the Dijkstra-based shortest path approach, information from FHWA (2011a) and Hranac *et al* (2006) on the effects of different weather conditions on travel speeds is used to simulate the slowdown traffic may experience due to weather. Rather than maintaining static weather over the entire simulation run, weather is programmed to be dynamic by changing the percent reduction in travel speed that traffic in the network experiences (eg 'Rain' caused a 15% reduction in travel speed in the Dijkstra-based approach). Modelling weather as such serves two purposes: first, it is more realistic as weather trends move and vary with respect to time; second, it allows us to study the effects of variation in weather patterns on travel times with respect to the VCSP problem. While running the simulation under 'Rain' and 'Snow' conditions, the weather will start at the 30-min mark and increase in strength every 2 min until coming to full strength 20 min later (15% slowdown for 'Rain', 40% for 'Snow'). This means every 2 min there will be a 0.75% increase in 'Rain' conditions and a 2% increase in 'Snow' conditions. The 'High' congestion condition is used along with the 'sensitive' EMS vehicle to test the VCSP approach.

4.6.2 Results and discussion

While there are other fire departments in the area that could respond to the crash, Depot C in Figure 4.2 is used to study the VCSP solutions as it allows for more path possibilities. To calculate the Variance-Covariance Matrix, data from the previous 20 min before the accident occurred provide 20 data sets at 1-min intervals that are used to determine the expected travel time along with the variance and covariance of these travel times. The unconstrained shortest path from the Dijkstra-based approach and the VCSP approach results from the simulation's first replication data set. These paths are replicated 10 times, where the movement (arrival time into the system, vehicle turning, etc) of other vehicles and the chute times vary between replications. This causes the EMS vehicle to experience different traffic signals, congestion and weather on the path.

4.6.2.1 'Clear' weather scenario results

Figure 4.3 presents the unconstrained shortest path (a) and the VCSP (b), as well as the starting location of the EMS vehicle (Depot 3) and the crash location. The first experiment run was under 'Clear' weather

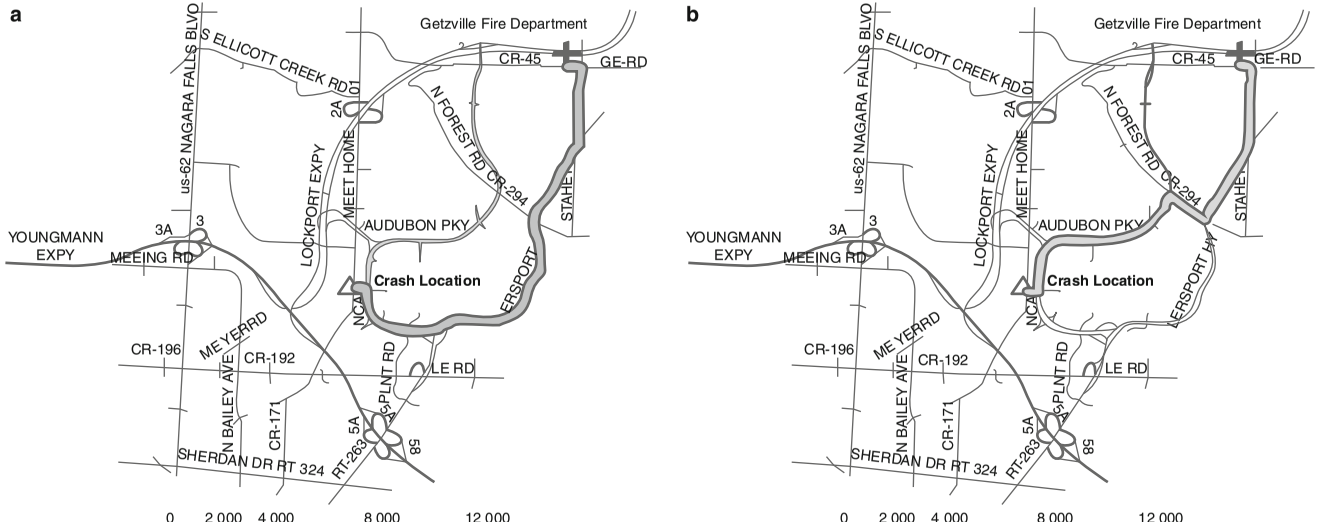


Figure 4.3 (a) Unconstrained shortest path (b) and VCSP

conditions, where no slowdown effects were experienced by the EMS vehicle. The average travel time for the unconstrained shortest path was 262 s, with a maximum travel time of 286 s. The average travel time for the VCSP path was 298 s, with a minimum travel time of 275 s.

After 20 replications, the VCSP actually had a slightly higher variance of 219 as opposed to 202 for the unconstrained shortest path. This is a result of multiple factors, such as the chute times varying between replications in addition to the varying individual link congestion and intersection signals, and the reaction of traffic to the EMS vehicle as it travels on the links. This causes the EMS vehicle to experience a different environment between replications of the experiment. A slightly higher calculated variance for the unconstrained shortest path results from use of data collected from the simulation's first replication along with assumptions used for signalized intersections.

4.6.2.2 'Rain' and 'Snow' weather scenario results

The paths resulting from data collected during a 'Rain' weather scenario are the same as shown in Figure 4.3. The average travel times were 309 and 359 s for the unconstrained shortest path and the VCSP, respectively. The maximum travel time for the unconstrained shortest path was 328 s while the minimum travel time for the VCSP was 332 s. This means that, in reality, the choice to follow the VCSP would not be the best to take even in a 'risk-averse' scenario, since the unconstrained shortest path would still perform better than the VCSP. In this scenario, the VCSP does have a smaller variance at 224, as opposed to 297 for the unconstrained shortest path, and thus the data collected allowed the VCSP to actually determine the 'risk-averse' path.

As with the 'Rain' scenario, the 'Snow' scenario also resulted in the paths shown in Figure 4.3 and had similar results when tested with the simulation model. The average travel times were 347 and 419 s for the unconstrained shortest path and the VCSP, respectively. The maximum travel time was 368 s for the unconstrained shortest path and the minimum travel time was 398 s for the VCSP, showing even more so than the 'Rain' scenario that the 'risk-averse' path would not be the best path to take. As with the 'Rain' scenario results, the 'Snow' scenario does, however, result in the VCSP having lower variance of 170 as opposed to the variance of 196 for the unconstrained shortest path.

4.6.2.3 Discussion of results

Table 4.2 summarizes the results for all three scenarios studied in the experimental analysis. The unconstrained shortest path (Figure 4.3(a))

Table 4.2 Summary of experimental scenario results

Scenario	Unconstrained shortest path			Risk-averse VCSP		
	Average	Maximum	Variance	Average	Minimum	Variance
Clear	262	286	202	298	275	219
Rain	309	328	297	359	332	224
Snow	347	368	196	419	398	170

is approximately 3.1 miles and travels through seven signalized intersections, and the risk-averse VCSP (Figure 4.3(b)) is approximately 3.4 miles and travels through six signalized intersections. The maximum travel times are shown for the unconstrained shortest path, while the minimum travel times are shown for the risk-averse VCSP. This demonstrates that over the simulated replications of the experiments, risk-aversion leads to routes that are at best still longer than the longest time experienced in the unconstrained shortest path replications.

The results of these experiments suggest that the variance experienced by an EMS vehicle may not be as easily categorized and accounted for by collecting data over a period of 20 min before an accident. Additionally, a better way of accounting for travel times through intersections is necessary as assuming an expected travel time and variance based on the signal timing and respective green and red signal travel times do not fully represent reality. This is because the travel time through one intersection is dependent upon what the EMS vehicle experienced at previous intersections and the estimated variance for each intersection is assumed to be independent of each other; a vehicle can generally travel through multiple green signals on a road before experiencing a red signal due to optimized signal timing.

In the 'Clear' weather scenario, the VCSP actually had greater variance than the unconstrained shortest path, a longer travel distance and fewer signalized intersections that it had to travel through. One possible cause of the greater variance in the 'Clear' scenario is due to the longer distances between the intersections in the VCSP solution, causing the EMS vehicle to encounter different states of the traffic signals across the replications. In the 'Rain' and 'Snow' scenarios, the EMS vehicle may have been slowed down enough over the path that it encountered nearly the same signal pattern over the multiple replications. Finally, the variance in congestion of regular traffic over the replications would also directly influence the variance in the EMS vehicle's travel time over the replications. Again, the 'Rain' and 'Snow' scenarios may have actually

provided an effect that caused the variance of congestion to be more constant between replications by slowing down all traffic and causing congestion to become steady.

4.7 Concluding remarks

The VCSP is a more sophisticated model that utilizes more data, but fails to deliver high-quality solutions. This is due to the unstable behaviour of the model that may yield solutions that are too risk-prone or too risk-averse due to the value of the parameter. The run time for the VCSP is also observed to be significantly longer given that it is an NP-hard problem. As shown with Google, incorporating both real-time data and historical data has provided better travel time estimates when showing the 'travel time in current conditions' (Dillow, 2012). Combining this with the VCSP approach, adding a confidence interval on the travel times would further improve the methods rather than either determining explicit 'optimal' paths or travel times. Further data fusion methods must be developed to present only necessary and actionable information to EMS responders, given the fast decision-making environment in which they operate.

The event-based congestion experimentation shows the need for a possible method of predicting future travel times and using them to route EMS vehicles. Doing so may provide a better route as it could reflect increasing levels of congestion resulting from a crash over time. This may be difficult given the short travel times of EMS vehicles to the scene of the crash and the need for real-time routing decisions. Trips from the scene of the crash to a hospital may benefit from this approach to a greater extent as they generally have longer travel times and can be rerouted over time if an event occurs on the current path they are travelling.

Predicting future levels of congestion as well as future timing of traffic lights may aid in determining better routes for responders. Expected travel times would be needed for each road segment and time interval. If a crash occurs on road segment r_c at time interval t_c, then for each (r, t), where t belongs to some set of time intervals after the crash and r belongs to the set of road segments in the network, an expected travel time $(E(r, t))$ must be computed. Using these expected travel times, a shortest path model that accounts for when the EMS will reach each road segment may then be developed to create the shortest path from the EMS's current location to the road segment where the crash is located (r_c). As with the methods presented in this paper, as the network becomes larger

adaptations are needed to maintain real-time solution capabilities such as Jotshi *et al* (2009) and Chou *et al* (1998) that use a hierarchy of networks to break apart the problem.

Acknowledgements

We thank the two anonymous reviewers who provided us with excellent advice and guidance in order to strengthen and improve our paper. This material is based upon work supported by the Federal Highway Administration under Cooperative Agreement No. DTFH61-07-H-00023, awarded to the Center for Transportation Injury Research, CUBRC, Inc., Buffalo, NY. Any opinions, findings and conclusions are those of the author(s) and do not necessarily reflect the view of the Federal Highway Administration.

References

Chou Y, Romeijn HE and Smith RL (1998). Approximating shortest paths in large-scale networks with an application to intelligent transportation systems. *INFORMS Journal on Computing* **43**(1): 163–179.

Dijkstra EW (1959). A note on two problems in connexion with graphs. *Numerische Mathematic* **1**(1): 269–271.

Dillow C (2012). Google maps re-launches its traffic mapping, incorporating real-time data from drivers with phones. *Popular Science*, http://www.popsci.com/cars/article/2012-04/google-maps-reboots-its-estimated-transit-time-estimates-incorporating-realtime-traffic-data, accessed 5 July 2012.

FHWA (2011a). *How Do Weather Events Impact Roads*? US Department of Transportation. Online. http://ops.fhwa.dot.gov/weather/ q1_roadimpact. htm, accessed 5 December 2011.

FHWA (2011b). *Traffic Analysis Toolbox Volume IV: Guidelines for Applying CORSIM Microsimulation Modeling Software: Appendix N*. US Department of Transportation. Online. http://ops.fhwa.dot.gov/ trafficanalysistools/tat_vol4/index.htm, accessed 5 December 2011.

Gorjestani A, Menon A, Cheng P, Shankwitz C and Donath M (2010). *The design of a minimal sensor configuration for a cooperative intersection collision avoidance system—Stop sign assist*. CICAS-SSA Report # 2. US Department of Transportation, Federal Highway Administration.

Handler GY and Zang I (1980). A dual algorithm for the constrained shortest path problem. *Networks* **10**(4): 293–309.

Henchey MJ, Batta R, Blatt A, Flanigan M and Majka K (2013). A simulation approach to study emergency response. *Journal of Simulation*, advance online publication 18 October, doi:10.1057/ jos.2013.20.

Hranac R, Sterzin E, Krechmer D, Hesham R and Farzaneh M (2006). *Empirical Studies on Traffic Flow in Inclement Weather*. US Department of Transportation, Federal Highway Administration: USA.

Huang H (2012). *Real-Time Information and Correlations for Optimal Routing in Stochastic Networks*. Open Access Dissertations. Paper 500. Online. http://

scholarworks.umass.edu/open_access_ dissertations/500, accessed 1 August 2012.

IAFC (2011). *Guide to IAFC Model Policies and Procedures for Emergency Vehicle Safety*. Online. https://www.iafc.org/files/1SAFEhealthSHS/VehclSafety_IAFCpolAndProceds.pdf, accessed 20 February 2011.

ITS UW (2012). *Real-Time Freeway Sensor Information over the Internet*. Online. http://www.its.washington.edu/software/loop_cli.html, accessed 5 July 2012.

Jotshi A, Gong Q and Batta R (2009). Dispatching and routing of emergency vehicles in disaster mitigation using data fusion. *Socio-Economic Planning Sciences* **43**(1): 1–24.

National Emergency Medical Service Information System (2008). *NEMSIS 2008 Database*, http://www.nemsis.org, accessed 18 November 2012.

Niagara International Transportation Technology Coalition (2012). *NITTEC Frequently Asked Questions*. Online. http://www.nittec.org/ about_us/faqs/index.html, accessed 1 August 2012.

Nie YM, Wu X, Dillenburg JF and Nelson PC (2012). Reliable route guidance: A case study from Chicago. *Transportation Research Part A: Policy and Practice* **46**(2): 403–419.

OHPI (2004). *Traffic data collection equipment for high-volume locations*. Traffic data for high volume routes: Best practices and guidelines final report. US Department of Transportation, Federal Highway Administration.

RITA (2011a). *Cooperative Intersection Collision Avoidance Systems (CICAS)*. US Department of Transportation. Online. http://www.its.dot.gov/cicas/index.htm, accessed 5 December 2011.

RITA (2011b). *About ITS*. Online. http://www.its.dot.gov/faqs.htm, accessed 23 October 2011.

Sivakumar RA and Batta R (1994). The variance-constrained shortest path problem. *Transportation Science* **28**(4): 309–316.

Yen JY (1971). Finding the K shortest loopless paths in a network. *Management Science* **17**(11): 712–716.

Appendices

A Nomenclature

T_c = congested link travel time (seconds)
T_f = free-flow link travel time (seconds)
v = link traffic volume (vehicles)
c = link capacity (vehicles)
α, β = volume/delay coefficients
S_c^r = regular vehicle congested link travel speed (feet/hour)
S_c^e = emergency vehicle congested link travel speed (feet/hour)
S_f = free-flow link travel speed (feet/hour)
u = utilization of road segment in Arena (percent)
w = weather slowdown effect (percent)
e = EMS vehicle slowdown effect (percent)
X = the set of candidate nodes
$\bar{X}$ = the set of all nodes not yet optimized
L = the set of all links in the network
l_j = travel time or cost from the origin to node j
l_{ij} = travel time or cost from node i to node j
p_j = predecessor of node j
R = performance measure of estimated travel time technique
A_e = Estimated EMS travel time (minutes)
A_s = Average EMS travel time from five simulation replications (minutes)
Min_s = Minimum EMS travel time from five simulation replications (minutes)
Max_s = Maximum EMS travel time from five simulation replications (minutes)
Var_s = Variance of EMS travel time from five simulation replications (minutes^2)
VSCP = Variance-Constrained Shortest Path
x_{ij} = decision variable in the VCSP problem (1 if link i to j is used, 0 otherwise)
λ = limiting value for the constraints in the VCSP problem
μ = dual variable of the VCSP used in solving the dual problem
v_{ij} = variance of link i to j in the VCSP problem
c_{ijkl} = covariance between links i to j and k to l

B Modified Dijkstra's Algorithm Pseudocode

This section presents the pseudocode for the modified Dijkstras algorithm employed in the shortest path calculations. This uses a graph $V = \{X, A\}$, where X is the set of nodes, with $0 \in X$ being the origin node and A is the set of arcs. This involves an initialization step, which sets the current shortest time to infinity for all nodes except the origin node that has a distance of zero. The main step then begins by finding candidates accessible from the origin, setting the appropriate shortest travel times and adding the best shortest travel time to the candidate

list. This repeats until the method reaches the destination node, in this study the location of the crash. The pseudocode also shows how the predecessors are tracked using the term p_j.

Algorithm 1 Initialize $\bar{X}$ = all nodes except origin node 0

```
for all node j in X̄ do [Cycles through all non-origin nodes]
l_j = ∞ [Assigns the travel time from the origin to j to infinity]
P_j = null [Assigns the predecessor as null (undefined) for j]
end for
t_0 = 0 [Assigns travel time at the origin to zero]
i = node 0 [Assigns node i to the origin node]
```

Algorithm 2 Main Step

```
while X̄ contains nodes and X does not contain the destination node do
bestQ = null [Initializes the best node to null, used to determine the next node
to consider]
bestT = ∞ [Initializes the travel time associated with the node above to infinity]
  for all node j in X̄ do
    for all link a in A do
      if node i connects to node j through link a then
        if l_i + l_ij < l_j then
        l_j = l_i + l_ij [Assigns the new travel time to node j]
        p_j = i [Assigns node i as predecessor to node j]
        end if
      end if
    end for
  end for
Remove node i from X̄
Add node i to X̂
  for all node j in X̄ do [Determines the shortest travel time of nodes in X̄]
    if l_j < bestT then
    bestT = l_j
    bestQ = j
    end if
  end for
i = bestQ [Assigns next candidate node i as the one with the shortest travel time]
end while
```

C "Insensitive" EMS in High Congestion

"Insensitive" EMS vehicle, α and β are 0.83 and 5.5, respectively (original BPR).

Table 4.3 "Insensitive" EMS in high congestion experimental results

Weather Effects	Crash Scene	EMS Depot	A_e	A_s	Min_s	Max_s	Var_s	R
Clear	1	A	4.00	3.71	3.39	4.00	0.05	1.08
	1	B	3.93	4.01	3.88	4.09	0.01	0.98
	1	C	3.94	3.93	3.87	4.14	0.01	1.00
	2	A	2.23	2.07	1.89	2.36	0.05	1.07
	2	B	2.73	2.61	2.29	2.90	0.05	1.05
	2	C	5.45	5.56	5.48	5.78	0.01	0.98
	3	A	3.22	3.14	2.82	3.29	0.03	1.03
	3	B	3.13	3.02	2.83	3.24	0.03	1.03
	3	C	4.71	4.45	4.08	4.93	0.08	1.06
Rain	1	A	5.25	5.08	4.73	5.28	0.05	1.03
	1	B	5.64	5.76	5.62	5.85	0.01	0.98
	1	C	5.30	5.20	5.14	5.42	0.01	1.02
	2	A	2.83	2.45	2.18	2.69	0.05	1.15
	2	B	3.74	3.46	3.41	3.65	0.01	1.08
	2	C	7.40	7.55	7.42	7.73	0.02	0.98
	3	A	4.57	4.51	4.08	4.88	0.07	1.01
	3	B	4.53	4.36	4.05	4.62	0.03	1.04
	3	C	6.39	6.04	5.82	6.29	0.02	1.06
Snow	1	A	7.33	6.75	6.33	7.37	0.18	1.09
	1	B	7.97	8.19	7.90	8.30	0.02	0.97
	1	C	7.48	7.32	7.02	7.42	0.02	1.02
	2	A	4.00	3.81	3.10	4.18	0.15	1.05
	2	B	5.31	4.94	4.46	5.15	0.08	1.07
	2	C	10.43	10.50	10.24	10.71	0.04	0.99
	3	A	6.47	6.12	5.75	6.88	0.22	1.06
	3	B	6.43	6.14	5.82	6.50	0.05	1.05
	3	C	9.05	8.57	8.21	9.03	0.09	1.06

D "Sensitive" EMS in High Congestion

"Sensitive" EMS vehicle, α and β are 30 and 5, respectively (as with regular vehicles in simulation). Routes and conditions are the same as in the "Insensitive" EMS case for comparison.

Table 4.4 "Sensitive" EMS in high congestion experimental results

Weather Effects	Crash Scene	EMS Depot	A_e	As	Min_s	Max_s	Var_s	R
Clear	1	A*	4.00	10.15	9.72	10.63	0.10	0.39
	1	B	3.93	5.73	5.24	6.34	0.19	0.69
	1	C	3.94	5.81	5.24	6.11	0.09	0.68

(*continued*)

Table 4.4 Continued

Weather Effects	Crash Scene	EMS Depot	A_e	As	Min_s	Max_s	Var_s	R
	2	A	2.23	2.52	2.21	2.80	0.04	0.88
	2	B*	2.73	18.90	18.55	19.79	0.21	0.14
	2	C*	5.45	23.02	22.10	23.93	0.35	0.24
	3	A	3.22	3.14	2.82	3.29	0.03	1.03
	3	B*	3.13	19.27	18.28	20.25	0.44	0.16
	3	C	4.71	4.45	4.08	4.93	0.08	1.06
Rain	1	A*	5.25	12.93	12.56	13.50	0.12	0.41
	1	B	5.64	8.69	8.06	9.42	0.24	0.65
	1	C	5.30	7.93	7.50	8.20	0.09	0.67
	2	A	2.83	3.29	2.72	3.63	0.09	0.86
	2	B*	3.74	26.97	25.30	27.62	0.76	0.14
	2	C*	7.40	33.08	32.34	34.12	0.49	0.22
	3	A	4.57	4.51	4.08	4.88	0.07	1.01
	3	B*	4.53	27.36	25.53	28.46	1.19	0.17
	3	C	6.39	6.00	5.82	6.23	0.03	1.07
Snow	1	A*	7.33	19.05	17.84	20.66	1.20	0.38
	1	B	7.97	12.01	10.58	13.87	1.27	0.66
	1	C	7.48	10.71	9.85	11.94	0.50	0.70
	2	A	4.00	4.94	4.01	6.02	0.65	0.81
	2	B*	5.31	37.79	35.16	39.34	2.20	0.14
	2	C*	10.43	47.20	45.89	48.15	0.54	0.22
	3	A	6.47	6.12	5.75	6.88	0.22	1.06
	3	B*	6.43	39.97	39.61	40.42	0.08	0.16
	3	C	9.05	8.45	8.21	8.72	0.05	1.07

*Indicates a route that would have changed if it was known to be "sensitive"

E Split Weather In Medium Congestion

These tests use Crash 1 (Figure 4.2 of the main paper), the three-lane crash, to gather results of sending the emergency vehicle from each of the three depot locations in varying weather conditions across the regions. Note: bold face indicates the best depot location from which the EMS vehicle should be sent based on estimated and simulated travel times.

Table 4.5 Results of three distinct regions of weather

Region 1	Region 2	Region 3	EMS Depot	A_e	A_s	Min_s	Max_s	Var_s	R
			A	4.84	4.56	4.27	4.83	0.05	1.06
Rain	Clear	Clear	B	4.18	4.00	3.68	4.18	0.03	1.04
			C	4.19	4.11	4.06	4.31	0.01	1.02
			A	3.87	3.69	3.39	3.90	0.04	1.05

(*continued*)

Table 4.5 Continued

Region 1	Region 2	Region 3	EMS Depot	A_e	A_s	Min_s	Max_s	Var_s	R
Clear	Rain	Clear	B	3.95	3.70	3.34	3.99	0.05	1.07
			C	4.38	4.12	4.06	4.35	0.01	1.06
			A	4.10	3.93	3.62	4.13	0.05	1.04
Clear	Clear	Rain	B	4.96	4.77	4.48	5.19	0.07	1.04
			C	4.52	4.49	4.43	4.68	0.01	1.01
			A	6.89	6.49	5.80	7.22	0.24	1.06
Snow	Rain	Rain	B	5.97	5.88	5.61	6.51	0.11	1.02
			C	5.68	5.47	5.40	5.71	0.01	1.04
			A	5.07	4.79	4.55	5.05	0.05	1.06
Rain	Snow	Rain	B	5.33	5.20	4.88	5.83	0.12	1.02
			C	6.15	5.57	5.41	5.82	0.04	1.10
			A	5.39	5.01	4.81	5.37	0.04	1.08
Rain	Rain	Snow	B	6.93	6.69	6.25	7.21	0.11	1.04
			C	6.16	6.02	5.95	6.25	0.01	1.02

F Event Based Congestion Results

The results presented in Table 4.6 are from the event based congestion experiments at Crash 1. Test 1 is the condition of a crash closing down three lanes for a duration of 1 minute, the two right lanes for a duration of 2 minutes and right-most lane for a duration of 30 minutes. Test 2 is the condition of a crash closing down three lanes for a duration of 2 minutes, the two right lanes for a duration of 5 minutes and right-most lane for a duration of 30 minutes. The emergency vehicle then leaves 2 minutes, 5 minutes and 10 minutes after the accident occurs (Departure Time), from Depot A or B.

Table 4.6 Event based congestion results of the two test configurations

Test Conditions	Departure Time	EMS Depot	A_e	A_s	Min_s	Max_s	Var_s	R
Test 1	2 minutes	A	3.99	3.71	3.39	4.87	0.34	1.08
		B	4.07	3.64	3.54	4.02	0.04	1.12
	5 minutes	A	4.43	4.08	3.65	4.50	0.13	1.08
		B	4.51	3.87	3.74	4.35	0.06	1.16
	10 minutes	A	4.51	3.56	3.39	3.60	0.01	1.27
		B	4.59	3.93	3.74	4.24	0.05	1.17
Test 2	2 minutes	A	4.00	3.90	3.70	4.32	0.05	1.03
		B	4.08	4.32	4.03	4.57	0.05	0.95
	5 minutes	A	4.54	3.83	3.65	4.11	0.03	1.19
		B	4.62	4.37	4.17	4.47	0.01	1.06
	10 minutes	A	5.43	4.43	3.59	5.29	0.33	1.22
		B	4.78	4.39	4.27	4.47	0.01	1.09

G VCSP Solution Methodology

The following binary integer program demonstrates the VCSP problem, minimizing the sum of the costs l in (1) of the path from node 1 to N while constraining the sum of the variance υ and covariance c to a provided value of λ in (2). This uses a graph $\mathbf{V} = \{\mathbf{X}, \mathbf{A}\}$, where $\mathbf{X}$ is the set of nodes, with $1 \in X$ being the origin node and $\mathbf{A}$ is the set of arcs. Constraints (3) and (4) ensure that the path starts at node 1 and ends at node N, with x_{ij} values being 0 or 1 (the decision variable on whether or not the arc is used). The specific enforcement of binary constraints on the x_{ij} variables becomes unnecessary due to the solution method using Yen's k-shortest-paths algorithm (1971) and Dijkstra's algorithm (1959) to generate solutions.

(VCSP) Minimize $$\sum_{(i,j)\in A} l_{ij}x_{ij} \tag{1}$$

Subject To

$$\sum_{(i,j)\in A} \upsilon_{ij}x_{ij} + \sum_{(i,j)\in A}\sum_{(k,l)\in A} c_{ijkl}x_{ij}x_{kl} - \lambda \le 0 \tag{2}$$

$$\sum_{j} x_{ij} - \sum_{k} x_{ki} = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{if } i = 2, \ldots, N-1, \\ -1, & \text{if } i = N \end{cases} \tag{3}$$

$$x_{ij} \in \{0,1\},\ (i,j) \in \mathbf{A} \tag{4}$$

Solving this problem requires use of a heuristic, where first $f(\mathbf{X})$ is set to be equivalent to (1) and $g(\mathbf{X})$ to the left hand side of (2), where $\mathbf{X}$ belongs to $\bar{\mathbf{X}}$, the set of nodes that satisfy constraints (3) and (4). Next, the VCSP is rewritten, where the objective function is $f^* = f(\mathbf{X}^*) = \min_{\mathbf{X}\in\bar{\mathbf{X}}} f(\mathbf{X})$ subject to $g(\mathbf{X}) \le 0$. Finally, the $g(\mathbf{X})$ constraint is relaxed, creating a Lagrangian relaxation function shown in (5) and a function for $L(\mu)$ shown in (6), where the dual variable $u \in \Re$. This makes it possible to formulate the dual VCSP (DVCSP) in (7) which can be solved via the algorithm shown in Handler and Zang's paper (1980) and provides a gap closing procedure when necessary to determine the optimal primal solution to the VCSP.

$$L(\mu, \mathbf{X}) = f(\mathbf{X}) + \mu g(\mathbf{X}) \tag{5}$$

$$L(\mu) = \min_{\mathbf{X}\in\bar{\mathbf{X}}} L(\mu, \mathbf{X}) \tag{6}$$

$$(DVCSP)\ L^* = L(\mu^*) = \max_{\mu\ge 0} L(\mu) \tag{7}$$

The VCSP is complicated by the non-linear constraint that replaces the linear constraint from the Constrained Shortest Path (CSP) with Equation (2) (the CSP did not have the $c_{ijkl}x_{ij}x_{kl}$ portion of the constraint). A series of manipulations and additional constraints are required to convert the non-linear constraint to

linear constraints which require introduction of additional dual variables, δ, γ and η making the relaxation function in terms of $L(\mu, \delta, \gamma, \eta)$. Because only a dual solution is needed for the gap-closing procedure to be efficient, (8) presents the necessary relaxation required to solve the dual problem, where the term $\in$ is defined in (9).

$$L(\mu,0,0,0) = \min_{X \in \bar{X}} \sum_{(i,j) \in A} (l_{ij} + \mu v_{ij}) x_{ij} + \mu(\in -\lambda) \tag{8}$$

$$\in = \sum_{(i,j) \in A} \sum_{(k,l) \in A} \min(c_{ijkl}, 0) \tag{9}$$

To close the gap a heuristic is required, as a direct solution of this problem would require large computation times if solved as a linear program. This is accomplished using Yen's *k*-shortest-path (1971) and Dijkstra's algorithm (1959), similar to the gap closing procedure of Handler and Zang's CSP problem (1980), but uses different equations to check for optimality and satisfaction of constraints. This heuristic generates first best, second best, third best, ..., solutions to (10) and uses the corresponding values of (11) to determine values of the lower bound. When the constraint on variance and covariance is satisfied, this creates an upper bound. While in most cases the lower bound becoming greater than the upper bound signifies optimality, it is necessary to ensure that optimality is not declared prematurely.

$$\min_{X \in \bar{X}} \sum_{(i,j) \in A} (l_{ij} + \mu v_{ij}) x_{ij} \tag{10}$$

$$\sum_{(i,j) \in A} (l_{ij} + \mu^{\star}\ v_{ij}) x_{ij} + \sum_{(i,j) \in A} \sum_{(k,l) \in A} \mu^{\star}\ c_{ijkl} x_{ij} x_{kl} \tag{11}$$

$$g = f + \mu^{\star} \sum_{(i,j) \in A} \sum_{(k,l) \in A} \min(c_{ijkl}, 0) \tag{12}$$

(12) provides the final check on optimality, where f is set to the value of (11) and the remaining solutions less than g must be examined. Sivakumar and Batta (1994) explain that this is true because solutions with values of (10) greater than or equal to g could not provide a feasible solution better than the current candidate optimal solution. This is because the most favorable scenario is to have a path which has link pairs with negative covariances only. The gap closing procedure then continues generating additional paths until a better feasible solution is found, at which time its value of (11) becomes f, and g is reevaluated. If a feasible solution is found that is greater than the value of g the process stops. An additional stopping criteria is that no feasible solutions are found after evaluation of all paths.

Please see Sivakumar and Batta's (1994) paper for a more exhaustive description of the methodology behind solving the VSCP problem. It is important to comment on their results with respect to solution time in a 50-node network. They showed that in some cases the solution methodology required more than

30 seconds to determine the solution on a VAX 11 / 780 from the early 1990s. While this is much better than the more than 200 seconds required by the *k* Shortest Path Based Exact Method, computation times for the solution methodology must be considered due to the nature of real-time solutions required by EMS responders. While the network used in this here is larger in terms of the number of links and nodes, given the advancement in technology in the nearly two decades since Sivakumar and Batta's paper, has made it possible to generate solutions in near real-time. On a machine with Windows 7 64-bit, an Intel Core i7-2600 CPU @ 3.40 GHz and 8 GB RAM, the Java program written to solve the VCSP took approximately 3 seconds to compute for an example where the starting and ending locations were placed at opposite corners of the transportation network. Further results of the computation time required to solve the VCSP for the example used throughout the main paper is presented in Section 6.

G.1 VCSP Solution Example

This section demonstrates an example solution to the VCSP problem, where the starting point of the EMS vehicle is the Getzville Fire Department (red cross) and the crash scene is at the intersection specified by an orange triangle in Figure 4.4. This example uses data captured using "Clear" weather conditions using traffic conditions reflecting an evening rush hour. The traffic flows into the area are calculated using average daily traffic counts and hourly percentages to determine average time between arrivals, described in Section 4 of the main paper.

The first stage of the solution method is Dual Optimization. In this stage, initial values for μ are set to zero and infinity, with solutions generated using a Dijkstra's algorithm with the objective function set to Equation (10) in Appendix G:

$$\min_{\mathbf{x}\in\bar{\mathbf{X}}} \sum_{(i,j)\in A} (l_{ij} + \mu v_{ij}) x_{ij} \tag{13}$$

The travel times are in seconds, with expected travel times and variances used for signalized intersections based on International Association of Fire Chiefs (IAFC) policy and signal timing (explained in Section 4.1 of the main paper). The value of λ was set to 200, which was selected based on determining the variance associated with the shortest unconstrained path and then setting λ to a greater value. The constraint values come from Equation (2) in Supplementary Appendix G:

$$\sum_{(i,j)\in A} v_{ij} x_{ij} + \sum_{(i,j)\in A} \sum_{(k,l)\in A} c_{ijkl} x_{ij} x_{kl} - \lambda \tag{14}$$

Figure 4.4 route (a) shows the shortest route with $\mu = 0$ that has a travel time of 238 seconds and an Equation (14) value of 24.3, meaning it did not satisfy the constraint as expected. Figure 4.4 route (b) shows the route solved with $\mu = \infty$ that has a travel time of 273 seconds and an Equation (14) value of –25, meaning it satisfies the constraint.

The second step of the heuristic (after setting $\mu = 0$ and ∞) is optimizing the values of μ by maximizing Equation (10) This is accomplished by using Dijkstra's algorithm to solve for Equation equ:PaperThreeCostFunctionm and them

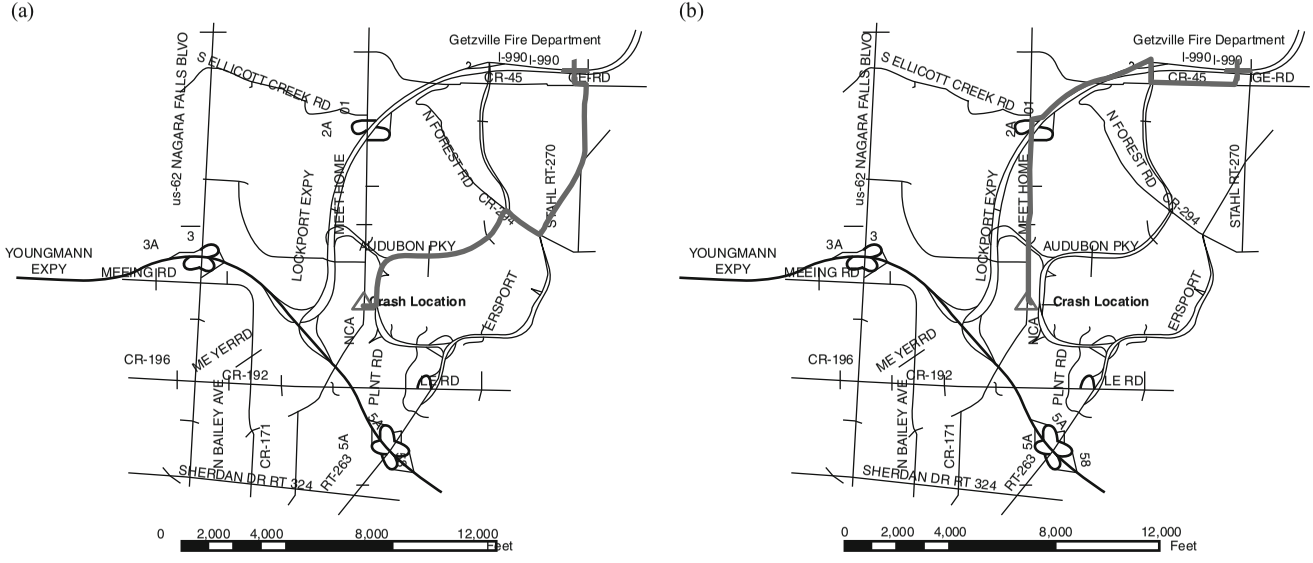


Figure 4.4 Routes from dual optimization at $\mu = 0$ (a) and $\mu = \infty$ (b)

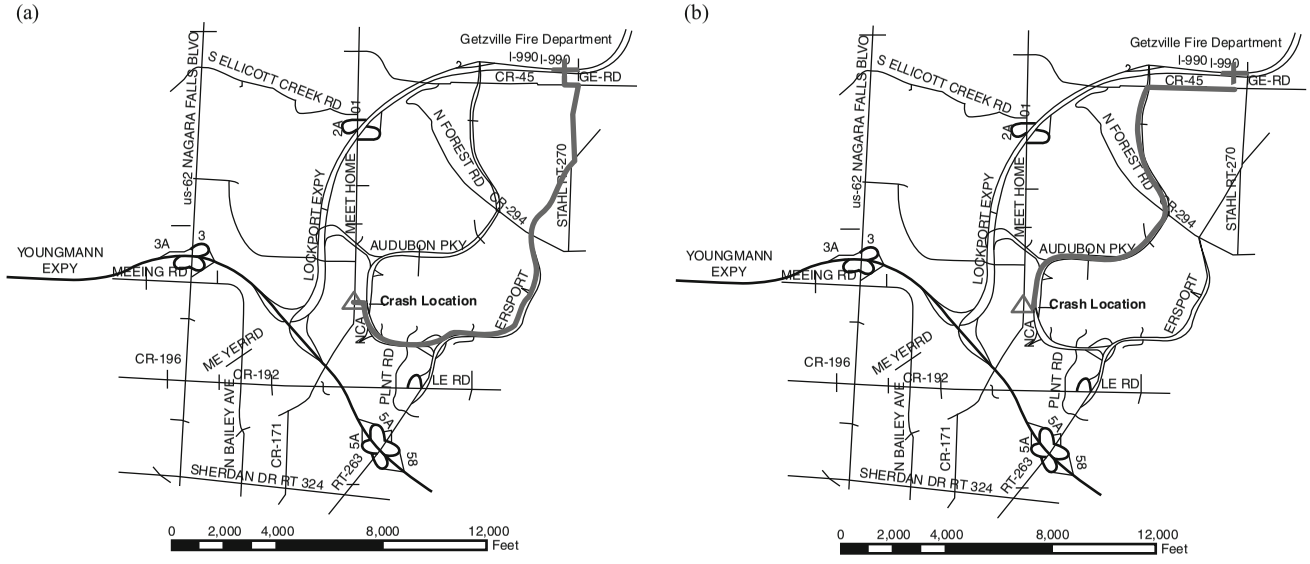


Figure 4.5 Optimal Route from $\mu^* = 0.43$ (c) and k-shortest route procedure (d)

maximizing equ:PaperThreeCostFunction by changing μ. Doing this iteratively then determines the optimal value of μ. The first iteration of this process for the example results in $\mu = 0.72$ with route (c) in Figure 4.5 which has an estimated travel time of 254 seconds and an Equation (14) value of –12.8. The second iteration results in $\mu = 0.43$ (again with route (c)) becoming the optimal value of μ. After this stage, the lower bound is 248 (from Equation (11))and the upper bound is 254. The gap closing procedure finds the route shown in Figure 4.5 route (d), which has a dual solution (Equation equ:PaperThreeLowerBoundFun ction)value of 259, a travel time of 247 seconds and an Equation (14) value of 28.4, meaning it did not satisfy the constraint. This sets the lower bound to 259, which is greater than the upper bound of 254, making the route from Figure 4.5 route (c) the optimal route.

While generating additional paths to ensure optimality (the final stage of the solution methodology) the next shortest path had a travel time of 262 seconds with an Equation (14) value of 62.9. The next feasible solution (Equation (14) ≤ 0) would have been the solution in Figure 4.4 route (b) with a travel time of 273 seconds and an Equation (14) value of –25. This demonstrates that the optimal route is indeed the route (Figure 4.5 (c)) with travel time of 254 seconds and an Equation (14) value of –12.8.

Table 4.7 summarizes the results of this example problem, presenting the estimated travel times and constraint values for each of the different routes along with the distance of the routes. Routes (a), (b) and (d) are nearly identical in length at approximately 3.1 miles, and route (c) is slightly longer at 3.4 miles. Using the estimated travel times, variances and covariances to determine the VCSP indicate that route (c) would actually be the better risk-averse route. Additional discussion on this example problem are made in the Experimental Analysis and Results section, showing what occurs when the unconstrained route (a) and the risk-averse route (c) are simulated.

Table 4.7 VCSP Solution Example Summary

Route	**Estimated Travel Time (seconds)**	**Equation (14) Value**	**Distance (miles)**
(a)	238	24.3	3.1
(b)	273	–25	3.1
(c)	254	–12.8	3.4
(d)	247	28.4	3.1

Part II
OR for Operational Planning in Emergency Services

5

Multi-Criteria Approach Using Simulation-Based Balanced Scorecard for Supporting Decisions in Health-care Facilities: An Emergency Department Case Study

W. Abo-Hamad and A. Arisha
Dublin Institute of Technology (DIT), Dublin, Ireland

Health research is a priority in every economy, and this research – set in the context of building a more sustainable and efficient health-care system – examines how operations management practices can be translated to clinical applications. Health-care systems in general (and emergency departments (EDs) in particular) around the world are facing enormous challenges in meeting the increasingly conflicting objectives of providing wide accessibility and delivering high-quality services efficiently and promptly. The framework proposed in this study integrates simulation modelling, the Balanced Scorecard, and multi-criteria decision analysis with the aim of providing a decision support system for health-care managers. Using the Analytic Hierarchy Process, simulation results are aggregated to achieve defined strategic as well as tactical and operational objectives. Communicating the significance of investigated strategies can encourage managers to implement the framework's recommendations in the ED within the partner hospital.

5.1 Introduction

Health-care managers are currently under constant pressure to control rapidly escalating expenses, while still responding to growing demands for both high-class patient service levels and medical treatment. Resolving

Reprinted from *Health Systems*, 3: 43–59, 2014, 'Multi-Criteria Approach Using Simulation-Based Balanced Scorecard for Supporting Decisions in Health-care Facilities: An Emergency Department Case Study', by W. Abo-Hamad and A. Arisha. With kind permission from Operational Research Society Ltd. All rights reserved.

such challenges requires a consistent understanding of health-care systems, which can be an overwhelming task, given the large number and diversity of the organisations involved and their high levels of uncertainty and interdependence. Moreover, health-care managers also face the challenge of intrinsic uncertainty of the demands and outcomes involved in healthcare systems; high levels of human involvement at both patient and staff level; limited budget and resources; and a large number of variables (e.g., staff scheduling, bed availability, etc.). As well as seeking high service quality levels, patients are, understandably, less and less prepared to wait in queues for essential health services, and thus the health-care service concept has shifted from optimising resource utilisation to finding the best balance between service for patients and efficiency for providers (Brailsford & Vissers, 2011). Dealing with the inevitable complexities in health-care processes and services and addressing the challenges involved in making informed decisions are the focus of this research. The objective of this paper is to develop a simulation-based decision support framework to improve planning and efficiency of health-care processes. A real-world case study of an emergency department (ED) in one of Dublin-Ireland largest university hospitals is investigated to help the hospital executive managers enhance patients' experience using the proposed framework.

5.2 Project background

Overcrowding in EDs has become a significant international crisis that negatively affects patient safety, quality of care, and patient satisfaction (Graff, 1999). Overcrowding in Irish EDs was declared a 'National Emergency' in Ireland in 2006. Several national reports have highlighted a growing demand for emergency care (1.2 million patients attending EDs annually) and a simultaneous decrease in the number of operating EDs. The results are increased crowding, high percentages of patients leaving EDs before completing their treatment episodes, and higher morbidity and mortality rates. In addition, prolonged waiting times have been reported with more than 500 patients on trolleys for hospital admission every day; 18% of patients are waiting more than 24 h and 40% between 10 and 24 h (Health Service Executive, 2010). Although Ireland is not alone in experiencing these figures (Schafermeyer & Asplin, 2003, Bond *et al,* 2007, Forero *et al,* 2010), it is important not to underestimate the sometimes catastrophic consequences this situation has for patients, staff, and the health-care sector.

This project was a joint effort involving hospital staff (managers, consultants, doctors, nurses, and administrators) and our institute research

team. The university hospital is an acute care public hospital in North Dublin. This 570-bed hospital provides a variety of health-care services, with a 24-h 'on-call' ED that receives over 55,000 patients annually. According to the task force report in 2007, the overall physical space of the ED and infrastructure were inadequate. The hospital – which was operating at approximately 99% occupancy – had difficulty in accommodating surges in ED admission numbers. Therefore, patients who required critical care (ICU/HDC) beds suffered from significant delays and the ED could not meet the national target of 6-h average length of stay (LOS) for patients. The ED figures show clear evidence of this overcrowding, with an average of 17% of its patients choosing to leave before being seen by the ED clinician. The report also indicated that the average time from ED registration to discharge was 9.16 h, that is, 3.16 h over the 0–6 h metric set by the HSE, and the average time from registration to acute admission was 21.3 h with a standard deviation of 17.2 h (i.e., 3.5 times higher than the same national metric). Obviously, patients who are admitted will usually experience longer LOS times than those who are discharged due to delays between admission referral by an ED doctor, the allocation of a bed, and time taken to transfer the patient to the bed.

To cope with these challenges, a joint collaborative work was established with the hospital management team to develop a decision support framework. This collaboration aims to identify performance bottlenecks and explore improvement strategies to meet the HSE targets.

5.3 Proposed methodology

5.3.1 Literature review

Over the past two decades, several performance measurement systems have been introduced with the objective of achieving the full potential of performance measurement approaches (Fitzgerald *et al*, 1991, Kaplan & Norton, 1992, Lynch & Cross, 1995, Neely *et al*, 2002). Assessing performance is essential because it provides the capability to identify performance bottlenecks and take corrective action before these problems escalate (Kueng, 2000).

The Balanced Scorecard (BSC) is one of the main performance measurement frameworks that use strategy-linked leading performance measures and actions for planning and implementing an organisation's strategy (Kaplan & Norton, 1996). The BSC was discussed as an appropriate tool for health-care organisations as early as 1994, when Griffith (1994) placed the BSC in the broader notion of championship management. Several papers have described financial success stories using the BSC in health-care organisations, whether by solving financial crises (Jones &

Filip, 2000, Meliones, 2000, Mathias, 2001) or by reducing costs (Berger, 2004, Colman, 2006). The BSC has also become a regular step in quality improvement within several health-care organisations (Moullin, 2004). It has become a tool for developing quality plans and for evaluating quality improvement processes (Colaneri, 1999, Peters & Ryan, 1999).

While the BSC has been applied successfully as a strategic management tool, there are many challenges in the design and implementation. The choice of performance perspectives and measures to be included in the BSC is one of the main challenges in designing BSCs in healthcare settings. Furthermore, the number of performance measures is challenged by the amount of resources tied up in the measurement process, in terms of data collection and analysis and the representation and interpretation of the measures (Gao *et al*, 2006). Finally, the interactions between the performance indicators within the BSC are, in most papers, assumed and treated as unproblematic issues (Aidemark & Funck, 2009), ignoring the fact that several indicators can oppose each other (Patel *et al*, 2008). Because of the large number of variables and high levels of uncertainty, the BSC has to integrate with other analytical tools.

Several studies have combined multi-criteria decision analysis (MCDA) techniques with the BSC in order to overcome these challenges. The Analytic Hierarchy Process (AHP) (Saaty, 1990) is applied by many authors for the selection of performance measures to be used in the BSC (Clinton *et al*, 2002, Searcy *et al*, 2004, Wu *et al*, 2009). Other multi-criteria analysis methods have also been applied in the design and evaluation of BSCs. For example, the Simple Multi-Attribute Rating Technique (SMART) can be used to select appropriate measures for the development of a BSC system in a financial institution (Valiris *et al*, 2005).

Despite the recognised importance of explicitly dealing with priorities and trade-off between different performance indicators (Banks & Wheelwright, 1979; Eccles & Pyburn, 1992; Da Silveira & Slack, 2001), limited literature has addressed the nature of the trade-offs between these measures and their inter-dependencies (Mapes *et al*, 1997, Neely *et al*, 2000). Understanding the causes of unsatisfactory performance levels and determining proper corrective actions requires, in most cases, understanding and detailed analysis of the underlined process and the consideration of trade-offs. However, the lack of analytical tools prevents decision makers from effectively processing all the information necessary in order to develop and implement better-informed decisions and plans. Consequently, modelling and simulation are required (Sterman, 1989; Senge, 1991).

Efforts to develop simulation models have advanced since the late 1980s when simulation was used to investigate the impact of key resources on

waiting times and patient throughputs (Saunders *et al*, 1989), and it has since been used to study the effect of a wide range of health interventions on health-care processes' performance (Dittus *et al*, 1996; Kim *et al*, 1999; Ingolfsson *et al*, 2003; Litvak *et al*, 2008). Simulation models can effectively be used as a predictive tool to predict the maximum demand level that ED staff can handle, and consequently determine the required staffing level to meet that increase in demand and at the same time to keep the average waiting time of patients under a certain threshold (Baesler *et al*, 2003). A balance in the utilisation of resources would be attained by analysing the arrival pattern of patients, which can significantly improve staffing planning and resource allocation (Sinreich & Marmor, 2005. The bed occupancy level has been found to be strongly correlated with average LOS of patients within the ED (Forster *et al*, 2003). By using simulation models, Elbeyli and Krishnan (2000) found that adding beds to other specialised units within the hospital decreased the average time of patients waiting to be admitted to the ED.

Most of the prior simulation studies have used a single-perspective performance measure. Given the current complexity of the health-care systems, multiple perspectives of performance are instrumental in operational and strategic decisions. The BSC, MCDA, and simulation modelling are approaches that have independently proven their potential to inform and support the decision-making process. There is also a clear potential for these approaches to be integrated and applied in a collaborative manner that can bring new insights to inform and support the different stages of the decision-making process.

5.3.2 Proposed framework

The main objective of this framework is to address the limitations in the literature and to provide health-care managers and planners with an integrated decision support tool that can be used in an effective and practical manner. This section discusses the aspects and requirements for developing such a framework. Figure 5.1 gives an overview of the framework, and the following sub-sections provide detailed descriptions of each component, and highlight the coordination between them and their points of integration.

5.3.2.1 Business process modelling

Defining the problem to be solved is one of the key elements in developing the framework. Health-care systems contain high levels of social interaction that are characterised by complexity, particularly at decision points, with the result that health-care service delivery and patient

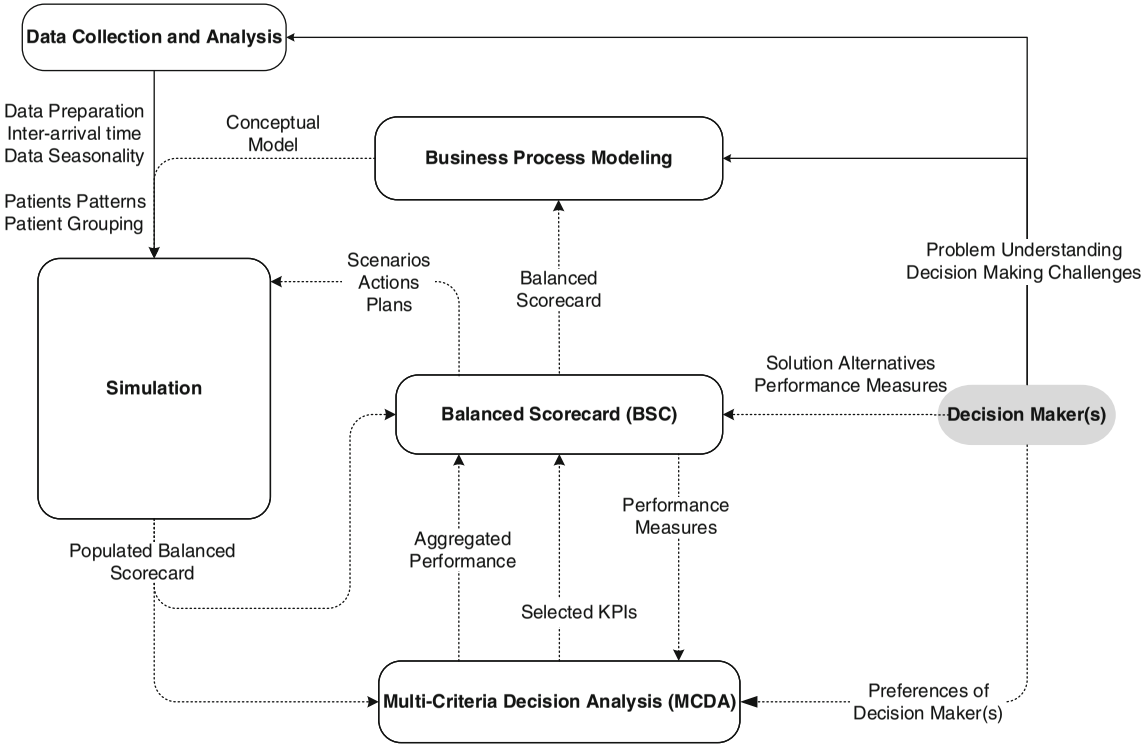


Figure 5.1 An overview of the integrated framework

flow management problems are usually hard to define. Gaining a better understanding of the health-care process is essential for making correct and justifiable decisions and providing effective solutions, and therefore modelling the underlined business process requires that problems be understood from the point of view of the individuals directly involved in service delivery.

In order to provide a holistic view about various aspects of the system, the data collection phase combined interviews, focus groups, and quality circles with experts and the underlined business processes were then mapped onto a conceptual process model using one of the well-developed modelling languages where sub-processes and activities are identified. The control flow definition is created by identifying the entities that flow through the system (e.g., patients, staff) and describing the connections that link different parts of the process, and resources are identified and assigned to activities where necessary. The process model must be verified to ensure that it is logically valid and does not contain errors.

5.3.2.2 Simulation modelling

The process model was combined with the analysed empirical data into a dynamic simulation model, so that both the data collection phase and the business process modelling take place within the context of developing the simulation model. The procedure is often referred to as model translation because it involves transforming an abstract conceptual model into a more detailed and complex executable simulation model. To ensure that the credibility of the simulation model can be guaranteed, it must be both verified and validated. Verification ensures that the transformation of the conceptual model has been applied correctly so that the model's logic reflects the underlying business process, while validation involves comparing the outcome data of the simulation model with the data obtained during the data collection phase (Balci, 1997). Once the simulation model is verified and validated, the decision makers can use the replicated model to investigate a number of decisions and alternatives (i.e., what-if scenarios) to foresee their consequences.

5.3.2.3 Integrating BSC and simulation

Although it can be applied in the context of health-care management, the full potential of the BSC cannot be realised in this context because of its limitations and the challenges involved in its implementation. In order to alleviate the BSC's limitations in terms of its measurement capabilities and its inability to identify cause-and-effect inferences between performance measures, an integration between the BSC and simulation

is proposed. Performance perspectives and performance measures are collected by interviewing senior managers of health-care facilities (e.g., an ED). This step is essential to align the facility's performance measures with the strategic objectives of the national health authorities (i.e., HSE), so that the simulation model will provide quantitative values of the performance measures, and qualitative measures (such as patient satisfaction) can be related to measurable indicators (such as average waiting and LOS times). Such integration allows for the evaluation of a wide range of actions and plans based on the recommendations of national reports and surveys, which can then be evaluated in the form of *what-if* scenarios, and the results used to populate the designed BSC.

The results are then evaluated and interpreted by decision makers, who provide guidance on the implementation of suggested decision alternatives and plans, and set benchmarks of the maximum performance that can be achieved using the available resources and staffing levels. Thus, integrating simulation and the BSC helps focus efforts on strategic visions to obtain desired outcomes, assists in making better decisions, improves communication within the organisation, provides continual feedback on strategies, promotes adjustments to changes, and assists both individuals and organisations in achieving their goals and objectives – and at the same time the simulation process can provide interesting information about the cause-and-effect relationships among performance measures.

5.3.2.4 Multi-criteria decision analysis

Although the BSC's measurement limitations can be resolved by integrating it with simulation, the large number of measures in the BSC delays the evaluation and analysis of the results, especially where they may be conflicting or even opposed to each other. MCDA tools can play an important role in addressing these challenges and overcoming the problems of selecting and evaluating the *key* performance measures during the design phase of the BSC. In the design phase of the BSC, MCDA methods can be applied for the selection of appropriate performance measures, and decision makers can evaluate and prioritise performance measures, which can then be illustrated in a *value tree* that represents the selected key performance indicators (KPIs). Following their selection, the resulting value tree is passed to the simulation model. MCDA can then effectively aggregate the simulation output (i.e., KPIs) into a marginal performance according to decision makers' preferences. This dual use of MCDA within the integrated framework can contribute greatly to making informed decisions for improving and managing the health-care business process.

5.4 Framework implementation

5.4.1 An ED – a case study

The ED of the hospital has 13 monitored trolley spaces, 3 of which are in a resuscitation area and are reserved for major trauma and critical care patients; an ambulatory care area (capacity 6 trolley spaces); two isolation rooms; a psychiatric assessment room; two rapid assessment triage bays; and two other triage rooms. The layout of the ED is shown in Figure 5.2. Five distinct areas can be identified: a waiting room for walk-in patients waiting for triage, a diagnostics area (X-ray and CT scan), an ambulatory care unit (ACU) area, an ED resuscitation area (CPR), and an ED major assessment area. Patients arriving by ambulance – usually in critical condition – are routed directly to the resuscitation area, whereas patients whose conditions require monitoring stay in the major assessment area. The ambulatory care area is for patients arriving on foot, who may be suffering from abdominal pain, headache, limb problems, wounds, head injuries, facial problems, and so on.

As a 24-h department, the ED has three consultants, two nursing managers, and 11 nurses during the day and nine nurses at night, divided into six types of nurse: advanced nurse practitioners (ANPs), triage nurses, resuscitation nurses, respiratory nurses, majors/minors nurses, and healthcare assistants. Physicians (excluding the three consultants who provide cover between 9 am and 5 pm (or 8 am and 8 pm) with 24/7 on-call provision) are divided into three types, registrar/specialist registrars; Senior House Officers (SHOs), and interns, and are distributed as follows when the roster allows: three registrars per day working 10-h shifts starting at 8 am, 12 pm and 10 pm; two interns working daily 8 am – 5 pm shifts Monday to Friday; and 12 SHOs working fixed shifts during the day and night to keep the ED running. Therefore, the number of doctors on duty varies between two and seven, depending on the time of day or night.

5.4.2 Process mapping

A variety of data collection methods such as interviews, focus groups, observations, and historical data were used to develop a comprehensive conceptual model for the ED. Four preliminary interview sessions with senior managers (two ED consultants and two nursing managers) were carried out in order to gain insights about the current challenges they face in managing their department. A better understanding of healthcare processes, activities, challenges, and variables was acquired with valuable insights into the challenges in the decision-making process. The interviews helped to develop significant inputs that critically

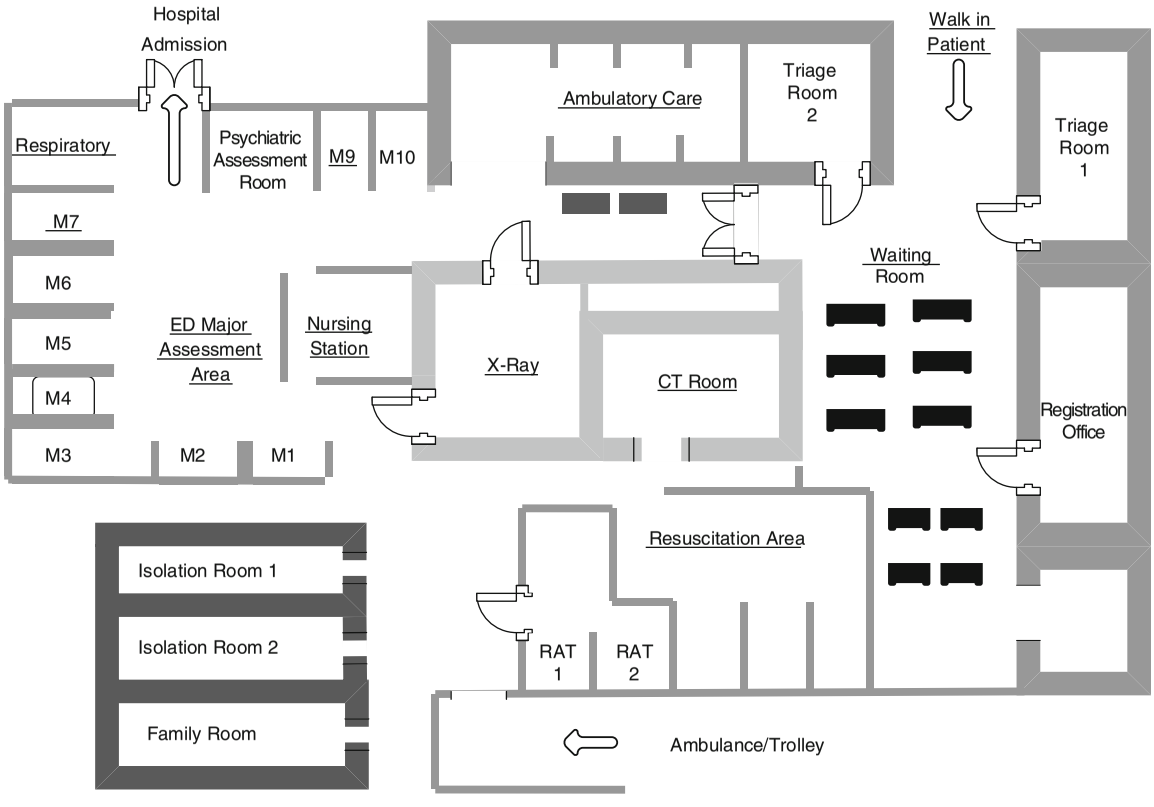


Figure 5.2 ED physical layout and main care areas

supported the development and validation phases of the proposed framework. This was followed by constructing a focus group of ED doctors (one registrar and three SHOs) and nurses (a triage nurse, one ANP, and two general nurses) and a weekly meeting was scheduled for discussing issues such as general patient care paths, categories of patients and their complexities, and resource availability and capacity issues. Meanwhile, a number of visits were made to the ED (i.e., site visits) with the objective of analysing the ED layout, which reflects how resources are allocated and utilised within the ED. A high-level understanding of the journey of the patient through the ED was acquired from the initial findings of the interviews. Upon the arrival of walk-in patients (self- or GP referred), they register and wait in the waiting area to be triaged. When their name is called (depending on triage staff availability) they are generally assessed by a triage nurse. On the basis of their condition and triage assessment, each patient is assigned a clinical priority (triage category) according to the Manchester Triage System (MTS), which is widely used in the United Kingdom, Europe, and Australia (Cronin, 2003) and uses a five-level scale for classifying patients according to their care requirements: immediate, very urgent, urgent, standard, and non-urgent. Once a triage category is assigned, the patient may be sent back to the waiting room until a bed or trolley is available in an area where they can be given treatment appropriate to the type and intensity of their care needs. Waiting times for patients will depend on their triage category and the availability of both medical staff (i.e., ED physician or ANP) and empty trolleys, which are a prerequisite for full and accurate assessment. After they have been assessed by an ED clinician, a decision is made either to discharge or to admit. These are the primary care stages that apply to all patients, whether they are discharged from or admitted to hospital. Secondary patient stages are those steps that may be involved in the care of some (but not all) patients such as diagnostics (e.g., X-rays, blood tests, etc), and further ED-doctor assessment or consultation with a medical/surgical specialty doctor to confirm whether a patient should be admitted or to gain advice on the best possible treatment for a patient being discharged. Figure 5.3 shows a detailed flowchart for patient journey through the ED.

The developed flow charts for patient flow were effective in fast and informal process representation, and therefore they are effective in communication and discussions between analysts and stakeholders. However, flowcharts use a sequential order of actions, do not support a breakdown of activities (Aguilar-Saven, 2004), and lack the necessary semantics to support more complex and standardised constructs

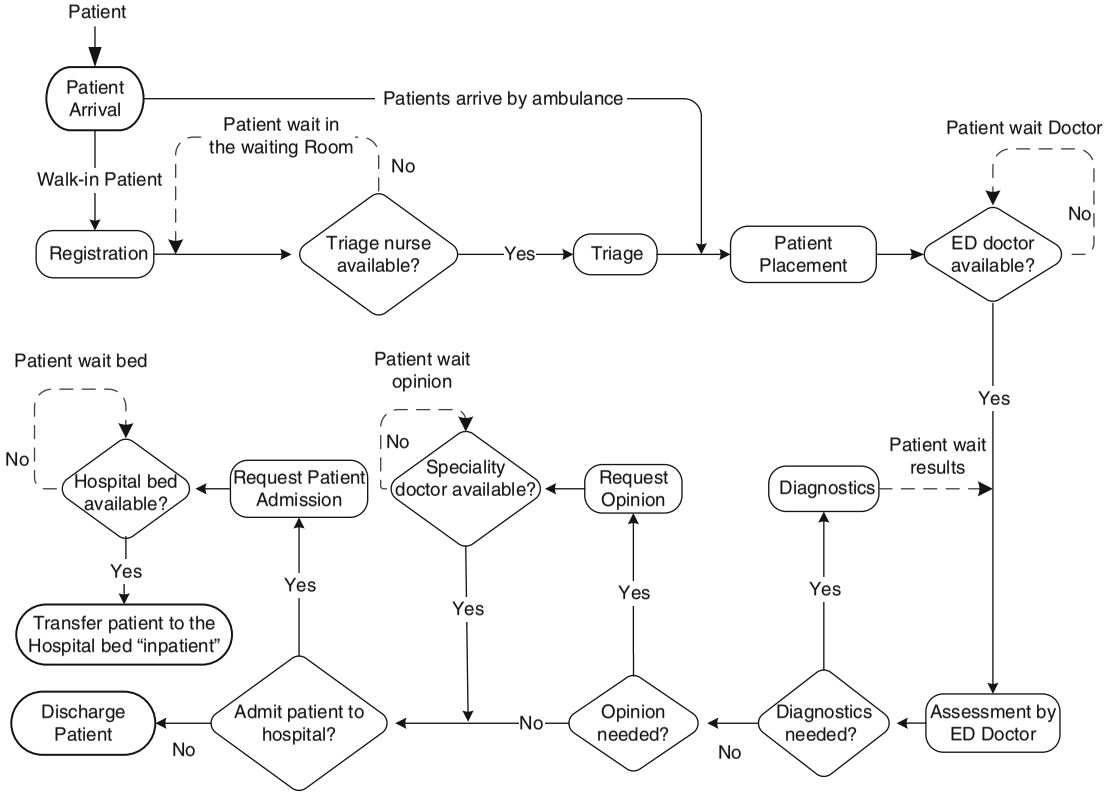


Figure 5.3 Detailed patient flow through the ED

(Havey, 2005). Therefore, different levels of detail about the patient flow were collected by the research team over a number of site visits. Site visits were carried out two times per week and different weekdays were selected at different hours (i.e., morning, afternoon, and night time). This was an essential step in order to observe the variability of care service demand (i.e., patient arrival) and to note the processes that the patient goes through. On the basis of the analysis of this stage, each ED process was broken down into smaller sub-functions, and key resources (e.g., staff and medical equipment) were identified at each care stage and detailed using IDEF0. IDEF0 is a structural graphical representation of processes or complex systems that allows the analysis and communication of the functional aspect of a system (NIST, 1993). Each process in IDEF0 is described as a combination of activities, inputs, controls, and mechanisms in a hierarchical fashion. At the highest level the representation may be of an entire process. The processes can be further decomposed to show lower-level activities. The breakdown of processes may continue until a point where sufficient detail is reached (Colquhoun *et al*, 1993). This hierarchical structure of IDEF0 keeps the model scope within the boundaries and allows the system to be easily refined into more detail until the model is as descriptive as necessary for the decision maker (Kim & Jang, 2002). The top level of the developed IDEF0 model for the ED is shown in Figure 5.4. The main unit of an IDEF0 model is an activity block that describes the process's main function, with ICOMs (Input, Control, Output, and Mechanism) represented by horizontal and vertical arrows. Process control (top arrow) can be patient information (e.g., arrival time, triage category, and presenting complaint), safety regulations, or national/international standards, whereas process mechanisms are usually the agents and/or physical resources that facilitate the activity (e.g., ED physicians, nurses, and beds/trolleys).

5.4.3 A BSC for the ED

In conjunction with the process mapping phase, a number of interviews with ED senior managers (two consultants and two nurse managers) took place to collect information about the performance areas and performance measures. Incorporating these measures at that stage was very useful for developing the BSC for the ED and setting the objectives of the simulation model. The findings of this stage resulted in the selection of four performance perspectives in the design of BSC: community, patient, internal business processes, and learning and growth. An overview of these perspectives is given in Figure 5.5 and brief details of performance measures in each perspective are discussed.

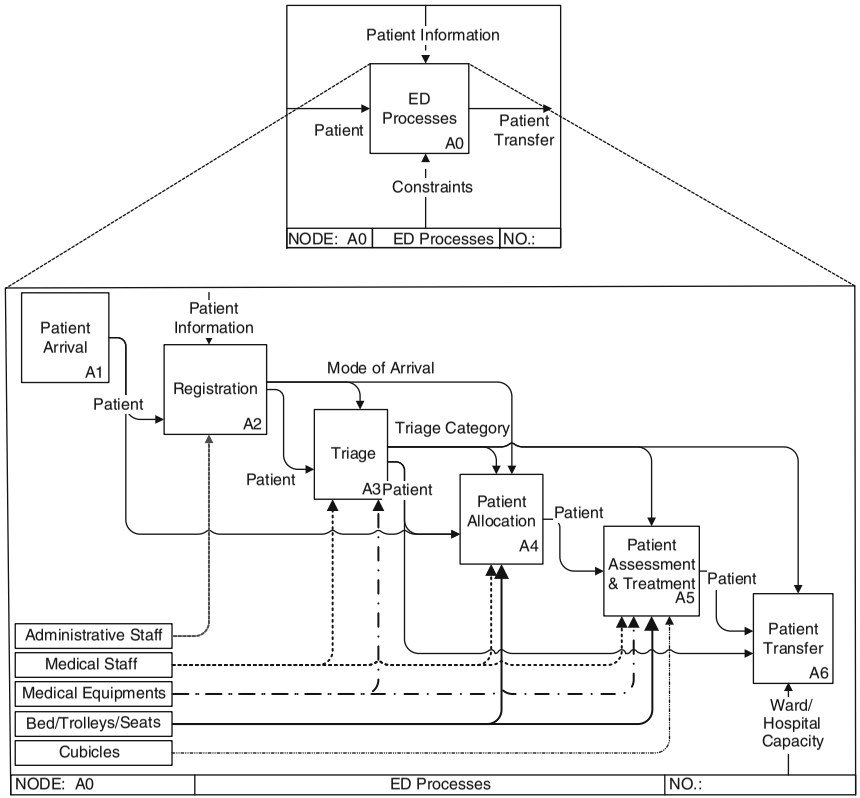


Figure 5.4 Mapping main ED processes

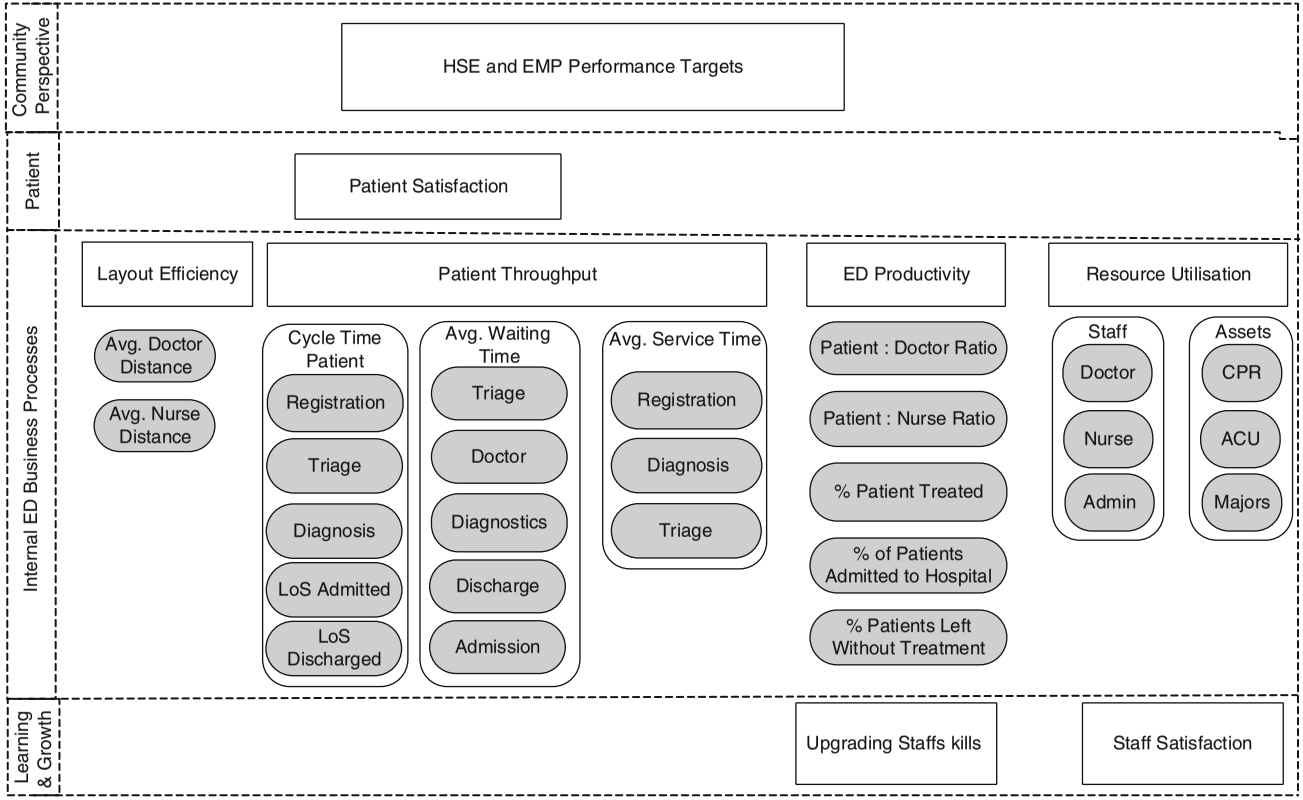


Figure 5.5 The ED Balanced Scorecard

Community engagement perspective: This perspective brings HSE performance targets and national Emergency Medicine Program (EMP) measures into the BSC. The HSE performance target is that all patients be processed through the ED within 6 h of arriving, before 'separation' (i.e., including discharge or admission where relevant). The overarching aims of the EMP are to improve the safety and quality of patient care in EDs and to reduce waiting times for patients. In designing the ED BSC, 'patient' was selected as a sole perspective and 'patient satisfaction' as its main measure. The efficiency of internal ED processes impacts patient satisfaction levels, and therefore average patient waiting and LOS times are connected to this performance measure. The main objective in the *Internal Business Processes* perspective is to improve the ED performance in terms of its layout efficiency, ED productivity, resource utilisation, and patient throughput. The layout efficiency measures the average daily distances travelled by doctors and nurses, while the ED productivity is measured in terms of five indicators: the ratio of patients per doctor, the ratio of patients per nurse, the percentage of patients treated, the percentage of patients admitted to the hospital, and the percentage of patients who leave the ED without treatment. Resource utilisation is measured for two types of resources: ED staff and ED assets such as major trolleys, ACUs, and resuscitation rooms (CPRs).

Patient throughput is measured via three dimensions: average patient cycle times, average patient waiting times, and average patient service (processing) times. The total patient cycle time is measured across the different stages of a patient's journey in the ED such as registration, triage, treatment, and diagnostics. This includes LOS for both admitted and discharged patients. Similarly, patients' average waiting times are detailed for each stage, for example, the average wait for triage, to be seen by ED physician, and for discharge or hospital admission.

Learning and growth perspective: Two main performance measures are selected in this perspective: staff development and staff satisfaction levels. The former is measured in terms of the effect of training the staff to do more than one task so that they can be allocated dynamically within the ED. The latter is related to the 'internal ED business processes' perspective through the following indicators: staff utilisation, ratio of patients per doctor, and ratio of patients per nurse.

5.4.4 Data analysis

A focus group for historical data collection was formed to discuss issues related to electronic patient records, existing information systems, and data entry procedures. The focus group included members from the

information system department in the partner hospital. The discussions with the focus group were supported by close observation of the data entry procedures through the patient journey and by a series of short interviews with the ED staff (e.g., registration staff, triage nurses, and physicians). A real-time patient tracking information system was used to track the patient's journey within the ED. Each patient record details the following patient-level variables: (1) the patient arrival mode, (2) the date/time the patient attended the ED, (3) the date/time of patient triage, (4) the triage category assigned to the patient, (5) the date/time the patient was seen by the doctor, (6) the medical complaint presented by the patient, and (7) whether the patient left without being seen, was discharged, or was admitted to the hospital. A total of 59,986 anonymous patient records were collected over a 16-month period. Patient records were analysed to extract quantitative information about their arrival patterns, patient groupings and allocations, and routing information. Patients were grouped based in their triage category. Table 5.1 summarises the analysis of patient information for each triage category along with arrival mode.

An estimated distribution of patient inter-arrivals was used to input arrival patterns for each patient group into the simulation model.

Regarding patient allocation data, Table 5.2 shows the analysis of the places/locations to which patients were allocated within the ED. The analysis shows that the ED staff members were unable to implement the MTS triage recommendations concerning the disposition of patients. Thus, 88% of 'Immediate' category patients were seen in the resuscitation room and 9% in the majors' cubicles, while 40% per cent of 'very urgent' patients were seen in inappropriate assessment areas (e.g., ACUs). Moreover, because of the overcrowded nature of the ED, the majority of standard and non-urgent patients were assessed and treated in inappropriate areas (e.g., chairs or waiting areas).

Table 5.1 Summary of the analysis of patient's records

Triage category	Percentage of patients	Arrival mode	
		Walk-in (%)	Ambulance (%)
IMM	1.1	5	95
VURG	16.5	40	60
URG	58	61	39
STD	23.9	81	19
NURG	0.5	72	28

Note: IMM: immediate, VURG: Very urgent, URG: Urgent, STD: Standard, NURG: Non-urgent.

Table 5.2 Analysis of patient allocation within the ED

ED areas	Triage category				
	IMM (%)	VURG (%)	URG (%)	STD (%)	NURG (%)
Resuscitation room	88	25	2	0	0
Majors area	9	15	8	1	0
Ambulatory care unit	0	12	10	20	11
Majors chairs	0	7	6	1	1
Rapid assessment Triage	3	12	7	2	2
Waiting room	0	14	56	74	85
X-ray sub-wait area	0	15	12	4	1

Note: IMM: Immediate, VURG: Very urgent, URG: Urgent, STD: Standard, NURG: Non-urgent.

5.4.5 Simulation model development and validation

A comprehensive simulation model was developed by the research team for the ED based on the ED business process model, the designed BSC, and the analysis of empirical data. The simulation model comprised a number of modules. These modules were linked together in the same way the blocks were linked in the conceptual flow chart; this facilitated the model construction phase. The top level of the simulation model defined the overall model structure and the sub-level blocks containing additional modules with greater detail. Object-oriented programming was used to customise pre-defined blocks for constructing the ED simulation model. Moreover, a relational database was used to save the measured KPIs after each simulation run, after which the populated BSC data were exported in tabular form for future analysis and validation. To reduce the time of the model development cycle and to increase the confidence of the ED simulation model results, verification and validation were carried out throughout the development phases of the model. Furthermore, each model development phase was verified and validated against the previously completed phases. The verification of the model's logic was carried out to ensure that patients in the simulation model follow the correct expected care paths; this was achieved by visually tracking patients (using animation) and checking intermediate output values such as queue lengths and waiting times between processes. The conceptual model was documented and validated by circulating it among ED senior managers and senior nursing staff, crucial steps to ensure that the logics of the model and ED activities were correct. All distributions determined from the data and used in the model were validated using the Kolmogorov–Smirnov goodness of fit test with a 5% significance level (Massey, 1951). Simulation variables were initialised using the empirical

data, the ED layout, and patient flow analysis given in previous sections. Queues at each stage of patient care (e.g., triage, seen by doctor, awaiting admission, and discharge) were set as empty and idle. A warm-up period of 2 months was found to mitigate any bias introduced by the initial conditions of the simulation model. The final results of the simulation model were validated using face validation and comparison testing. Face validation was performed by interviewing ED senior managers and nursing staff to validate the final results of the simulation model. Comparison testing involved comparing the output of the simulation model with the real output of the system under identical input conditions (Balci, 1997). Three main KPIs are used in this approach: average waiting times until seen by doctor, average LOS for discharged patients, and average LOS for admitted patients. In addition to the overall averages for all patients, detailed data for each KPI were also calculated for three triage categories: VURG, URG, and STD. On the basis of the comparison testing approach, the deviation between actual and simulated results for these KPIs ranged from 1 to 11% with an average of only 6% (Figure 5.6).

The comparison in Figure 5.6 shows that waiting times for urgent patients (URG) has the largest deviation (11%), which is reflected in

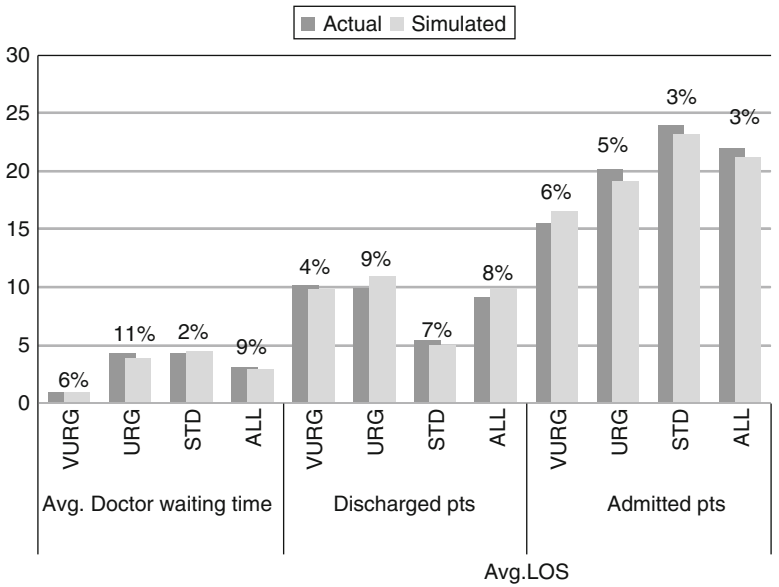


Figure 5.6 Validation of simulation results against actual data
Note: IMM: Immediate, VURG: Very urgent, URG: Urgent, STD: Standard, NURG: Non-urgent.

the total average LOS for the same group of patients (9% for discharged patients and 5% for admitted patients). According to the ED consultants, urgent patients (who represent 60% of patients attending the ED) are the most challenging and diverse group of patients with a wide range of medical complaints and ageing conditions. The underlying assumptions used to build the simulation model have also factored in such deviation. For example, only staff activities related to direct contact with patients were considered; other routine work and break times could not be considered because of the high level of variations in these activities and lack of accurate data.

5.4.6 KPIs selection

The BSC developed for the ED in the previous section includes qualitative measures – such as patient satisfaction, staff skills upgrading, and staff satisfaction – as well as quantitative measures. Although these measures cannot be measured directly in the simulation, they are directly related to the quantitative performance measures in the 'Internal Business Processes' perspective, which can be directly measured in the simulation model. Nevertheless, there is a level of redundancy between the performance measures in the internal ED business process perspective – for example, 'percentage of Patients Treated' and 'percentage of Patients Leaving without Treatment' are clearly complementary. Several ED measures – such as staff utilisation and staff satisfaction – may actually conflict; thus, maximising staff utilisation may reach burnout levels (i.e., 85% utilisation), which then decreases staff satisfaction levels. Consequently, to narrow down the list of the measures and to achieve a useful trade-off between conflicting objectives, MCDA tools are used to select the main KPIs systematically.

The selection process is based on SMART (Barron & Barrett, 1996) to identify the *alternatives* and *criteria* that are relevant to the decision problem. SMART begins with identifying the alternatives (in this case, performance measures in the BSC) and specifying the criteria to be used for evaluating these alternatives. The SMART procedure is applied to the performance measures in the 'Internal ED business processes' perspective. Therefore, the 26 performance measures within this perspective are considered as the 'decision alternatives' for the SMART procedure, and are then evaluated against the main ED performance drivers, namely, layout efficiency, patient throughput, ED productivity, and resource utilisation. Once the alternatives and criteria were identified, a value tree was produced (as shown in Figure 5.7): the root of the tree represents the ED performance, the first level represents the evaluation criteria, and the second level represents the candidate alternatives.

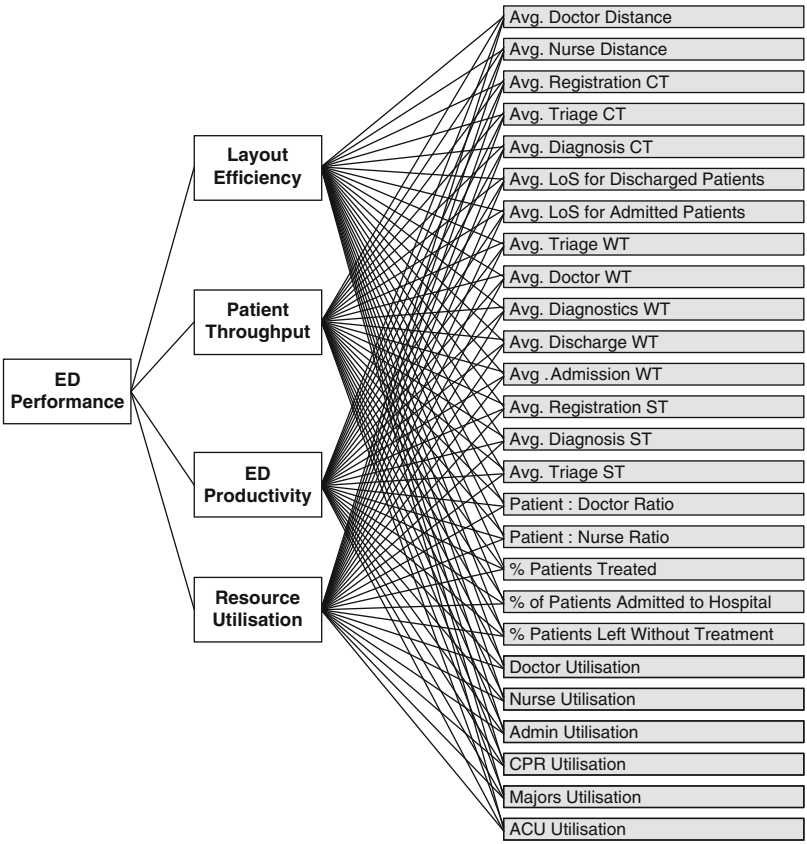


Figure 5.7 Alternatives value tree

The ED managers were asked to rank the alternatives with respect to each criterion in order, from the most to least preferred, on an easy-to-use value scale (Valiris *et al*, 2005). For each criterion, a value of 100 was given to the most relevant measure and 0 to the least relevant. With respect to the 'layout efficiency' criterion, for example, the 'average distance travelled by doctors' within the ED was seen as the most relevant and the 'average patient registration service time' as the least relevant. Table 5.3 represents preferences for each of the four main criteria where the bold values in each column are given to the most relevant measure to that column.

The remaining set of alternatives are then rated regarding the most relevant and the least relevant and assigned a value that ranges from 0 to 100. As the evaluation criteria were not of equal importance, their

Table 5.3 ED senior managers' rating of alternatives for each criterion

Alternatives	Evaluation criteria			
	Layout efficiency	Patient throughput	ED productivity	Resource utilization
Average doctor distance	**100**	50	70	90
Average nurse distance	90	50	70	80
Average registration CT	10	60	20	20
Average diagnosis CT	20	60	50	20
Average triage CT	20	30	50	10
Average LOS for discharged patients	30	**100**	30	70
Average LOS for admitted patients	40	90	30	70
Average triage WT	30	20	0	10
Average doctor WT	60	70	40	40
Average diagnostics WT	10	0	10	10
Average admission WT	20	20	10	20
Average discharge WT	20	10	10	0
Average registration ST	0	10	10	20
Average diagnosis ST	20	10	20	10
Average triage ST	20	10	10	20
Patient to doctor ratio	40	80	90	90
Patient to nurse ratio	40	80	80	90
Percentage of patients treated	50	90	**100**	80
Percentage of patients admitted	30	60	20	70
Percentage of patients left without treatment	20	30	30	30
Doctor utilisation	70	70	90	**100**
Nurse utilisation	70	70	90	90
Administrator utilisation	10	20	30	20
CPR trolleys utilisation	70	70	80	80
Majors trolleys utilisation	80	70	80	80
ACU trolleys utilisation	60	60	70	70

Bold value in each column refers to the most relevant measure to that column.
Note: CT: Cycle time, WT: Waiting time, ST: Service time, LOS: Length of stay.

relative importance to the overall ED performance was ranked by the ED consultants as shown in Table 5.4.

The normalised weighting is calculated by dividing the 'value score' for the particular criterion by the total for all value scores, that is, for Rank 1, 100/270 = 0.37. The total score is then calculated for each

Table 5.4 The relative importance of the evaluation criteria

Rank	Criterion	Value score	Normalised weighting
1	Patient throughput	100	0.37
2	ED productivity	80	0.29
3	Resource utilisation	60	0.22
4	Layout efficiency	30	0.11

Table 5.5 Aggregated weights and values for 'percentage of patients treated'

Criterion	Value score	Criterion weight	Alternative weight
Layout efficiency	50	0.11	5.56
Patient throughput	90	0.37	33.33
ED productivity	100	0.29	29.63
Resource utilisation	80	0.22	17.78
Total			86.30

Table 5.6 Final scores and alternative rankings using SMART procedure

Alternatives	Total score	Rank	Alternatives	Total score	Rank
Percentage of patients treated	86.30	1	Percentage of patients admitted	47.04	14
Doctor utilisation	82.59	2	Average diagnosis CT	43.70	15
Patient to doctor ratio	80.74	3	Average registration CT	33.70	16
Nurse utilisation	80.37	4	Average triage CT	30.37	17
Patient to nurse ratio	77.78	5	Percentage of patients left without treatment	28.89	18
Majors trolleys utilisation	76.30	6	Administrator utilisation	21.85	19
CPR trolleys utilisation	75.19	7	Average admission WT	17.04	20
Average doctor distance	70.37	8	Average diagnosis ST	14.07	21
Average nurse distance	67.04	9	Average triage ST	13.33	22
ACU trolleys utilisation	65.19	10	Average triage WT	12.96	23
Average LOS for discharged patients	64.81	11	Average registration ST	11.11	24
Average LOS for admitted patients	62.22	12	Average discharge WT	8.89	25
Average doctor WT	53.33	13	Average Lab WT	6.30	26

Note: CT: Cycle time, WT: Waiting time, ST: Service time, LOS: Length of stay.

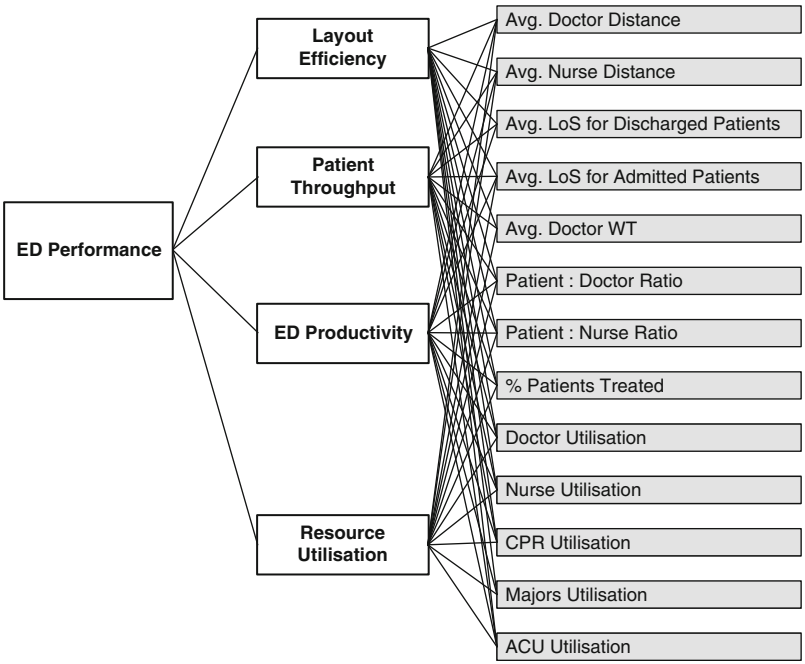


Figure 5.8 ED KPIs

alternative as the weighted average of the value scores for all of its criteria. For example, Table 5.5 shows the aggregated weights and values for 'percentage of Patients Treated'.

Table 5.6 summarises the final weighted scores for all the alternatives and specifies the rank of each alternative.

Finally, the consultants set a cut-off level of 50 for the total score for the alternatives to highlight the most important factors and leave the others out of consideration to make the results simpler to use (Figure 5.8). These final sets of alternatives were then passed to the simulation model as the simulation output.

5.5 Real-time strategies for the ED

5.5.1 Scenario design

The main scenarios introduced involved increased clinical assessment resources (adding six extra trolley cubicles), increased availability of clinical assessors (adding one extra SHO shift at night), and absolute compliance with the national 6-h admission target (zero tolerance)

Table 5.7 Simulation variables for baseline system and scenarios

Scenarios	Decision variables		
	Access block	**Physical capacity**	**Staff**
Base line	Yes	13	—
Capacity Expansion	Yes	19	—
Increasing staff	Yes	13	1 SHO [9 pm to 7 am]
Zero tolerance	No	13	—

(Table 5.7). Each scenario runs for 3-month blocks, a period identified by ED managers as being generally associated with stable ED staffing levels.

The scenarios were suggested by the ED senior managers to evaluate the proposed new hospital extension that was intended to include rebuilding of key parts of the facility including the ED. As expanding the ED capacity was likely to eventually necessitate corresponding increases in its staffing levels, hospital managers and those planning the new ED were interested in evaluating the effects of capacity and staffing levels expansion against the effects of unblocking critical performance bottlenecks such as the 'access block' from the ED to the hospital.

5.5.2 Analysis of results

The results of the simulation model (see Table 5.8) showed that adopting Scenario 3 (absolute enforcement of the national 6-h admission target) had the most significant impact on the average LOS at every stage of patients moving through the ED. Average LOS for patients who are ultimately discharged directly from the ED decreased from 10.23 h to 5.3 h (48% improvement in LOS).

The changes that followed from the first two scenarios, namely, 'capacity expansion' and 'increasing staff', resulted in fewer improvements that were neither clinically significant nor improved patient experience (i.e., had a negligible impact on average LOS for admitted patients). The 'zero-tolerance' scenario improved how the department utilised its physicians, and was also expected to improve the average LOS of patients waiting to be admitted to the hospital.

In order to consider the preferences of the ED managers in the analysis of these scenarios, AHP was used. Table 5.9 presents the AHP comparison matrix for the four main ED performance criteria and their corresponding weightings.

A comparison matrix for each criterion was then constructed to obtain the weights of individual KPIs. Table 5.10 shows the comparison matrix

Table 5.8 Simulation results of Scenarios 1, 2, and 3

Key Performance Indicators (KPIs)		Base line	Capacity expansion		Increasing staff		Zero tolerance	
			O/P	↑↓ (%)	O/P	↑↓ (%)	O/P	↑↓ (%)
Patient throughput	AWT doctor (h)	2.96	2.50	15	2.80	5	1.80	39
	Average LOS Dis. Pts. (h)	10.23	8.40	18	9.80	4	5.30	48
	Average LOS Adm. Pts. (h)	21.30	18.20	15	19.80	7	5.70	73
Resource utilisation	Doctor utilisation	81%	84%	4	73%	10	86%	7
	Nurse utilisation	82%	87%	7	83%	1	74%	10
	CPR utilisation	91%	86%	6	91%	0	87%	5
	Majors utilisation	94%	82%	13	92%	2	85%	10
	ACU utilisation	93%	75%	19	94%	2	83%	11
Layout efficiency	Average doctor distance (km/d)	3.24	3.63	12	2.83	13	3.91	21
	Average nurse distance (km/d)	6.48	7.32	13	6.55	1	5.34	18
ED productivity	Patient:Doctor Ratio	7.34	7.52	2	7.14	3	7.9	8
	Patient:Nurse Ratio	9.84	10.22	4	10.16	3	10.8	10
	Percentage of patients treated	83%	85%	2	90%	8	96%	16

Note: AWT: Average waiting time, LOS: Length of stay, Dis. Pts.: Discharged patients, Adm. Pts.: Admitted patients.

Table 5.9 AHP comparison matrix for main KPIs in ED performance criteria

	LE	PT	PR	RU	Resulting AHP weight
LE	1	0.125	0.167	0.25	0.046
PT	8	1.000	3.000	6.00	0.581
PR	6	0.330	1.000	3.00	0.285
RU	4	0.167	0.330	1.00	0.116

Note: LE: Layout efficiency, PT: Patient throughput, PR: ED productivity, RU: Resource utilisation.

for the three KPIs representing the 'Patient Throughput' criterion and their AHP weightings.

The same process of comparing pairs of KPIs for each main criterion was repeated until the last level was reached. Figure 5.9 shows the final weights for all the levels in the performance value tree. After calculating the relative weightings, ED managers determined the acceptable range

Table 5.10 The comparison matrix for the KPIs of the patient throughput criterion

	Average LOS distance	**Average LOS administrator**	**A.W.T. doctor**	**AHP weight**
Average LOS discharged patients	1	0.33	4	0.304
Average LOS admitted patients	3	1	3	0.575
AWT doctor	0.25	0.33	1	0.121

Note: AWT: Average waiting time, LOS: Length of stay.

for each KPI. For example, staff utilisation (for nurses and doctors) was given a range between 50 and 85% to avoid burnout levels (of 85%). Similarly, a range between 0 and 6 was specified for the LOS KPI to measure the levels achieved by each scenario while keeping to the 6-h maximum HSE target. After the acceptable ranges had been assigned, a value function was attached to each individual KPI to describe the desirability of achieving different performance levels.

Given the results of the simulation model reported in Table 5.8, and the AHP preference model in Figure 5.10, the final value for each scenario including the baseline scenario (which was set as the current ED operation) was aggregated and summarised in Table 5.11.

The analysis of the AHP results showed that if the hospital implements the 'zero-tolerance' strategy this will have the most significant impact on the throughput of patients (producing 54.6% increases), and on the overall ED performance (Figure 5.10).

5.5.3 Sensitivity analysis

A one-way sensitivity analysis was conducted to explore how sensitive each decision alternative (i.e., strategy or scenario) is to variations in performance measures. In one-way sensitivity analysis, single-attribute value functions or attribute ratings for decision alternatives are varied, one at time, to see how sensitive the model is to those changes. The total values of decision alternatives are drawn as a function of the variable under consideration. Figure 5.11 shows the sensitivity of proposed scenarios to the variation in average LOS. The zero-tolerance scenario gives the highest overall value for the ED performance. The baseline gives the lowest value among all the alternatives as the average LOS increases. The increase of the average LOS for the ED above 6 h will deteriorate the performance of the current ED at all levels, which necessitates the addition of more staff and the expansion of the ED at this stage. However, enforcing the 6-h

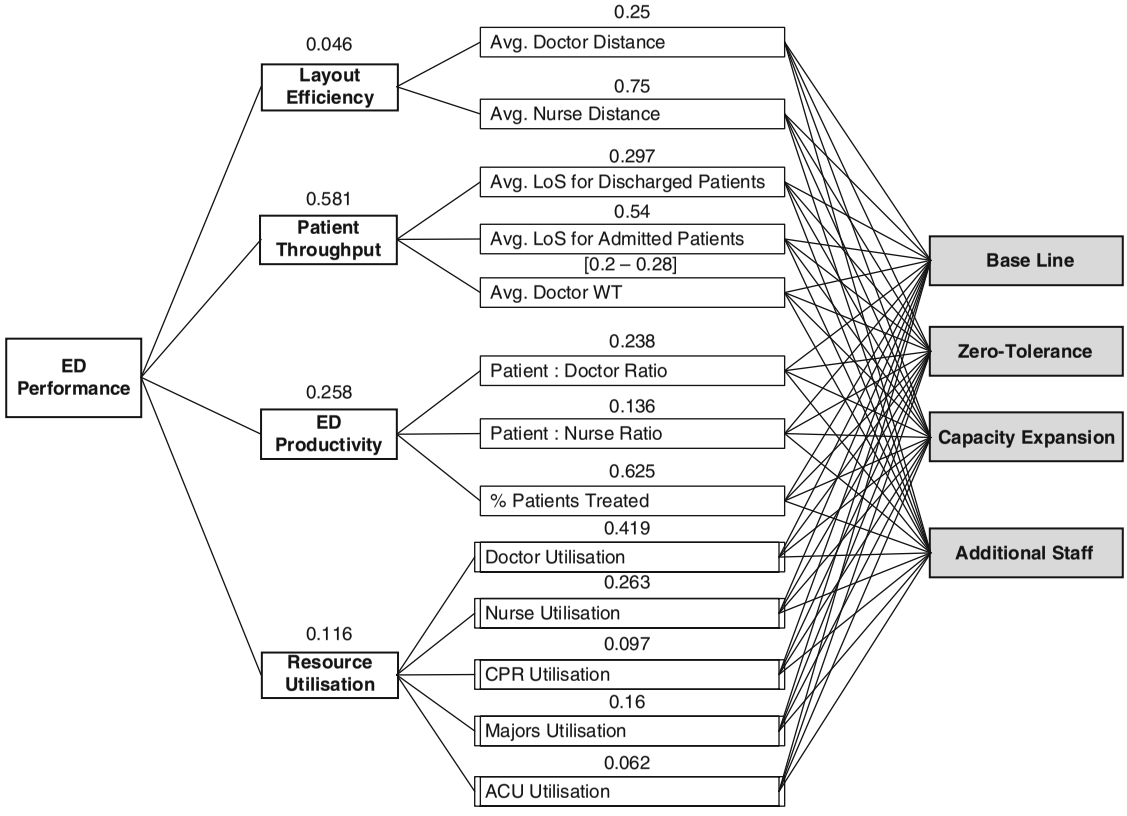


Figure 5.9 Full AHP weighted value tree

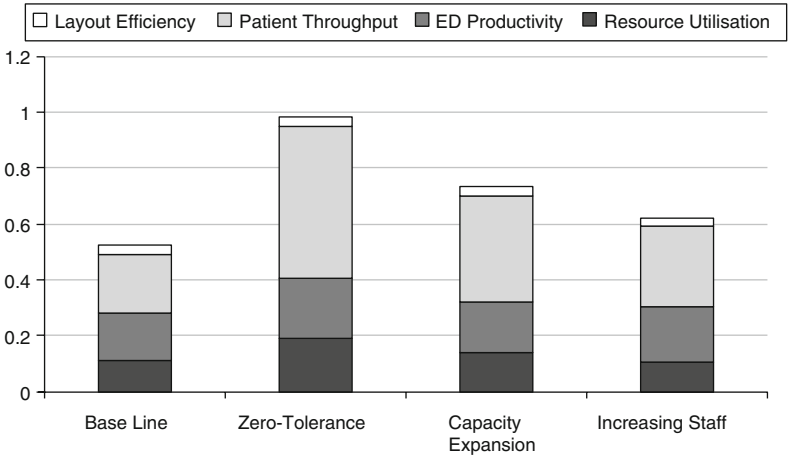


Figure 5.10 The ED performance for all the scenarios against the current ED

Table 5.11 Weighted results for all scenarios against the baseline scenario

	Baseline	Capacity expansion	Increasing staff	Zero-tolerance
Resource utilisation	0.11	0.14	0.108	0.19
ED Productivity	0.169	0.18	0.194	0.215
Patient throughput	0.214	0.38	0.289	0.546
Layout efficiency	0.031	0.034	0.031	0.029
ED performance	0.524	0.734	0.622	0.98

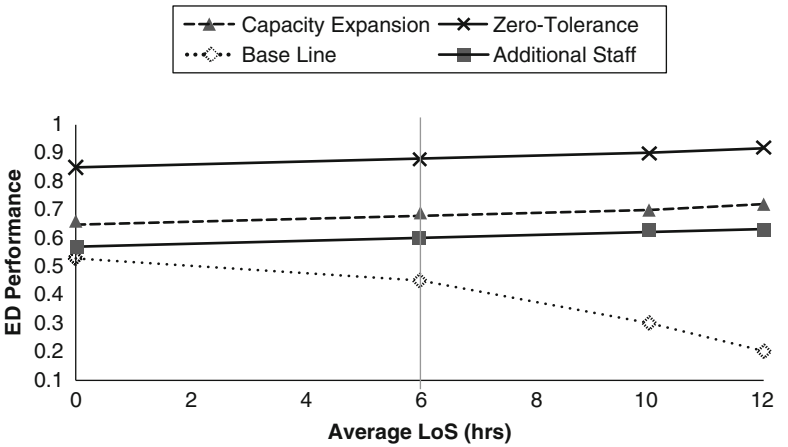


Figure 5.11 The change in ED performance with average LOS for all scenarios

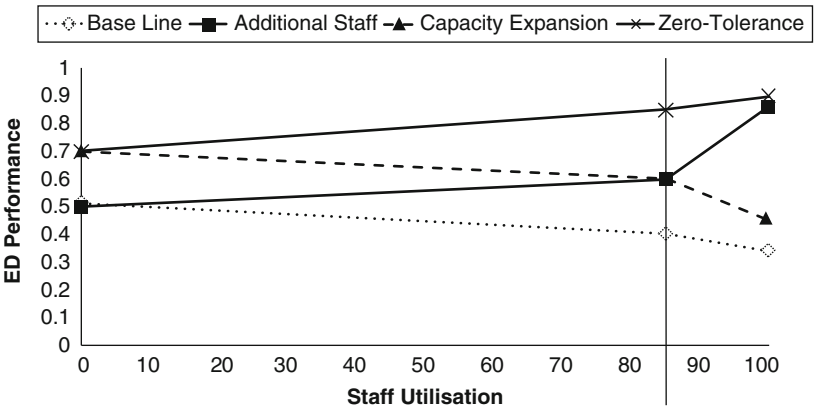


Figure 5.12 ED performance showing staff utilisation and burnout levels

target (i.e., zero-tolerance scenario) outperformed these more expensive scenarios (i.e., capacity expansion and additional staff).

Similarly, the zero-tolerance scenario gives the highest overall value for the ED performance in Figure 5.12 when a one-way sensitivity analysis is performed for staff utilisation. If the staff utilisation is less than 85% (staff burnout level), the capacity expansion alternative is preferable to adding more staff. However, if the burnout level is reached, additional staff is recommended. On the other hand, the performance of baseline deteriorates when staff are over-utilised to the point where they reach their burnout level at 85% utilisation. The sensitivity analysis suggests that this risk can be better mitigated by increasing the ED's staffing levels than by expanding its physical capacity, which does not reduce the workload for individual staff members.

This strategy enables ED management to use their resources (e.g., physical beds and trolleys) for helping not only patients awaiting admission but also newly arrived patients. Moreover, nursing staff that are frequently monitoring patients in critical medical conditions in corridors or trolleys can be reallocated to other areas in the ED.

Therefore, the hospital management team has suggested three initiatives to implement the recommendations of this study: first, executing an escalation plan that includes placing of additional beds on inpatient wards for moving patients quickly who are waiting in the ED; second, contracting cheaper beds in community care for elderly patients – this will shorten the average LOS in the whole hospital and consequently

more beds will be available for ED admissions; and finally, the hospital executives initiated lean training across all departments in the hospital. The aim of this training is to increase the coordination level between hospital units and achieve better utilisation of hospital resources. However, the implementation of these initiatives is still at its early stage and the 6-h target may or may not be achieved.

5.6 Conclusion

The proposed framework presented in this paper brought together scientists and clinicians to resolve many challenges that face health-care managers in the ED setting. Developing a detailed and comprehensive model that simulated a real process allowed managers to use a *what if* analysis approach to examine strategies and enhance their decision making.

The proposed framework has been well received by the ED managers and the hospital senior decision makers and was acknowledged as a sustainable tool to support their strategies. A number of factors have contributed to this positive perception from the management team. First, the development of a high-level process model before the development of the simulation model has greatly helped in the collection of relevant information on the operation of the system (i.e., data collection), and therefore reduced the effort and time consumed to develop the simulation model. The utilisation of IDEF for process modelling has not only improved the quality of the simulation model, but it has also enhanced the level of communication between decision makers and the staff through modelling the underlined workflow, decision points, and processes in a hierarchical form. Second, the integration between simulation modelling and BSC established a clear link between the strategic objectives of the hospital and the daily activities within the department, which gave decision makers deep insights regarding performance bottle-necks and potential corrective plans. Finally, the combination of MCDA tools along with simulation and BSC contributed significantly to the decision-making process by explicitly dealing with priorities and trade-offs between different performance measures.

The recommendations of the framework have been considered by the executive board of the partner hospital where the framework is currently used to model other hospital processes that affect the flow of patients to achieve the required alignment and coordination between hospital departments.

Although the proposed framework has successfully encompassed many factors that affect decision making, there is still room for

improvement. The key limitation of the proposed framework is the cost factor of the decision. Incorporating the cost element was not possible in this study because of two main reasons: (1) lack of cost-related information to support the analysis phase and (2) the variability in the cost model in various public hospitals in Ireland created a high level of complexity in modelling the financial element. The proposed framework is also limited to Discrete-Event Simulation, and other simulation and modelling methods such as system dynamics and agent-based simulation are emerging as potential tools for analysing the interconnected relationships between health-care components at the macro-level of the system.

References

Aguilar-Saven RS (2004) Business process modelling: Review and framework. *International Journal of Production Economics* **90**(**2**), 129–149.

Aidemark LG and Funck EK (2009) Measurement and health care management. *Financial Accountability and Management* **25**(**2**), 253–276.

Baesler FF, Jahnsen HE and Dacosta M (2003) Emergency departments I: the use of simulation and design of experiments for estimating maximum capacity in an emergency room. In *Proceedings of the 35th conference on Winter simulation* (Chick S, Sanchez PJ, Ferrin D and Morrice DJ, Eds) New Orleans, Louisiana Winter Simulation Conference, pp 1903–1906.

Balci O (1997) Verification validation and accreditation of simulation models. In *Proceedings of the 1997 Winter Simulation Conference* (Andradóttir S, Healy KJ, Withers DH and Nelson BL, Eds), pp 135–141, Atlanta, GA. IEEE Computer Society, Washington DC, USA.

Banks RL and Wheelwright SC (1979) Operations versus strategy–trading tomorrow for today. *Harvard Business Review* **57**(**3**), 112–120.

Barron FH and Barrett BE (1996) The efficacy of SMARTER – simple multi-attribute rating technique extended to ranking. *Acta Psychologica* **93**(**1**), 23–36.

Berger SH (2004) 10 ways to improve healthcare cost management. *Healthcare Financial Management Journal* **58**(**8**), 76–80.

Bond K et al (2007) Frequency, determinants and impact of overcrowding in emergency departments in Canada: a national survey. *Healthcare Quarterly* **10**(**4**), 32–40.

Brailsford S and Vissers J (2011) OR in healthcare: a European perspective. *European Journal of Operational Research* **212**(**2**), 223–234.

Clinton D, Webber SA and Hassell JM (2002) Implementing the balanced scorecard using the analytic hierarchy process. *Management Accounting Quarterly* **3**(**3**), 1–11.

Colaneri J (1999) A balanced scorecard approach to quality improvement in a renal transplant program. *Nephrology News & Issues* **13**: 19–26.

Colman R (2006) Improving the health care pulse. *CMA Management* **80**(**5**), 38–40.

Colquhoun GJ, Baines RW and Cro S (1993) A state of the art review of IDEF0. *International Journal of Computer Integrated Manufacturing* **6**(**4**), 252–264.

Cronin J (2003) The introduction of the Manchester triage scale to an emergency department in the Republic of Ireland. *Accident and Emergency Nursing* **11**: 121–125.

Da Silveira G and Slack N (2001) Exploring the trade-off concept. *International Journal of Operations & Production Management* **21**(**7**), 949–964.

Dittus RS, Klein RW, Debrota DJ, Dame MA and Fitzgerald JF (1996) Medical resident work schedules: design and evaluation by simulation modeling. *Management Science* **42**(**6**), 891–906.

Eccles RG and Pyburn PJ (1992) Creating a comprehensive system to measure performance. *Management Accounting* **74**(**4**), 41–44.

Elbeyli S and Krishnan P (2000) In-Patient flow analysis using ProModel™ simulation package. In *Proceedings of FREC SP00-02*, University of Delaware, Newark, DE.

Fitzgerald L, Johnston R, Brignall TJ, Silvestro R and Voss C (1991) *Performance Measurement in Service Businesses*. Chartered Institute of Management Accountants Publishing, London.

Forero R, Hillman K, Mccarthy S, Fatovich D, Joseph A and Richardson D (2010) Access block and ED overcrowding. *Emergency Medicine Australasia* **22**(**2**), 119–135.

Forster AJ, Stiell I, Wells G, Lee AJ and Van Walraven C (2003) The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine* **10**(**2**), 127–133.

Gao T, Gurd B and Baulderstone J (2006) Lives in the balance: managing with the scorecard in not for profit healthcare settings. In *Proceedings of the Australia and New Zealand Third Sector Research conference*, Australia, Flinders University, pp 1–26.

Graff L (1999) Overcrowding in the ED: an international symptom of health care system failure. *The American Journal of Emergency Medicine* **17**(**2**), 208–209.

Griffith JR (1994) Reengineering health care: management systems for survivors. *Hospital & Health Services Administration* **39**(**4**), 451–470.

Havey M (2005) *Essential Business Process Modeling*. O'Reilly Media, Sebastopol, CA.

Health Service Executive. (2010) Performance Monitoring Report. [WWW document] http://www.hse.ie/eng/services/Publications/corporate/performancereports/November_2010_Performance_Report.pdf (accessed October 2013).

Ingolfsson A, Erkut E and Budge S (2003) Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society* **54**(**8**), 736–746.

Jones ML and Filip SJ (2000) Implementation and outcomes of a balanced scorecard model in women's services in an academic health care institution. *Quality Management in Healthcare* **8**(**4**), 40–51.

Kaplan RS and Norton DP (1992) The balanced scorecard – measures that drive performance. *Harvard Business Review* **70**(**1**), 71–79.

Kaplan RS and Norton DP (1996) Using the balanced scorecard as a strategic management system. *Harvard Business Review* **74**(**1**), 75–85.

Kim SC, Horowitz I, Young KK and Buckley TA (1999) Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research* **115**(**1**), 36–46.

Kim SH and Jang KJ (2002) Designing performance analysis and IDEF0 for enterprise modelling in BPR. *International Journal of Production Economics* **76**(**2**), 121–133.

Kueng P (2000) Process performance measurement system: a tool to support process-based organizations. *Total Quality Management* **11**(**1**), 67–85.

Lebcir MR (2006) Health care management: the contribution of systems thinking. University of Hertfordshire, Business School Working Papers UHBS 2006–2007, Hatfield, UK.

LITVAK N, Van RIJSBERGEN M, BOUCHERIE RJ and VAN HOUDENHOVEN M (2008) Managing the overflow of intensive care patients. *European Journal of Operational Research* **185**(**3**), 998–1010.

LYNCH RL and CROSS KF (1995) *Measure Up! How to Measure Corporate Performance*. Blackwell, Oxford, UK.

MAPES J, NEW C and SZWEJCZEWSKI M (1997) Performance trade-offs in manufacturing plants. *International Journal of Operations & Production Management* **17**(**10**), 1020–1033.

MASSEY FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* **46**(**253**), 68–78.

MATHIAS JM (2001) Balanced scorecard helps ORs to weigh clinical, fiscal issues. *OR Manager* **17**(**9**), 18–22.

MELIONES J (2000) Saving money, saving lives. *Harvard Business Review* **78**(**6**), 57–65.

MOULLIN M (2004) Evaluating a health service taskforce. *International Journal of Health Hare Quality Assurance* **17**(**4–5**), 248–257.

NEELY A et al (2000) Performance measurement system design: developing and testing a process-based approach. *International Journal of Operations & Production Management* **20**(**10**), 1119–1145.

NEELY A, ADAMS C and KENNERLEY M (2002) *The Performance Prism: The Scorecard for Measuring and Managing Business Success*. Prentice Hall, London.

NIST. (1993) Federal information processing standards 183 – Integration definition for function modelling (IDEF0). NIST Computer Systems Laboratory. National Institute of Standards and Technology, Gaithersburg, MD.

PATEL B, CHAUSSALET T and MILLARD P (2008) Balancing the NHS balanced scorecard! *European Journal of Operational Research* **185**(**3**), 905–914.

PETERS K and RYAN H (1999) An integrated dialysis delivery network in Ontario. *The Journal of the Canadian Association of Nephrology Nurses and Technicians* **9**(**1**), 20–23.

SAATY TL (1990) How to make a decision: the analytic hierarchy process. *European Journal of Operational Research* **48**(**1**), 9–26.

SAUNDERS CE, MAKENS PK and LEBLANC LJ (1989) Modeling emergency department operations using advanced computer simulation systems. *Annals of Emergency Medicine* **18**(**2**), 134–140.

SCHAFERMEYER RW and ASPLIN BR (2003) Hospital and emergency department crowding in the United States. *Emergency Medicine* **15**(**1**), 22–27.

SEARCY DL (2004) Aligning the balanced scorecard and a firm's strategy using the analytic hierarchy process. *Management Accounting Quarterly* **5**(**4**), 1–10.

SENGE PM (1991) *The Fifth Discipline, The Art and Practice of the Learning Organization*. Doubleday/Currency, New York, NY.

SINREICH D and MARMOR Y (2005) Emergency department operations: the basis for developing a simulation tool. *IIE Transactions* **37**(**3**), 233–245.

STERMAN JD (1989) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(**3**), 321–339.

VALIRIS G, CHYTAS P and GLYKAS M (2005) Making decisions using the balanced scorecard and the simple multi-attribute rating technique. *Performance Measurement and Metrics* **6**(**3**), 159–171.

WU HY, TZENG GH and CHEN YH (2009) A fuzzy MCDM approach for evaluating banking performance based on Balanced Scorecard. *Expert Systems with Applications* **36**(**6**), 10135–10147.

6

Combining Data Mining and Discrete Event Simulation for a Value-Added View of a Hospital Emergency Department

R. Ceglowski, L. Churilov and J. Wasserthiel
Monash University, Melbourne, Australia

While simulation models have furthered understanding of the operations of emergency departments (EDs) and the dynamics of the ED within the health-care system, they only model patient treatment implicitly, tracing the paths patients follow through the ED. By identifying the core patient treatments provided by the ED and incorporating them into a Discrete Event Simulation model, this paper provides insight into the complex relationship between patient urgency, treatment and disposal, and the occurrence of queues for treatment. The essential characteristics of the presented model are used to indicate a generally applicable methodology for identifying bottlenecks in the interface between an ED and a hospital ward.

6.1 Introduction

Like similar facilities in the United Kingdom, Canada, The United States and several other countries, Australian emergency departments (EDs) are becoming the dominant source of primary care and one of the main routes for admission into hospitals. (Bolton *et al,* 2000). There have been large increases in presentations to EDs in recent years (Lane *et al,* 2000; Cooke *et al,* 2002), leading to longer waiting times, blockage, ambulance bypass and other issues that have the potential to affect the ability of the

Reprinted from *Journal of the Operational Research Society*, 58(2): 246–254, 2007, 'Combining Data Mining and Discrete Event Simulation for a Value-Added View of a Hospital Emergency Department', by R. Ceglowski, L. Churilov and J. Wasserthiel. With kind permission from Operational Research Society Ltd. All rights reserved.

ED to save lives. This patient increase has coincided with reduced healthcare budgets and pressure on hospitals to 'do more with less' (http://www.dhs.vic.gov.au/ahs/archive/emerg/ accessed 24 January 2006). The reasons for the increase in patient numbers are complex (Baggoley, 1998; Liaw *et al,* 2001) and will not be discussed here, but the fact remains that EDs have experienced pressure to improve efficiency. This paper describes the experience at the ED in one of Melbourne's teaching metropolitan hospitals, and then generalizes this experience through validation and verification in order to provide a new methodology for developing data-informed Discrete Event Simulations for more general systems settings.

6.1.1 The nature of Ed operations

ED operations can be put in context by looking at three main functions of EDs:

1. availability for patients seeking care, regardless of time of day and number of patients;
2. reception and management (including treatment) of patients (both urgent and non-urgent); and
3. disposition of patients once their treatment is complete.

These main functions of EDs address the relationship between the demand for the ED services and the supply of services by the ED.

The 'always open' and 'ready for any eventuality' nature of EDs make demand forecasting extremely complex and uncertain. While there is a well-recognized pattern to daily demand, the relative predictability of the average number of patient presentations each hour does not simplify demand estimation (Jiang *et al,* 2004). Even if patient numbers can be determined, the demographic mix of patients is usually wide and can vary, particularly in the instances of itinerant groups such as holiday resorts or sports stadiums in the region served by the ED. Patients may be of any age or either sex, have a full spectrum of ailments and injuries from life-threatening to minor and range from lucid to unresponsive (Coleridge *et al,* 1993; Liaw *et al,* 2001).

On the supply side, patient severity is first evaluated and classified into urgency levels. A qualified medical (triage) practitioner determines how long the patient may wait for medical attention. This 'triage' procedure is standard in almost all EDs worldwide and in Australia triage classifications range from 1 (most urgent) for 'immediate attention required' to 5 (least urgent) for 'attention within 2h' (Australasian School of Medicine, 1994). Patients are shown to the first available treatment site (most commonly a bed in the ED, but it may be a

'treatment chair' in the case of less serious presentations) and treatment initiated through the attention of qualified medical staff. During treatment phase, a decision is also made whether to admit the patient to hospital or whether to discharge them home. Specific 'disposal' codes are also used for patients who leave before treatment is complete and for patients who die in the ED.

6.1.2 Models of Ed operations

There has been a lot of research into ED operations over the past 30 years, so it is pragmatic to refer only to the most recent works. Fatovich (2002) summarized the developments in EDs, focusing on the Australian situation but also considering international trends. A series of articles describe measures taken in Australia to reduce patient time in the ED (Asplin *et al,* 2003; Richardson, 2003; Richardson *et al,* 2003). Jun *et al* (1999) surveyed the uses of Discrete Event Simulation over the past 20 years in healthcare clinics that range from individual practices to EDs. The review maintains that the breadth and scope of units within hospitals make it impossible to simultaneously address the inputs (patient flow and resource availability) and outputs (patient length of stay and resource utilization).

Simulation studies have formed a large component of the drive to understand and improve ED operations within the healthcare system. System dynamic simulations have looked at the interaction of ambulance services with the ED and the role of hospital policy in ED patient treatment time (Lane *et al,* 2000). Discrete Event Simulation of EDs have concentrated on either the scheduling of resources or the reduction of patient residence times (most commonly referred to as ED *length of stay*) to simulate activities at steady state. Discrete Event Simulation models have been able to assist with the identification of bottlenecks and other outcomes (Miller *et al,* 2004).

Discrete Event Simulation studies in EDs commonly break the ED into sub-units, assign patients to urgency categories and use these to prioritize access to resources. They generally approximate patient arrival rates and regulate patient flow by events such as completion of triage, admittance to an ED bed and review by doctors (Baesler *et al,* 2003; Brailsford *et al,* 2003; Mahapatra *et al,* 2003; Miller *et al,* 2003; Miro *et al,* 2003; Samaha and Armel, 2003; Wiinamaki and Dronzek, 2003; Connelly and Bair, 2004).

6.1.3 Reducing systems complexity through patient grouping

In trying to model the uncertain nature of ED operations, both healthcare and OR/MS experts aim to simplify the situation by grouping ED

patients. Clinicians have grouped patient cases under the Casemix principle, suggesting that similar cases will be treated alike and utilize a particular set of resources (Schneider *et al,* 1988; Jelinek, 1995; Bond *et al,* 1998). Casemix constitutes the clinicians' perspective on 'supply' processes but has received little attention in simulation studies, possibly because it suits situations where the range of cases is small, such as specialist departments, but has been less successful in EDs.

The majority of OR/MS specialists have used generalized distributions to describe arrival rates, lengths of stay and treatment times in simulation and optimization models (Bagust *et al,* 1999; Hoffenberg *et al,* 2001; Harper, 2002). Some Discrete Event Simulation studies have grouped patients to better emulate ED operations. In Sinreich and Marmor's work (2004), five patient groups were characterized across five hospitals based on interviews and time studies. Unique process charts were developed for each patient group. These included the duration of activities such as imaging and tests, and the frequency of connections between the activities. The paper concluded that ED processes are better characterized by patient group than by hospital. Evans *et al* (1996) had good success in duplicating patient length of stay by grouping patients into 13 categories according to diagnosis and building separate process flows for each group.

Overall, it is fair to note that the prevailing response to the complexity of the ED system has been to gather more data and build increasingly complex models—in essence, refining the tools without re-evaluating the logic employed. This contrasts with the original (Morse and Kimball, 1951) concepts of OR, which emphasized understanding of the data and its context in problem-driven analyses.

6.1.4 Redefining the OR problem

Demand and supply perspectives use patient throughput and resource utilization as important measures of ED performance, but they do not address the key value adding function of EDs—the transformation of ill and injured patients to a state of comparably good health. *The real OR problem in EDs is how to support patient treatment.* This does not mean that OR practitioners need to tell health professionals how to treat patients, in much the same way that Morse and Kimball (1951) did not attempt to tell pilots how to fly aeroplanes. Rather, the ED simulation problem may be reformulated as 'supporting the efficient and effective treatment of patients in the ED'. This problem formulation forces focus to shift from length of stay to 'access to treatment' and 'quality of treatment'. It is difficult and controversial to suggest measures of treatment quality because such a move implies usurping the clinician's prerogative, but it is reasonable to facilitate

patients' access to treatment. Access to treatment can be analysed by investigating the nature of queues into, and within, EDs.

> It is the objective of this paper to show how patient treatment groups may be extracted from existing data using non-parametric methods and then used to build a relatively simple, yet powerful simulation of the emergency department. This model intends to support timeous patient treatment through improving understanding of the relationship between patient urgency, treatment and disposal, and the occurrence of queues for treatment.

This paper intentionally limits description of algorithms, concentrating instead on the underlying logic and data analysis. There are three main parts to this paper. In the first part, the grouping of patients according to similarity of treatment is described. These patient groups are incorporated into a Discrete Event Simulation model in the second part of the paper. These parts systematically deal with the concept, realization and validation of (1) non-parametric patient grouping and (2) treatment-based Discrete Event Simulation. The third part of the paper provides insights that arose in the course of constructing the model, in particular, describing a generally applicable method for identifying bottlenecks between the ED and hospital wards.

6.2 Treatment-based patient groups

This section describes how Data Mining principles not readily accepted as part of the OR/MS practitioner's toolkit can be used to identify groups of patients. Data Mining has all the qualities required by OR/MS. It is the process by which repositories (often vast) of data may be systematically analysed to provide new insights. Descriptive statistics, and parametric and non-parametric methods are often combined to give non-intuitive grouping of variables in the data. The use of non-parametric methods for grouping of patients is relatively new (Isken and Rajagopalan, 2002; Ceglowski *et al,* 2004c), so this section will introduce the 'self-organized mapping' method and describe the results when it was applied to ED patient data.

6.2.1 Concept of data-driven patient grouping

The simulation studies described in the previous section were generally driven through consultation with experts about patient flows or by *in situ* observations. These studies are inherently limited by the extent of the

analysts' knowledge (knowledge sampling) and constrained by the data that can effectively be gathered (data sampling). The resulting models either aggregate patients according to expert or observers' perceptions or encapsulate extensive detail from a non-universal sample of patients. Data-driven modelling offers an alternative approach. No assumptions need to be made about the distribution of patient characteristics and *all* patients can be included in the analysis. Non-obvious groupings of patients can be identified by use of non-parametric techniques. These groupings, being based on data for every patient, suffer less from the depth and breadth limitations of traditional data- or knowledge-sampling approaches.

6.2.2 Concept realization

A generic Data Mining project proceeds in several stages (Berry and Linoff, 2000). Initially, the objectives and requirements are specified as they relate to the business. An understanding of the data is then developed, especially its source, scope, accuracy and completeness. The data is subsequently prepared through the elimination of obvious errors, coding into appropriate formats and other activities. With the problem understood and data prepared, modelling is undertaken through the application of (often several) algorithms that are appropriate to the problem. The models are evaluated (sometimes against one another) for technical accuracy and suitability. Findings are then deployed into the business, occasionally stimulating another project.

For EDs, the assertion that patient treatment is the key value-adding activity should be remembered as it provides the focus for the analyses described in this section. ED patient data were obtained from a major metropolitan hospital, which is a partner in this research project. The data consisted of 56 906 de-identified records of all ED presentations in 2002. The records contained demographic information plus details of the visit such as 'presentation problem', key time points and 'disposition'. Data on medical procedures undergone by patients were combined with the records of ED presentations so that each record contained demographic and visit information, as well as all medical procedures performed during that visit.

A number of data investigations were undertaken. These included the use of descriptive statistics, classification and regression trees and neural networks. These efforts culminated in the hypothesis that patients could be grouped according to the medical procedures most often performed together. A non-parametric method called SOM (Kohonen, 1995) was employed to detect whether such 'treatment groups' of patients could be identified.

The SOM method is driven by a non-parametric algorithm and relies on data, rather than domain-specific expertise. The objective of the method is similar to k-means clustering used by statisticians (Kennedy *et al,* 1998). It is to find groups of patients with minimal intra-group diversity and maximal inter-group separation. SOM generally employs large data sets, works well with many input variables and produces arbitrarily complex models unlimited by human comprehension (Kennedy *et al,* 1998). SOMs provide a visual understanding of patterns in data through a two-dimensional representation of all variables. The SOM algorithm repeatedly repositions records in the map until a classification error function is minimized. Records that have similar characteristics are adjacent in the map, and dissimilar records are situated at a distance determined by degree of dissimilarity. Viscovery SOMine, the software tool used in this analysis, employs a variant of Kohonen's Batch-SOM (Kohonen, 1995) guided by Ward's classic Hierarchical Agglomeration algorithm (Ward, 1963) to determine the optimal number of clusters.

Distinct groups of patients who underwent particular groups of procedures were identified (Figure 6.1). These groups of medical procedures represented the 'core treatment pathways' (Ceglowski *et al,* 2004c). Nineteen groups of procedures accounted for treatment of all patients whose treatment involved two or more procedures. (Although this was 78% of all patient presentations, it provides for the bulk of resource consumption because of the complexity of the treatment relative to patients who receive a single procedure or no procedures.) A sample of the grouping is given in Table 6.1. Each of the groups, or clusters, represents a pattern of treatment. Since there is a high degree of confidence in the quality of data collection at this ED, it is considered that these pathways depict an 'as is' picture of ED treatment.

6.2.3 Validation and verification

The resulting clustering model was *validated internally* in three ways. It was first evaluated by the project's ED clinician. Secondly, similar clusters were obtained using the *k*-means method (Ceglowski *et al,* 2004a). Lastly, the clusters were aligned with patient *presentation problem* through application of text mining techniques to the record of patient symptoms. This analysis revealed that patients were being consistently classified into treatment clusters appropriate to their presenting problem. For example, the patients who complained of a foreign object in an eye were given an eye examination procedure and were clustered into an eye treatment pathway; the patients with broken limbs received X-rays and Plaster of Paris and were clustered into a Plaster of Paris

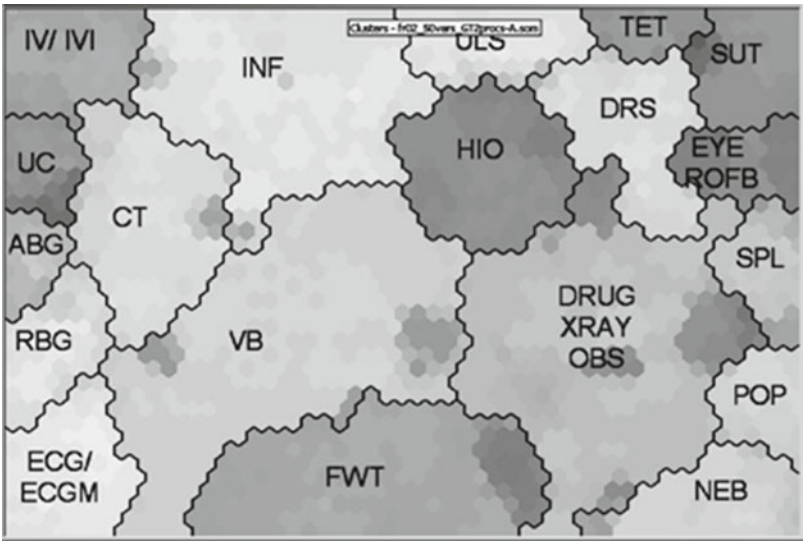


Figure 6.1 Screenshot of the SOM treatment clusters in Viscovery SOMine. The 50 input variables have been compressed into two dimensions and separated by boundaries. The clusters are labelled according to the procedure that is dominant (typically all patients in that cluster have that procedure and other, allied procedures). Treatments on the right-hand side relate to accident victims, with treatments including tetanus injections (TET), dressings (DRS), sutures (SUT), eye injuries (EYE, ROFB), splints (SPL) and Plaster of Paris (POP). Those on the left relate more to illness. Examples are treatments that include tests of arterial blood gases (ABG) or random blood glucose (RBG), monitoring of echocardiograms (ECG/ECGM) and intravenous drug infusion (IVI)

treatment pathway; while the patients with cuts and bruising received wound dressings and were clustered accordingly.

The treatment-based grouping methodology was then *validated externally* by investigating several data sets originating from various EDs over different time periods. Similar clustering results were obtained for patient samples from other years at the same hospital and for a number of other Melbourne metropolitan hospitals of comparable size and nature (Ceglowski *et al,* 2004b). Although the exact grouping of procedures differed from hospital to hospital, distinct, logical groupings were identified in every case.

In comparing treatment clusters across a number of different EDs, it was seen that the selection and ratio of procedures was sometimes different. This may be due to clustering, coding or treatment issues. Clustering differences usually be detected and overcome by adjustment of parameters. Coding issues relate to interpretation of procedure definitions.

Table 6.1 Selected treatment clusters showing proportion of patients (1.00 = 100%) in each cluster that undergo each of the medical procedures (Ceglowski *et al,* 2004c)

Procedure (abbreviation)	**Cluster**									
	Ill 1	**Ill 2**	**Ill 3**	**Ill 4**	**Ill 5**	**Ill 6**	**Asthma**	**Inj1**	**Inj2**	**Inj3**
Full ward test (FWT)	0.01		0.27	**0.98**	0.09	0.21	0.05	0.01	0.11	0.01
Head injury observation (HIO)	0.01								**1.00**	0.01
Electrocardiogram monitoring (ECGM)	0.01				**1.00**	0.12	0.07		0.13	
Electrocardiogram (ECG)	0.48	0.11	0.32	0.21	0.99	0.56	0.28	0.03	0.38	0.03
Random blood glucose (RBG)	0.82	0.06	0.80	0.56	0.87	0.79	0.40	0.10	0.50	0.12
Venipuncture (VB)										
CT scan (CT)	0.01					**0.99**	0.01		0.21	0.01
Ultrasound (ULS)										0.01
X-ray (XRAY)	0.41	0.39	0.44	0.35	0.73	0.55	0.49	0.37	0.46	0.89
Peripheral intravenous catheter (IV)	0.56	0.04	0.80	0.31	0.68	0.71	0.34	0.12	0.37	0.18
Observation (OBS)	0.74	**0.99**	0.54	0.66	0.47	0.64	0.72	0.57	0.22	0.55
Drug administration (oral, topical, etc) (DRUG)	0.52	0.75	0.60	0.63	0.61	0.54	0.71	0.57	0.58	0.48
Infusion IV fluid (ex blood products) (INF)	0.03	0.02	**0.99**	0.03	0.31	0.36	0.11	0.05	0.25	0.05
Nebulized medication (NEB)							**1.00**		0.01	
Dressing (DRS)		0.01				0.01		**1.00**	0.10	0.11
Suture (SUT)									0.08	0.02
Plaster of Paris (POP)										**1.00**
% *Patients*	14.5	22.0	10.5	8.4	5.8	4.2	4.6	3.7	3.5	2.0

Note the differences in the treatment of ill (first sis columns) and injured patients (last three columns).
Bold figures indicate the most significant procedures in the cluster.

These may vary from hospital to hospital, but can be addressed by clearer, agreed guidelines. Treatment differences may arise because of different patient profiles or innovation in treatment at certain campuses. Cross-campus comparisons provide the opportunity for knowledge management and knowledge transfer across campuses so that clinical decision support systems can be constantly updated. Associated with this knowledge management concept is the idea that treatment can be 'scored' and consistency of treatment measured for training and performance management (Gunning and Rowan, 1999).

The validation of results gave confidence that the groups of patients were clinically sound and provided a good representation of treatment activities within the ED. The next section describes how these patient treatment groups were incorporated into a Discrete Event Simulation model.

6.3 Discrete Event Simulation of patient treatment

Typically, ED simulation models are scaled-down representations of the facility. While scale models appeal because of their link to the observed world and because the one-to-one correspondence aids model verification, to quote Bernard Roy 'the map is not the territory' (Roy, 1989). Such a standard discrete event model had been previously built for the ED in this study and supplemented by formalized process diagrams of all ED activities (Djorhan and Churilov, 2003; Liew *et al,* 2003). These models do focus on patient flows within generalized processes, but fail to provide insight into variation of patient treatment and its effect on the ED system. The model described in this section seeks to complement conventional scale models by providing a high-level, abstracted view of ED operations. The treatment-focused Discrete Event Simulation approach encourages a systems-wide view by ignoring the (distracting) physical movement of patients and instead concentrating on how patient and treatment differences affect queue times. The treatment grouping introduced in the previous section is useful in this abstraction of ED utilization.

6.3.1 Concept of treatment-focused Discrete Event Simulation

Since patient registration and triage are well understood and largely optimized, it is reasonable to model only the stage between patient placement in a treatment bed and their physical departure from the ED. This simplifies the system to consideration of whether treatment sites (most commonly ED beds) are physically occupied. Queues develop if all sites are occupied. The benefit of this simplified 'state-based' view is that many variables become extraneous. For instance, patient bed times

may vary according to the people involved in the treatment (interns or experienced doctors, for example), admission of patients to virtual 'short stay units' within the ED so that the ED meets its performance obligations or other issues. Variability owing to doctors' different abilities and gaming are difficult to cater for in a conventional ED model. By using total bed time, it becomes unnecessary to gather these data, provided a high-level view of the ED is acceptable. If existing data distributions are sampled, rather than generalized (approximate) distributions of variables, the model can readily be used to retell the past.

6.3.2 Concept realization

The model was designed to generate a large variety of patient types according to urgency, treatment and disposal. The use of urgency and disposal variables were occasioned by Casemix studies that had indicated the importance of these on patient grouping (Cleary *et al,* 1994). Patients arrived in the ED bed queue at rates dictated by the data. They were apportioned urgencies and disposal within urgency according to historic distribution. Patients of each urgency/disposal type were streamed into one of the 20 treatment pathways according to the distribution profiles noted for that urgency/disposal/treatment combination. The 20 pathways consisted of 19 treatment pathways from non-parametric clustering and a single pathway for patients who had no procedures or only one procedure. A separate process was provided for patients classed as 'dead on arrival'.

At this point, the patient carried urgency, treatment and disposal labels that jointly defined the *patient type*. Patient type provided a framework for building the model and subsequent analysis. In theory, there were 401 patient types (possible combinations of five urgencies, 20 pathways and four disposal codes as well as patients categorized as 'dead on arrival'). In reality, though, data were sparse for some patient types (eg patients who died in the ED), so discrete distributions were specifically developed for 161 patient types (99% of patients) and generalized distributions were used for the remaining 1% of patient types that occurred rarely (Table 6.2).

Relationships between urgencies, treatment procedures and disposals that were hidden in the data became apparent during the model building process. As illustrated in Table 6.2, out of 401 theoretically possible patient types, only 251 occurred in the data and just 161 accounted for 99% of patients. Note also that 80% of patients could be accounted for by just 47 combinations of patient types based on the combination of urgency, treatment and disposal.

Table 6.2 Number of patient types existing in the data

% of patients	Number of urgency, disposal, treatment combinations
100	251
99	161
95	96
80	47
60	24

In the model, as in real life, patients queued for suitable beds if all beds were occupied. If access to treatment has been compromised at any time, a queue for beds develops. Waiting time may then exceed the thresholds stipulated by the national triage scale for a given triage category. Patient bed time (the total time for which they occupy a bed) was drawn from historic distributions for that urgency, treatment and disposal combination. Bed time was an input to the system and queue time was regarded as an output of the system. The model may be viewed as a sequence of sampled distributions (Figure 6.2).

Neither exact number of treatment sites within the ED nor the time it took for bed changeover from one patient to the next was known with any degree of certainty. The number of treatment sites was uncertain because staff adapt processes to facilitate patient throughput, particularly when the ED is reaching capacity. For instance, the number of 'treatment chairs' used for non-urgent patients may be increased, or patients awaiting transport to hospital wards may be placed in trolleys in corridors, freeing up treatment beds. Data about the number of patients actually in treatment at any given time were not available, so it was impossible to determine the extent to which such flexibility had been employed. Similarly, the available data did not capture how long it took to change a bed between patients. Generalized distributions had to be developed for bed turnover based on expert opinion. Variation of the median time in these distributions did not have a very significant impact on queue length.

While the logic of the model was straightforward, its construction in Simul8 (Version 11 from the Simul8 Corporation) was complicated by the presumption of patient flow inherent in the software. The model was implemented through sequences of virtual workstations and queues. These were used to query the data, update patient characteristics and ensure that patients received priority to beds based on urgency. Patients had zero residence time at all sites apart from beds or queuing for beds.

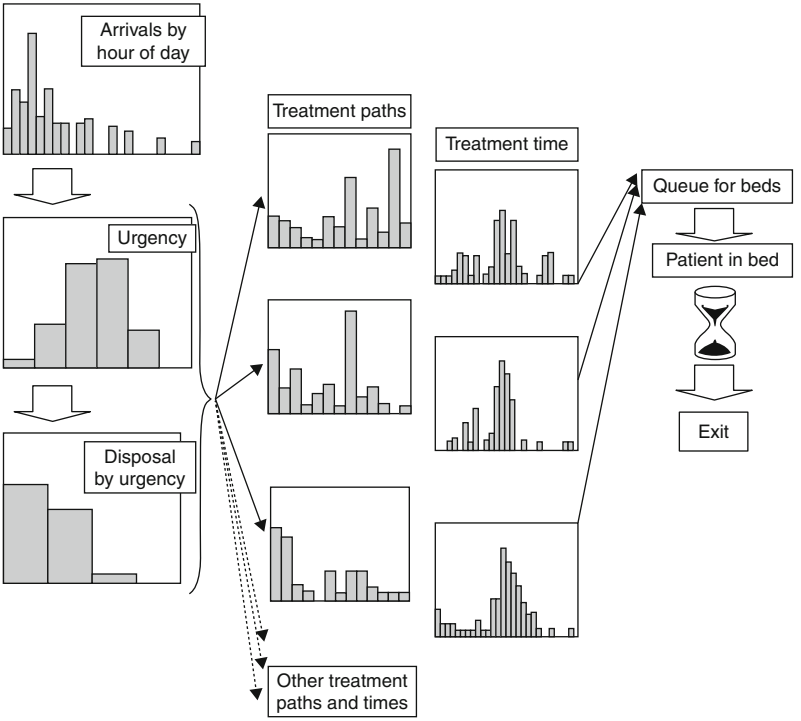


Figure 6.2 Schematic view of the simulation. Rather than following the physical movement of patients, the simulation tracks the state of ED treatment sites as being 'occupied' or 'free'. Queues result when all treatment sites are occupied (irrespective of other resource considerations). The bulk of the simulation is dedicated to allocation of appropriate urgency, disposal, treatment cluster and bed time labels to patients. While patient types are generic by urgency, disposal and treatment cluster, patient bed time is individual

The model had run times of minutes on a standard 'gaming' PC, despite the heavy sampling overhead.

6.3.3 Validation and verification

The infrequency of certain events suggested that a long warm-up period would be appropriate. The simulation model was warmed up for a year and run for a year. Both runs sampled historic data from the same year. Validation was performed by ensuring that the model generated the correct proportions of patient types and that the bed times were appropriate. Early incarnations of the model used generalized distributions

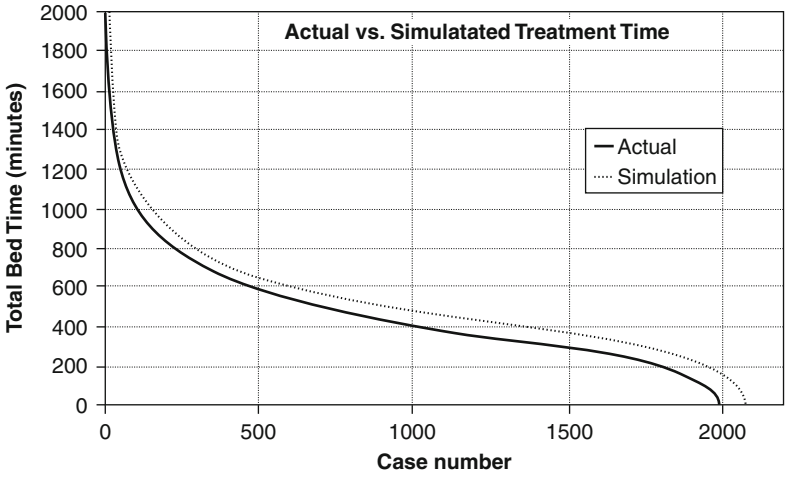


Figure 6.3 Comparison of patient bed times (sorted from highest to lowest) against actual times for patients of urgency 3 who will be admitted to a hospital ward for treatment of symptoms typically related to intake induced vomiting and diarrhoea. This graph intentionally depicts a run where simulated times do not match actual times

for bed times. These proved inadequate during validation, so historic data were sampled without approximating the time distributions. This strategy vastly improved the validation results.

Simulated bed times by patient type were graphically compared to actual data (Figure 6.3). This gave a quick and easy way of searching for errors and inconsistencies. Ultimately, the similarity of the simulated and actual data was confirmed using a Kolmogorov–Smirnov test. The Kolmogorov–Smirnov test detects differences in both the locations and the shapes of the distributions. It is based on the maximum absolute difference between the observed cumulative distribution functions for both samples. When this difference is significantly large, the two distributions are considered different.

Once the 'inputs' to the model were validated, verification of model 'output' was done. The queues that resulted when all beds were full were the output of the system since they were not explicitly modelled but arose from the interaction of patient bed times. The simulation model was verified by comparing the simulated number of patients queuing for beds and the time for which they queued with the actual number and time. Once again the results were graphically compared initially and later confirmed using a Kolmogorov–Smirnov test.

6.4 Summary and Discussion

This paper set out to describe a unique approach to simulation of hospital EDs. The intention was to facilitate simulation of ED treatment operations without onerous data collection requirements or excessive model complexity. Reformulation of the ED simulation problem from 'reduction in patient lengths of stay' to 'support for patient treatment in the ED' stimulated a fresh look at the simulation approaches used to date and their utilization of data.

A return to OR/MS roots suggested that historic data be explored for new understanding of ED operations in general and patient treatment in particular, rather than perpetuating the duplication and exhaustive extension of existing data. Data Mining led to patients being grouped by similarity of treatment. These groups were used in an abstract representation of the ED as a system that was either available or full. Factors that impacted on access to treatment could be identified by analyzing what happened when the system was full.

It was noticed that queues formed often but were generally not long either in duration or in number of patients. However, in several instances long queues formed that were analogous to those experienced in the ED when the system became blocked (unable to accept any new patients for treatment). In studying these instances, it became apparent that system blockage depended on the combination of patient types within the system. Patient types that were characterized by long bed times were implicated in the blockage, as would be expected. The unexpected finding was that the *combination* of number of patients and long bed time was also significant.

A simple weighting of the number of patients in each patient type with the average bed time for that patient type showed that certain patient types were occupying ED beds for a disproportionate time. Table 6.3 presents a sample of these. It is notable that the heaviest users were all awaiting admission to a hospital ward. The data records that the decision to admit these patients was made early in their treatment, but the ED was 'forced' to continue treating them because of the delay in moving them to a hospital ward. The treatment and symptoms of these patients give an indication of which wards were implicated in the admission delay.

It is worth mentioning that patient treatment clusters have applications apart from simulation. Several industrial engineering approaches that have been of only limited use in ED to date, such as Statistical Process Control, Fish Diagrams, Process Charts, Flow Diagrams and

Table 6.3 Patient types with the highest cumulative weighted impact on the ED

Urgency	Treatment cluster (typical symptoms)	Disposal	Average bed time in minutes (T)	Number of patients (N)	Weighted impact (T × N)
3	3 (Intake related vomiting, Diarrhoea)	Admit	476	2003	953 428
3	1 (general malaise)	Admit	438	1828	800 664
2	8 (cardiac or respiratory)	Admit	524	1516	794 384
2	1 (general malaise)	Admit	457	1252	572 164
4	3 (Intake related vomiting, Diarrhoea)	Admit	505	980	494 900
3	5 (collapse, mental)	Admit	541	877	474 457
3	4 (Fever Vomiting, Diarrhoea)	Admit	445	1012	450 340
4	1 (general malaise)	Admit	454	909	412 686
4	2 (Injury to limb or head)	Discharge	115	3368	387 320
4	20 (1 or no procedures)	Discharge	107	3538	378 566

The symptoms in brackets implicate which hospital wards might be involved in the admission.

Methods Engineering, can be facilitated by the treatment orientation of the clusters (Ceglowski *et al,* 2004c).

Hospital administrators will see the value that the process-based treatment groups have for activity-based costing (ABC). Whereas proxies such as 'doctors time' have been commonly used as cost drivers in ABC models, treatment groups now provide true 'activity' cost drivers for costing models. An illustrative costing model that uses Monte Carlo sampling of the range of procedure costs combined with information about procedures in treatment groups and patient presentations has been built as a part of this project. This model can be easily modified as a more accurate picture of procedure costs, procedure synergies and treatment-specific resource requirements become apparent.

6.5 Conclusion

The implication of certain hospital wards in queues for ED treatment has previously been elegantly described in a system dynamics model by Lane *et al* (2000), but it may be concluded from this paper that, although highly advisable, it is not essential to build a simulation model to identify

ED/ward bottlenecks. Any ED with adequate data can identify ED/ward bottlenecks in three steps. The first step involves the gathering and preparation of data. Secondly, patients must be divided into homogeneous groups and average bed times calculated (with due diligence to skewness and spread of the distribution). The urgency, treatment and disposal groups and non-parametric clustering used in this paper are just one way of doing this. Correlations between age and urgency may be included or treatment replaced by diagnosis and the coarse discharge/admit disposal replaced by more exact details of destination such as the designated hospital ward. The last step is a relatively simple analysis of the bed time for each of the patient groups and the number of patients in each of the groups. These can be used to provide a utility function of the demand placed on the system by each patient type. The end result will be an overview of ED workload and the most critical ED/hospital ward interactions.

Caution must be exercised in experimenting with the utility function described above without the benefit of simulation since averaged bed times are unlikely to produce accurate results in the inherently non-linear ED system (the complexity of the ED system is discussed in Ceglowski *et al,* 2005). The form of the treatment-focused simulation described in this paper makes it simple to test selectively the impact of faster admission procedures. The bed times for admitted patient may be reduced to emulate faster admission to ward (and so less waiting in the ED bed). Times may be reduced only for patients who have longer than average bed times, akin to a selective performance improvement for 'long stayers' in the ED.

The work described in this paper came at a time when the ED was being criticized for its patients' length of stay performance. The results proved useful in reinforcing the linkage between hospital policy and access to care, as opposed to addressing ED management in isolation. The financial and operational viability of 'bed changing teams' have been raised as a cost-effective way of reducing ED to ward transfer times. It can be seen from this paper that the reduction of these times for certain groups of patients or wards can free up beds in the ED, hence promoting access to treatment without physically increasing the number of beds in the ED.

References

Asplin BR *et al* (2003). A conceptual model of emergency department crowding. *Ann Emerg Med* **42**(2): 173–180.

Australasian School of Medicine (1994). National Triage Scale. *Emerg Med Australasia* **6**: 145–146.

Baesler FF, Jahnsen HE and DaCosta M (2003). The use of simulation and design of experiments for estimating maximum capacity in an emergency room. In: *Proceedings of the 2003 Winter Simulation, Conference*. INFORMS, New Orleans, Louisiana.

Baggoley C (1998). Primary care patients—what's the problem? *Emerg Med Australasia* **9**(2): 95–100.

Bagust A, Place M and Posnett JW (1999). Dynamics of bed use in accommodating emergency admissions: Stochastic simulation model. *BMJ* **319**(7203): 155–158.

Berry MJA and Linoff GS (2000). *Mastering Data Mining*. John Wiley and Sons: New York.

Bolton P, Mira M and Sprogis A (2000). Oranges are not the only fruit: The role of emergency departments in providing care to primary care patients. *Aust Health Rev* **23**: 132–136.

Bond M *et al* (1998). Urgency, disposition and age groups: A casemix model for emergency departments. *Emerg Med Austral*. **10**: 103–110.

Brailsford S, Churilov L and Liew SK (2003). Treating ailing emergency departments with simulation: An integrated perspective. In: *Proceedings of International Conference on Health Sciences Simulation*. The Society for Modeling and Simulation International: Orlando: FL.

Ceglowski A, Churilov L and Wasserthiel J (2004a) Lessons learnt from comparison of Self Organised and *k*-means clustering on hospital emergency department data. In: *Proceedings of International Conference on Optimization: Techniques and Applications*. Pacific Optimization Research Activity Group: Ballarat Australia.

Ceglowski A, Churilov L and Wasserthiel J (2004b). Process focused patient clustering for a hospital emergency department. In: *Proceedings of Casemix 2004,* Sydney.

Ceglowski A, Churilov L and Wasserthiel J (2004c). Data driven process modelling for a hospital emergency department. In: *Proceedings of Computer Supported Activity Co-ordination Workshop at ICEIS 2004,* Porto, Portugal.

Ceglowski A, Churilov L and Wasserthiel J (2004c). Process mining informed industrial engineering in hospital emergency departments. In: *Proceedings of APIEMS 2004,* Gold Coast, Australia.

Ceglowski A, Churilov L and Wasserthiel J (2005) Don't panic—prepare: Towards crisis-aware models of emergency department operations. *Clin Invest Med* **28**(5): 320–322.

Geary M *et al* (1994). The future of casemix in emergency medicine and ambulatory care. *Med J Aust* **161**(Suppl): S30–S33.

Colendge JT *et al* (1993). Profile of emergency department attendances. *Emerg Med Austral* **5**(1): 18–25.

Connelly LG and Bair AE (2004). Discrete event simulation of emergency department activity: A platform for system-level operations research. *Acad Emerg Med* **11**(11): 1177–1185.

Cooke MW, Wilson S and Pearson S (2002). The effect of a separate stream for minor injuries on accident and emergency department waiting times. *Emerg Med J* **19**(1): 28–30.

Djorhan V and Churilov L (2003). Business interactions in an acute emergency department in Australia: A clinical process modeling perspective. In: *Proceedings of the Sixth Pacific Asia Conference on Information Systems (PACIS 2002)*. The Japan Society for Management Information (JASMIN): Tokyo.

Evans GW, Unger E and Gor TB (1996). A simulation model for evaluating personnel schedules in a hospital emergency department. In: *Proceedings of the 1996 Winter Simulation Conference*. INFORMS.

Fatovich DM (2002). Emergency medicine. *BMJ* **324**(7343): 958–962.

Gunning K and Rowan K (1999). ABC of intensive care: Outcome data and scoring systems. *BMJ* **319**(7204): 241.

Harper P (2002). A framework for operational modelling of hospital resources. *Health Care Mngt Sci* **5**(3): 165–173.

Hoffenberg S, Hill MB and Houry D (2001). Does sharing process differences reduce patient length of stay in the emergency department?. *Ann Emerg Med* **38**(5): 533–540.

Isken MW and Rajagopalan B (2002). Data mining to support simulation modeling of patient flow in hospitals. *J Med Syst* **26**(2): 179–197.

Jelinek GA (1995). A Casemix Information System for Australian Hospital Emergency Departments. A Report to the Commissioner of Health, Perth, Western Australia.

Jiang H *et al* (2004). Predicting emergency department status. In: *Proceedings of International Conference on Optimization: Techniques and Applications*. Pacific Optimization Research Activity Group: Ballarat, Australia.

Jun JB, Jacobson SH and Swisher JR (1999). Application of discrete-event simulation in health care clinics: A survey, *J Opl Res Soc* **50**(2): 109–123.

Kennedy R *et al* (1998). *Solving Data Mining Problems Through Pattern Recognition.* Prentice-Hall: Englewood Cliffs, NJ.

Kohonen T (1995). *Self-organizing Maps*. Springer: Berlin.

Lane DC, Monefeldt C and Rosenhead J (2000). Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *J Opl Res Soc* **51**: 518–531.

Liaw S-T *et al* (2001). Emergency and primary care at a Melbourne hospital: Reasons for attendance and satisfaction. *Aust Health Rev* **24**(2): 120–134.

Liew D, Liew D and Kennedy MP (2003). Emergency department length of stay independently predicts excess inpatient length of stay. *Med J Australia* **179**: 524–526.

Mahapatra S *et al* (2003). Pairing emergency severity index 5-level triage data with computer aided system design to improve emergency department access and throughput. In: *Proceedings of the 2003 Winter Simulation Conference.* INFORMS: New Orleans, Louisiana.

Miller MJ, Ferrin DM and Szymanski JM (2003). Simulating Six Sigma improvement ideas for a hospital emergency department. In: *Proceedings of the 2003 Winter Simulation Conference*. INFORMS: New Orleans, Louisiana.

Miller MJ, Ferrin DM and Messer MG (2004). Fixing the emergency department: A transformational journey with EDSIM. In: *Proceedings of the 2004 Winter Simulation Conference*. INFORMS, Washinton, DC.

Miro O *et al* (2003). Analysis of patient flow in the emergency department and the effect of an extensive reorganisation. *Emerg Med J* **20**(2): 143–148.

Morse PM and Kimball GE (1951). *Methods of Operations Research.* Technology Press of Massachusetts Institute of Technology Wiley: Cambridge/New York.

Richardson DB (2003). Editorial: Reducing patient time in the emergency department. *Med J Aust* **179**: 516.

Richardson D *et al* (2003). Responses to access block in Australia. *Med J Aust* **178***:* 103–111.

Roy B (1989). Main sources of inaccurate determination, uncertainity and imprecision in decision models. *Math Comput Model* **12**(10/11): 1245–1254.

Samaha S and Armel WS (2003). The use of simulation to reduce the length of stay in an emergency department. In: *Proceedings of the 2003 Winter Simulation Conference*. INFORMS: New Orleans, Louisiana.

Schneider K, Lichtenstein J and Freeman J (1988). Ambulatory visit groups: An outpatient classification system. *J Ambul Care Mngt* **11**: 1–12.

Sinreich D and Marmor YN (2004). A simple and intuitive simulation tool for analyzing emergency department operations. In: *Proceedings of the 2004 Winter Simulation Conference*. INFORMS: Washington, DC.

Ward J (1963). Hierarchical grouping to optimise an objective function. *J Am Stat Assoc* **58**: 236–244.

Wiinamaki A and Dronzek R (2003). Using simulation in the architectural concept phase of an emergency department design. In: *Proceedings of the 2003 Winter Simulation Conference*. INFORMS: New Orleans, Louisiana.

7

Uncovering Effective Process Improvement Strategies in an Emergency Department Using Discrete Event Simulation

O. H. Choon[1], *Z. Dali*[1], *P. T. Beng*[2] *and C. P. Y. Magdalene*[2]
[1]*Centre for Health Services Research, Singapore Health Services Pte Ltd., Singapore*
[2]*Department of Emergency Medicine, Singapore General Hospital, Singapore*

This study aimed to develop a discrete event simulation (DES) model that realistically represented the process flow of selected non-fever patients in Singapore General Hospital (SGH) Emergency Department (ED). This model was then to be employed to evaluate the impact of different operational strategies on two performance measures, namely, time to first consult (TFC) and turnaround time (TT) of patients at ED. The DES model is used to represent ED sub-processes, including screening, registration, triage, consultation, observation ward visits, procedure room processes, laboratory and diagnostic radiological testing. Upon validating the model by comparing its performance projection with historical data, it was then employed to evaluate effectiveness of several new operational strategies. In the validation, we use the data of ED patients from May 2011 to July 2011 to characterize the load, arrival and resource-utilization pattern of its patients. Medical records of randomly sampled patients from the same period were also reviewed to evaluate the utilization levels of laboratory, diagnostic radiological, observation ward resources among ED patients. On the basis of DES model projections, redeployment of doctors with their hourly availability matching patient arrival pattern would yield the best improvement among the evaluated operational strategies. The manpower redeployment strategy required no additional man-hour compared to baseline requirement. In addition, via numerical results for a set of potential operational strategies, we show that there exists ample room for improvement in TFC and TTs of selected non-fever patients at SGH ED.

Reprinted from *Health Systems*, 3(2): 93–104, 2014, 'Uncovering Effective Process Improvement Strategies in an Emergency Department Using Discrete Event Simulation', by O. H. Choon, Z. Dali, P. T. Beng and C. P. Y. Magdalene. With kind permission from Operational Research Society Ltd. All rights reserved.

7.1 Introduction

The emergency department (ED) of a local public hospital has to contend with three major challenges. First, patient attendance of this ED had grown on average 3.6% annually over the last 10 years (2003–2012) (see Figure 7.1). Though this growth is likely to slow down in the near future due to opening of two new public hospitals in 2014 and 2018, respectively (Khaw, 2010), the challenge of designing effective infrastructure layout and process flow design, allocating appropriate manpower and equipment resources remains of great relevance to this ED. Moreover, elderly patients with chronic medical conditions, high rates of co-morbidities and home accident injuries (Foo *et al*, 2009; Ng, 2010; Paul *et al*, 2010) are expected to make up an increasing proportion of the future ED's workload due to Singapore's ageing population and their complex health-care needs. It has been reported that Singapore residents of age 65 and above are projected to triple to 900,000 in 2030 (The National Population & Talent Division, 2013). Together with the growing patient expectation, the anticipated changes in patient volume and profile need to be addressed in an effective and timely manner to ensure future ED patients continue to benefit from good quality of care and clinical outcomes.

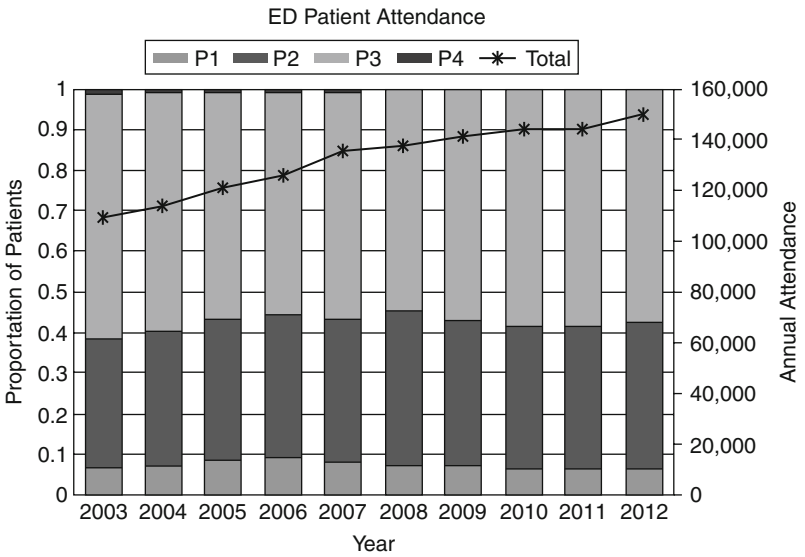


Figure 7.1 Historical ED attendance and proportions of four patient acuity categories (see Table 7.1 for their definitions)

The second challenge that is almost perpetually faced by ED is pertinent to the unfavorable working conditions attributed to high patient volume. If these conditions are not managed appropriately, they can compromise quality of care and may lead to medical errors, poor outcomes and even unnecessary patient deaths (Gordon *et al*, 2001; Committee on Pediatric Emergency Medicine, 2004). Lengthy waiting times that are attributed by high patient attendance also mean prolonged pain and suffering for patients and a risk that their condition may deteriorate while waiting to be seen. For ED staff, the high-stress environment induced by high patient volume working condition can also contribute to staff burnout and higher turnover rates, worsen staff productivity, and impair ongoing clinical teaching (Derlet, 2002). Therefore, it is clinically critical that effective facility layout plan, process flow design, and adequate manpower and equipment resources are deployed accordingly at the ED.

Lastly, ED of interest also faces a major strategic challenge due to its plan to construct a new facility to replace the existing one in 2020–2025. This offers the stakeholders an opportunity to implement resource-allocation plan, new facility layout and process flow design, which may further enhance experience of patients who visit the new facility. However, there are likely to be multiple permutations of new infrastructure layout and process flow designs, and resource-allocation plans that decision makers have to evaluate. The underlying complexity of such an evaluation exercise is also likely to be compounded by the uncertainty over the combined effects of different strategies.

Fortunately, the abovementioned challenges could be effectively addressed with the aid of discrete event simulation (DES). The latter offers a scientific evaluation platform for determination of good resource-allocation strategies, infrastructure and process flow design which can (1) enhance patient experience through reduction of waiting time to consultation and (2) cope effectively with future work load of ED via improved manpower and process planning according to patient volume. Such an approach is especially relevant when it is expensive, disruptive or impractical to implement changes to a complex and critical system such as the ED. Essentially, DES entails mathematical representation of a real-world system where changes in the system state occur at discrete points over time. These variables change at discrete instances (also known as event times) as the system state changes. On the basis of historical characteristics of a system and given a predetermined system design, DES projects the performance of the system involved. With this projection, inferences can then be made with regard to the operating characteristics of the real system and/or the effectiveness of one or more process improvement measures. Since

the introduction of DES in the late 1950s, it has evolved to become one of the most commonly used modeling techniques in industries that involve manufacturing/assembling processes, transportation of living and non-living objects, and provision of services. For example, automotive giants like Toyota and Volvo have applied DES to support their production plant design decision-making processes while United Stated Marine Corps and Canadian Forces have employed DES to enhance their manpower training continuum, manpower and equipment deployment decision-making processes. DES has also been used in EDs for the analysis of patient flows and throughput time (Mahapatra *et al*, 2003; Ruby *et al*, 2003; Simon *et al*, 2003; Soemon & Hiroko, 2004; Hupert *et al*, 2007; Duguay & Chetouane, 2007; Hung *et al*, 2007; Levin *et al*, 2008; Hung & Kissoon, 2009; Maull *et al*, 2009; Wang *et al*, 2009; Bair *et al*, 2010; Brenner *et al*, 2010; Joshi & Rys, 2011; Rosmulder *et al*, 2011; Bowers *et al*, 2012; Gul & Guneri, 2012; Klein & Reinhardt, 2012; Paul & Lin, 2012; Rasheed *et al*, 2012; Wang *et al*, 2012; Zeng *et al*, 2012; Abo-Hamad & Arisha, 2013; Ashour & Kremer, 2013; Day *et al*, 2013; Hamrock *et al*, 2013; Lim *et al*, 2013). There are also reported applications of DES in estimating future capacities of new ED facilities (Allan & Rainer, 2003; Felipe *et al*, 2003; Hoot *et al*, 2008). Moreover, the flexibility of DES has also enabled researchers to integrate simulation with other techniques like Integer Linear Programming, Six Sigma, data mining and so on for specific applications (Felipe *et al*, 2003; Martha *et al*, 2003; Martin & David, 2003; Ceglowski *et al*, 2007).

7.1.1 ED modeling overview

Similar to previous work in the literature, this project entails development of a DES model to represent the interaction of entities (patients, doctors, nurses etc) in the ED of interest. Essentially, this simulation model serves as an evaluation platform in determination of good resource-allocation strategies that keep time to first consult (TFC) and turnaround times (TTs) of patients involved to the minimal. Before the task of simulation model development is embarked, it is crucial to establish understanding of the journey of patients at the ED via process mapping. As such, we describe in the following section the sequence of visits made by the patients to different stations within the ED of interest during each episode of care from the time they enter ED and to when they leave the purview of ED.

7.1.2 Process flow

The ED of interest is demarcated into fever and non-fever areas that serve patients with fever and no fever, respectively. Within each of these two areas, it is further demarcated into triage, registration, critical care, observation ward and waiting areas. Within the non-fever area,

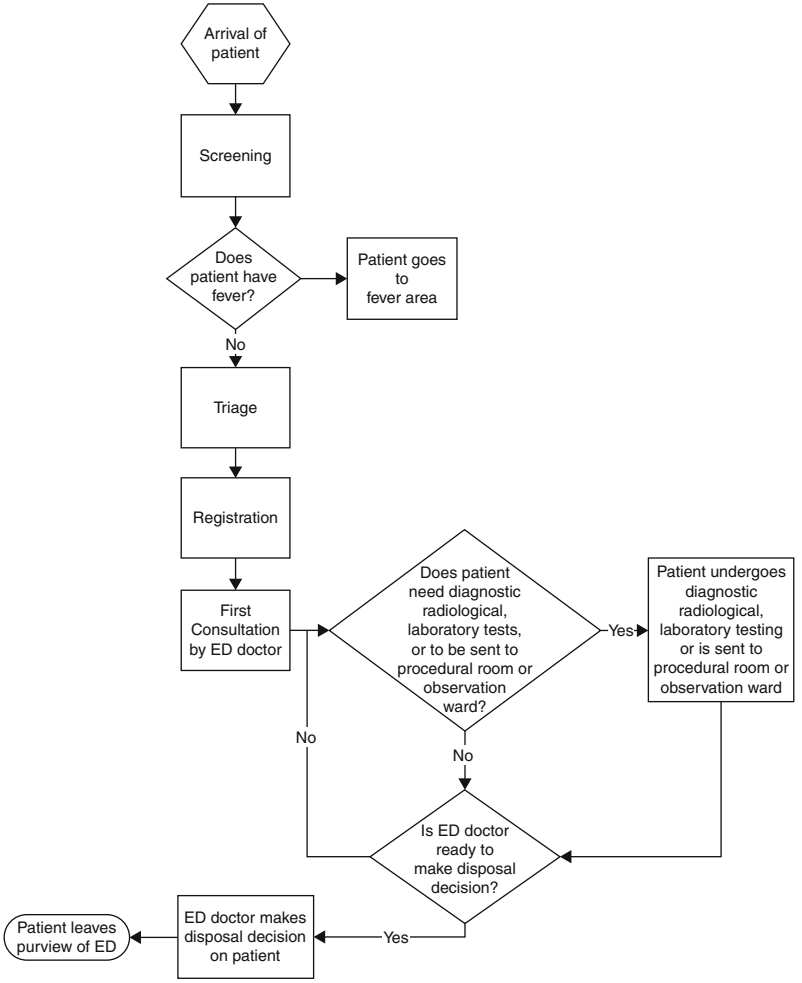


Figure 7.2 Simplified schematic diagram of patient flow at ED

there is an additional area that serves only resuscitation and critically ill patients. Figure 7.2 illustrates a simplified journey of an ED patient from the point of arrival to the point when the patient leaves the purview of ED. Upon arrival at the ED via various transportation modes including ambulance, taxis and private vehicles, a patient will be screened for fever and queried about the patient's recent travel history. With the exception of resuscitation and critically ill patients (who

will be subsequently assigned with patient acuity category of highest possible priority), all fever patients will be directed to the fever area. All resuscitation and critically ill patients regardless of them having fever or not will be attended by ED doctors at the resuscitation area after their screening processes. In contrast, all other post-screening patients will follow through the triage, registration, consultation and other downstream processes in either the fever and non-fever areas. Typically, ED patients with local identification or work permit documents will proceed to the triage station after the screening process. Those without identification or work permit documents or do not carry with them these documents will need to go through the registration process before the triage process. The latter entails undertaking several key activities by a trained nurse on an ED patient and these activities may include medical history taking, diagnostic measurements of blood pressure and pulse rate, blood capillary and urine tests. Depending on their medical conditions, the triage process may require the ED patients to have the x-ray imaging or electrocardiogram done before they are first seen by an ED doctor. Essentially, every ED patient will be assigned with one of the patient acuity categories listed in Table 7.1 after the triage process. To this end, it is important to note that patients who arrive to ED via ambulances will typically have their triage done within the critical care area. Patients assigned with patient acuity category of P2 after triage will be attended by ED staff at the critical care area. In contrast, P3 and P4 patients will be directed to a common waiting area where they will wait for their respective turns to be seen by ED doctors who are assigned to attend patients of these acuity categories.

The registration process basically entails verification and electronic entry of patient particulars by the registration counter staff. It must be highlighted that the registration process need not be done in the presence of the ED patients involved. Friends and relatives who accompanied the patients to ED can register them on the patients' behalf. In addition, patients who arrive at ED via public ambulances will have their registrations done by the ambulance paramedics. Generally, the latter will proceed with the registration processes after handing over the patients from their respective ambulances to the ED staff.

Depending on the medical conditions of a patient, there are multiple possible scenarios after a patient is seen by an ED doctors for the first time. The simple scenario entails the ED doctor discharging the patient so that the latter leaves the purview of ED (see Figure 7.2). The more complex scenarios will see ED doctors ordering laboratory or diagnostic radiological tests on the patients, sending patients to the observation

Table 7.1 Patient acuity categorization at the ED

Patient acuity category	Patient conditions
P1	These are patients who are either in a state of cardiovascular collapse or in imminent danger of collapse and would therefore be required to be attended to without delay. They would most likely require maximum allocation of staff and equipment resources for initial management
P2	These patients are ill, non-ambulant and in various forms of severe distress. They would appear to be in a stable state on initial cardiovascular examination and are not in danger of imminent collapse. The severity of their symptoms requires very early attention. Failure to do so will likely result in early deterioration of their medical status. They would be trolley-based
P3	These patients have acute symptoms but are ambulant and have mild to moderate symptoms and require acute treatment, which will result in resolution of symptoms over time
P4	These are non-emergency patients. They should not be presented at ED and should be managed in a primary health-care setting, such as by general practitioners or in polyclinics. They may have an old injury or condition that has been present for a long time. They do not require immediate treatment. There is no immediate threat to their life or limb.

ward or undertaking minor medical procedure on the patients at the procedural room. Patients who are directed to observation ward will be periodically reviewed by ED staff. Typically, ED doctors rely on results of laboratory or diagnostic radiological tests in their disposal decision-making processes. As such, ED patients who require laboratory or diagnostic radiological tests at ED may have more than one consultation round with the ED doctors before they leave the purview of ED. However, there are also scenarios where ED doctors make disposal decisions on their patients without waiting for the results of their laboratory or diagnostic radiological tests.

On the whole, ED doctors can make three disposal decisions on their patients. First, the ED patients are discharged with or without specialist outpatient referrals. Second, the patients are referred to doctors of other disciplines where on-call doctors from departments of Ear Nose and Throat, Orthopedic, Cardiology and so on are activated to attend to these patients. Third, the patients are admitted to the hospital inpatient wards. Once ED doctors make the disposal decisions, the patients involved leave the purview of the ED.

7.1.3 Modeling Scope and Features

In this study, we focused on the processes that took place between the arrival of patients at the screening station and the instants when their respective disposal decisions were made by the ED doctors. Since non-fever P2 and P3 patients constituted more than 90% of ED's daily attendance, we limited our modeling scope so that the model would simulate the aforementioned processes of these selected ED patients. Essentially, this study aimed to evaluate the impact of different manpower allocation plan on non-fever P2 and P3 patients with the aid of a DES model. The impact was assessed in terms of patient's TFC and TT. TFC of an ED patient is defined as the time interval between the starts of screening process and first consultation (i.e., patient sees ED doctor for the first time during an ED visit) of the patient. On the other hand, an ED patient's TT is defined as the time interval between the start of screening process of the patient and the instant when disposal decision is made on the patient by ED doctor. To this end, it is also important to note that the patient attendance at the ED on Monday was distinctly higher than other days by up to 21.8% (see Figure 7.3). Thus, both our DES model and subsequent analysis were developed and done, respectively, with distinction of different patient arrival pattern for Monday and other days.

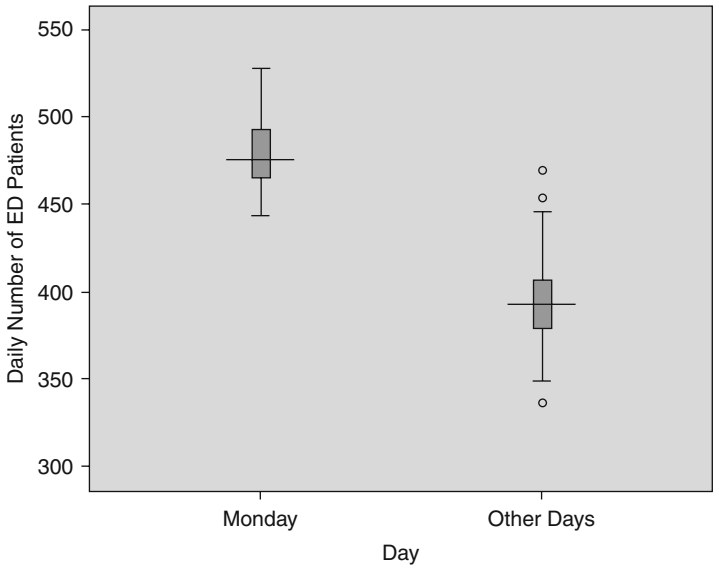


Figure 7.3 Daily load of ED

To simulate the journey of selected patients who visit the ED, key processes that were needed to be accounted for in the simulation model included the arrival of all ED patients for screening, triaging, registration, doctor consultation (first and review consultations), patients undergoing diagnostic measurements, tests, surgical procedures, patients being admitted to observation ward and physical movements of patients between stations in ED. In addition, the profiles of the ED patients had to be accounted accordingly. These profiles include proportion of patient types (based on patient acuity category), proportion of patients who require registration before their triaging, proportion of patients who require laboratory and/or radiological diagnostic tests after triage, doctor consultation and so on. Relevant data that characterized the above processes and patient profiles needed to be collated to serve as input to the simulation model. Table 7.2 lists the key data that were captured for our model development. Generally, these data were retrieved from electronic database, time motion study (pioneered by industrial engineer Frederick Winslow Taylor) or based on ED staff input and manual activity records. One assumption that we made in our simulation model development was that the collated data such as service rate distributions of ED staff at various ED stations, ED patient profiles, their resource needs were representative of underlying processes and patient characteristics, and they were time-independent.

With all the necessary data collated, a DES model that represented the flow of patients at ED was developed using SIMUL8 2011 Professional (SIMUL8 Corp.) installed in a computer with Intel®Core™ i5 M580 @ 2.67 GHz CPU, 1.17 GHz, 2.92 GB RAM and Windows XP OS system. Essentially, this software permits modular representation of processes where the sequential flow of patients to these processes can be predefined accordingly to simulate the historical-based patient flow in the ED. The warm-up and result-collection periods of resultant DES model were fixed at 2 and 8 weeks, respectively.

On the basis of the reported simulation model validation techniques in the literature (Sargent, 1999), two key steps were performed to ascertain operational validity of resultant simulation model. First, we verified from the model output that the chronological sequences of process flow of all patients were in accordance with the characteristics of the patients. Second, we performed a subjective validation study (Sargent, 1996) where we verified that the profiles of several key process characteristics and outcome measures generated by the simulation model matched closely to those of ED's historical data. For example, the patient arrival pattern generated in the DES model was found to match that of historical data (see Figure 7.4). Similarly, the cumulative probability distributions of TFC and

Table 7.2 List of data and their sources for DES model development

Process	Data	Source	Remark
Arrival	Inter-arrival interval	ED electronic database	Three-month historical data
	Transportation mode	ED electronic database	Three-month historical data
Screening	Service interval	Time–motion study	280 samples were analyzed
	Manpower schedule	Input of ED nurse clinician	NA
	Proportion of fever patients	ED electronic database	Three-month historical data
Triage	Service interval	Time–motion study	261 samples were analyzed
	Test requirements[a]	ED electronic database	1940 samples were analyzed
	Manpower schedule	Input of ED nurse clinician	NA
	Proportions of P1–P4 patients	ED electronic database	Three-month historical data
Registration	Service interval	Time–motion study	240 samples were analyzed
	Manpower schedule	Input of ED nurse clinician	NA
	Proportion of foreign patients	ED electronic database	Three-month historical data
Consultation	Service interval	Time–motion study	268 samples were analyzed
	Test requirements[a]	ED electronic database	1940 samples were analyzed
	Manpower schedule	Input of ED nurse clinician and roster record	Two weeks of doctor roster were analyzed
	Test TTs[a]	ED electronic database	1940 samples were analyzed
Observation ward	Proportion of Emergency Observation Ward (EOW) and non-EOW patients[a]	ED electronic database	1940 samples were analyzed
	TT[a]	ED electronic database	1940 samples were analyzed
	Manpower schedule	Input of ED nurse clinician	NA
Procedure room	Proportion of patients	ED electronic database and manual record	Three-month historical data
	TT	ED manual record	Three-month historical data

[a] These data were captured electronically in free text. As such, manpower were deployed to review the electronic case notes of ED patients and a total of 1940 were sampled.

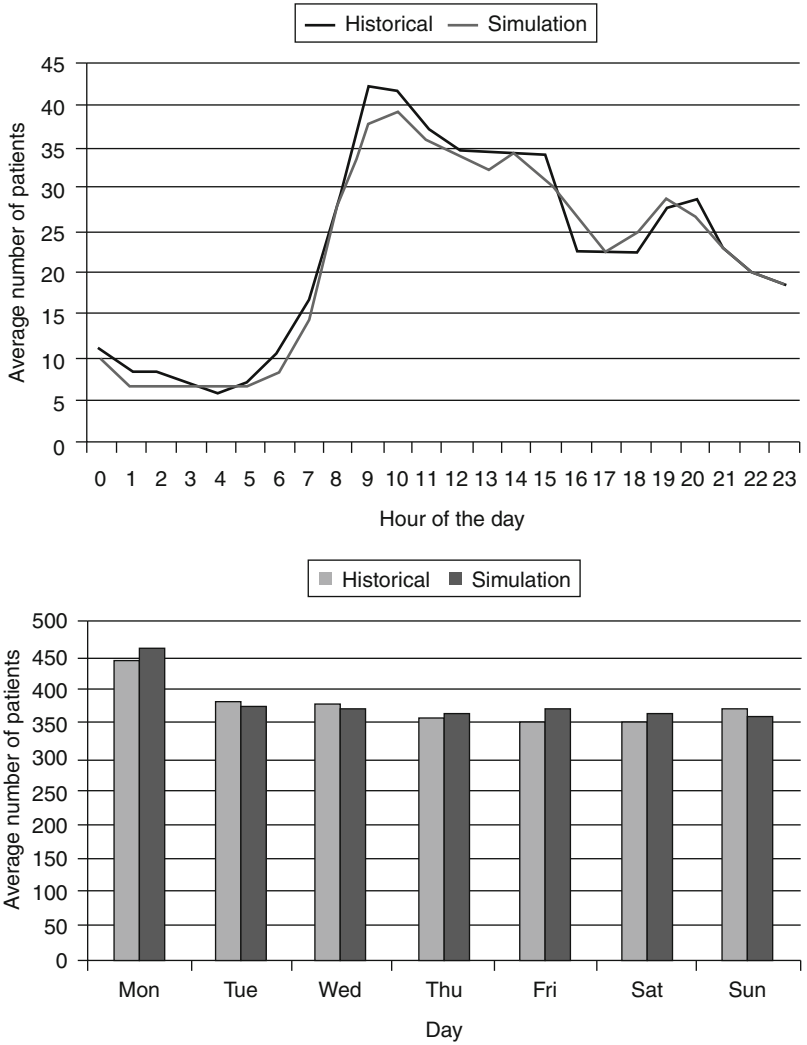


Figure 7.4 Arrival patterns of ED patients

TT of non-fever P2 and P3 patients projected by the DES model were also found to match closely to those of historical data (see Figures 7.5 and 7.6). On the basis of the overall results of the abovementioned validation steps and the endorsement by ED manager, it was affirmed that the simulation model represented the process flow within the ED of interest realistically.

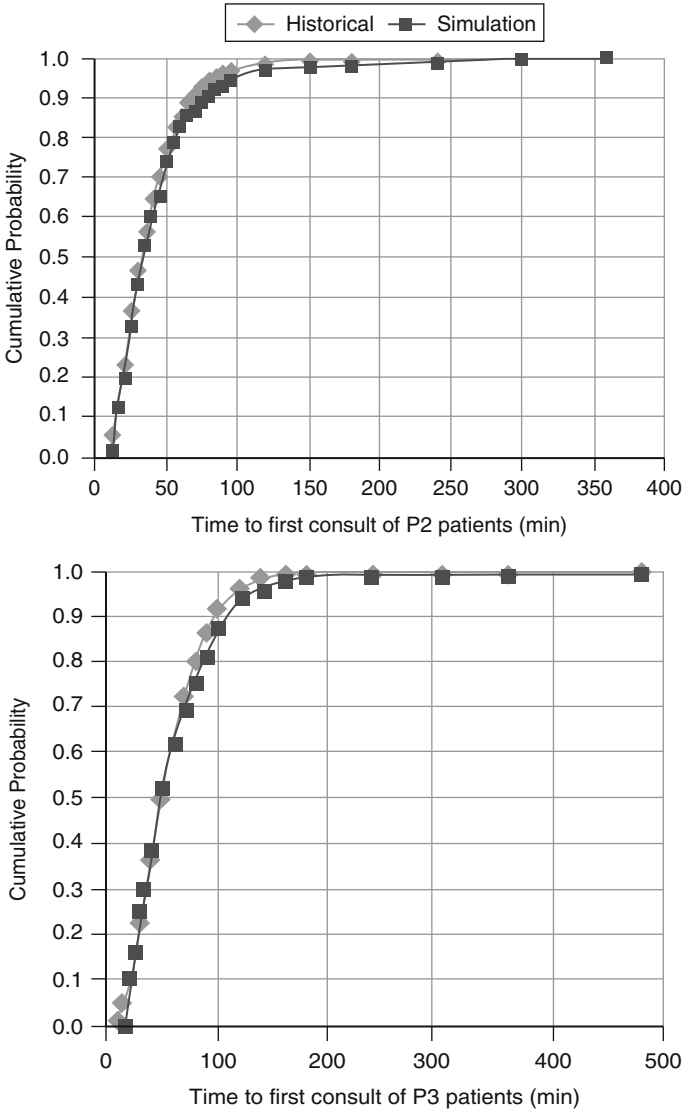


Figure 7.5 TFC times of non-fever P2 and P3 patients

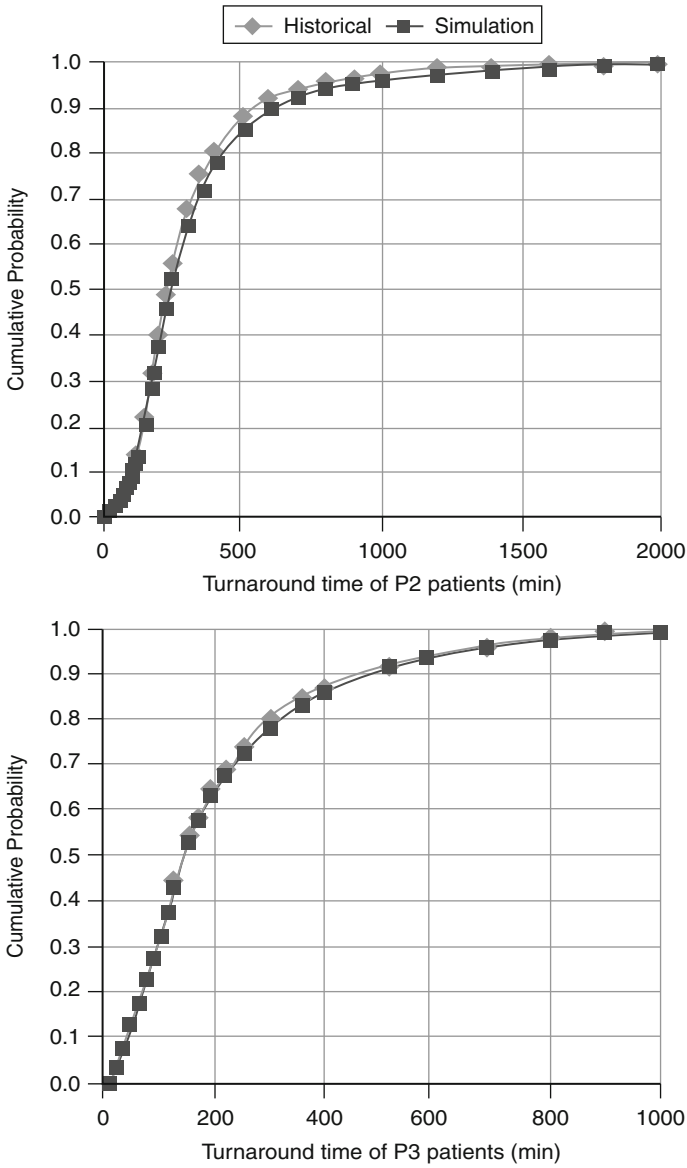


Figure 7.6 TTs of non-fever P2 and P3 patients

7.1.4 What-if Analyses

With the validated DES model, we went on to employ it to project the impact of several new operational strategies in our what-if analyses. These new strategies were conceptualized after evaluation of all postulations on how TFC and TT of ED patients could be improved. In general, they could be broadly classified into three types of strategies based on how they affected the doctor manpower schedules, laboratory test and diagnostic radiological test processing times. Details of new operational strategies evaluated in our study were described in the following sections.

7.2 New Doctor Manpower Schedules

When the patient arrival pattern was plotted against the existing doctor manpower allocation schedule as shown in Figure 7.7, it was evident that the doctor manpower availability profiles did not match the patient arrival pattern. As such, it was postulated that TFC and TT of ED patients could be reduced if the doctor manpower availability profiles could better match the patient arrival pattern than the existing arrangement. In total, three new doctor manpower scheduling strategies were evolved and their respective impact on TFC, TT of ED patients and overall manpower requirements were projected with aid of the DES model.

The first new doctor manpower schedule was derived by matching doctor manpower availability profiles more closely with patient arrival patterns (as illustrated in Figure 7.8) with minimal change in the total man-hours. The total man-hours based on the new (baseline) manpower scheduling plans for P2 doctors are 188.0 (188.5) and 181.0 (183.1) on Monday and non-Monday, respectively, while those of P3 doctors are 75.0 (75.5) and 79.0 (78.9) on Monday and non-Monday, respectively. With this first new doctor manpower scheduling plan, it was projected by the DES model that the median (95th percentile) TFC of P2 and P3 patients could be reduced by 21% (20%) and 20% (18%), respectively, compared to those of baseline arrangement. The TT of P2 and P3 patients could also be reduced by 22% (21%) and 20% (18%), respectively.

From the first doctor new manpower schedule, it was noted that the total man-hours of P3 doctors on Monday were lower than those of non-Monday even though historical data indicated a higher P3 patient volume on Monday than non-Monday by about 10%. As a result, the second new doctor manpower schedule was a revision of the first new

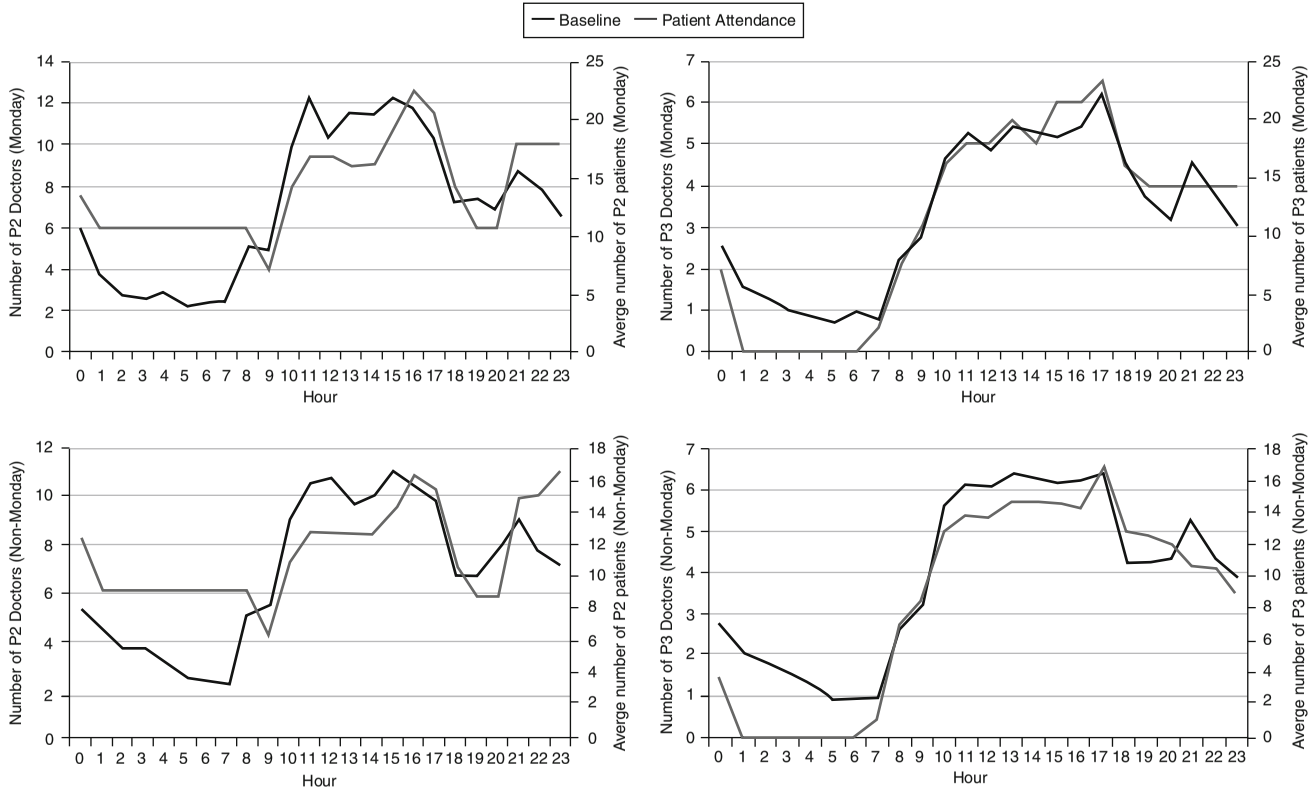


Figure 7.7 Baseline hourly profiles of P2 and P3 patients and doctors

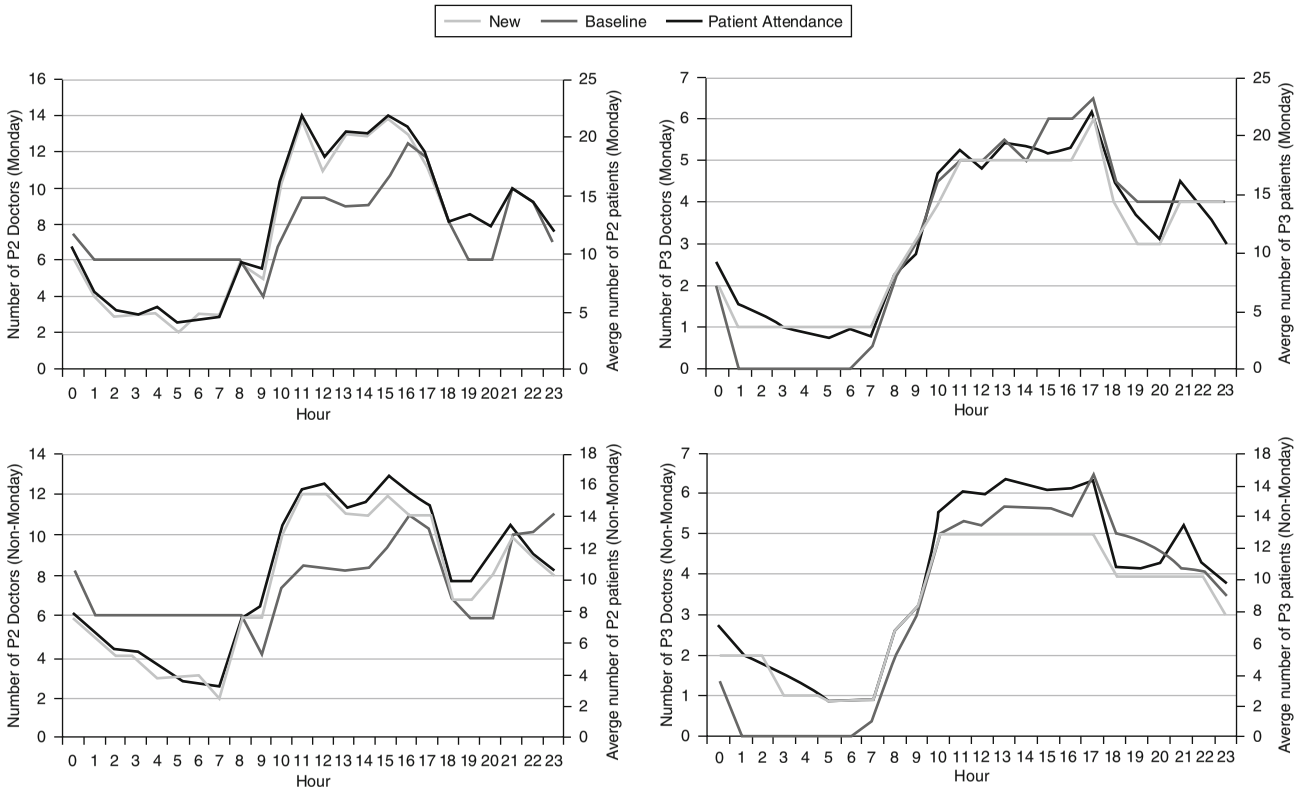


Figure 7.8 Baseline and new hourly profiles of P2 and P3 patients and doctors

one where the manpower schedules of P3 doctors on Monday and non-Monday were adjusted as shown in Figure 7.9 so that the total man-hours of P3 doctors were 86.0 and 77.0, respectively, with the former being higher by about 11.7%. In addition, the total weekly man-hours of the second new manpower schedule differed from that of first new schedule by only 1. With this new doctor manpower scheduling plan,

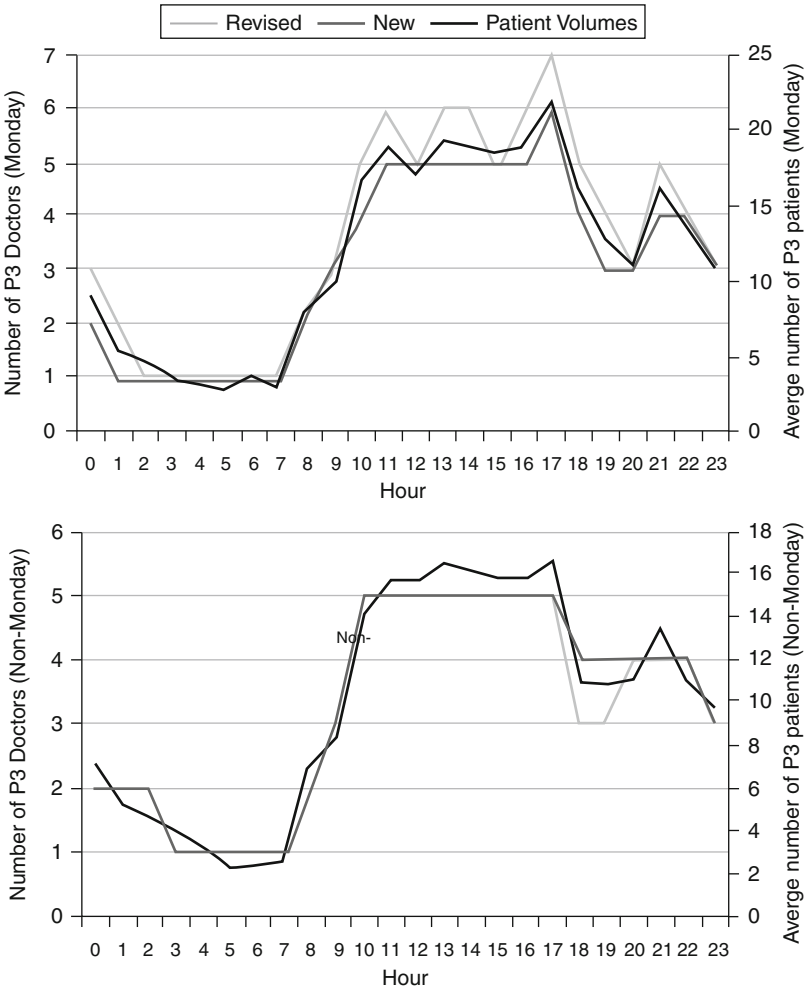


Figure 7.9 Hourly profiles of P3 patients and doctors in the second new manpower schedules

it was projected by the DES model that the median (95th percentile) TFC and TT of P3 patients could be reduced by 22% (20%) and 25% (21%), respectively, compared to the baseline performance. Note the manpower profiles of P2 doctors in both the first and second new doctor manpower schedules were the same.

Essentially, the third new doctor manpower schedule was a variation of the second one where the former accounted for the existing shift hours of the doctors. At the ED of interest, the existing shift periods of doctors were 8 am–5 pm, 10 am–6 pm, 3 pm–11 pm, 4 pm–12 am, 5 pm–1 am, 9 pm–8 am. On the basis of these shift hours, the second new doctor manpower schedule was derived by assigning doctors to these shifts so that the sum of hourly absolute deviation in the number of available doctors in second new doctor manpower schedule from that of first schedule was minimized. Thus, if N_{ih} denotes the number of P2 or P3 doctors in hour h ($h \in \{0, 1, \ldots, 23\}$) of schedule i ($i \in \{1, 2\}$), the number of P2 or P3 doctors in the third new doctor manpower schedule (i.e., N_{2h}) was derived by determining the values of N_{2h} which yielded the minimum function of $z = \Sigma_h |N_{1h} - N_{2h}|$. This assignment was done with the aid of the Solver add-in in Microsoft Excel 2010 and the details of the third new doctor manpower schedule were tabulated in Table 7.3. The profile of P2 doctors in the second doctor manpower schedule on non-Monday was illustrated in Figure 7.7. With this third new doctor manpower scheduling plan, it was projected by the DES model that the median (95th percentile) TFC of P2 and P3 patients could be reduced by 15% (14%) and 18% (14%), respectively, compared to those of baseline arrangement. The TT of P2 and P3 patients could also be reduced by 17% (14%) and 13% (14%), respectively. Relative to the baseline arrangement, the second new doctor manpower schedule resulted in a net reduction of 4 man-hours per week. It was important to note the aforementioned DES model was run with a constraint that a maximum of five P3 patients could be seen by doctors at any time to reflect the infrastructural limit of five consultation rooms for non-fever P3 patients at the ED of interest.

7.3 New laboratory test processing times

On the basis of historical data, 54.8% of ED patients required collection of their blood samples for laboratory tests. The median (90th percentile) time interval between ordering of laboratory test by ED doctors and the point when the corresponding laboratory test results are ready for review is 1.0 (3.3) h. This time interval encompasses several activities

Table 7.3 Numbers of P2 and P3 doctors in existing shifts in the second new doctor manpower schedule

Day	Doctor types	Shift hours						Total man-hours
		21–8	8–17	10–18	15–23	16–0	17–1	
Monday	P2	3	6	5	3	1	4	188.0
		(3)	(5)	(5)	(3)	(2)	(4)	(188.5)
	P3	1	3	2	1	1	1	78.0
		(1)	(2)	(3)	(0)	(1)	(2)	(75.5)
Non-Monday	P2	3	6	5	1	1	4	175.0
		(3)	(5)	(5)	(2)	(2)	(4)	(183.1)
	P3	1	3	2	0	1	3	86.0
		(1)	(2)	(2)	(0)	(2)	(2)	(78.9)

Note: Numbers in brackets denote baseline assignment (rounded to nearest integer).

that include blood sample collection, delivery of samples to laboratory, preparation of blood samples before the analytical processes which are typically done by laboratory equipment. There were postulations that reduction of the aforementioned time interval via procurement of more laboratory equipment, onsite laboratory test devices and/or expansion of laboratory staff could potentially improve the TT performance of ED. But there was uncertainty over the impact of such a move relative to the new doctor manpower schedules. As such, six scenarios that depicted different degrees of laboratory test processing time reduction were simulated in the DES model and their respective impact on TT of ED patients were subsequently projected with aid of the DES model. Four scenarios emulated the presence of onsite laboratory test devices that, respectively, allowed 10, 20, 50 and 100% of all laboratory tests order by ED doctors to be done onsite so that their corresponding times were reduced by 80% relative to baseline requirements. The other two scenarios emulated the procurement of more laboratory equipment and/or expansion of existing laboratory staff so that the laboratory test processing time was reduced by 20 and 50%, respectively, relative to baseline requirements.

On the basis of the DES model's projection, the impact of these six scenarios on the TT of P2 and P3 patients was summarized in Table 7.4. The best two improvements in TT were achieved when 100% of all laboratory tests were performed onsite and 50% reduction in all laboratory test processing times was achieved. The amount of improvements that could be potentially realized for P2 patients in these two scenarios

Table 7.4 Projected reduction in TT of patients under selected scenarios

Scenarios	Percentage reduction in TT of P2 patients		Percentage reduction in TT of P3 patients	
	Median (in percentage)	95th Percentile (in percentage)	Median (in percentage)	95th Percentile (in percentage)
10% of all laboratory tests were performed onsite[a]	4	2	3	4
20% of all laboratory tests were performed onsite[a]	8	11	6	10
50% of all laboratory tests were performed onsite[a]	14	15	19	17
100% of all laboratory tests were performed onsite[a]	19	20	27	24
20% reduction in all laboratory test processing times	9	3	6	11
50% reduction in all laboratory test processing times	16	17	22	20
20% reduction in all diagnostic radiological processing times	7	1	8	1
50% reduction in all diagnostic radiological processing times	12	7	12	14

[a]Presence of onsite laboratory test devices allowed 80% reduction in laboratory test processing time relative to the baseline performance.

was comparable to those of aforementioned three new doctor manpower schedules. In terms of the TT of P3, the amount of improvement was 6–9% lower in the new doctor manpower schedules relative to the two best scenarios. Evidently, new doctor manpower schedules were more attractive since they do not incur additional significant capital

expenditures as in the two best scenarios. Moreover, the third new doctor manpower schedule is expected to incur less operational cost compared to the baseline expenditure.

7.3.1 New diagnostic radiological processing times

According to the ED's historical data, 53.0% of ED patients required x-ray imaging. The median (90th percentile) time interval between ordering of X-ray imaging was made by the ED doctors till the point when the corresponding X-ray images are ready for review are about 12 (41) min and 22 (35) min respectively for non-fever P2 and P3 patients. This time interval encompasses activities like physical movement of patients to the locations with x-ray imaging equipment and preparation of patients for x-ray imaging. In our study, two scenarios that emulated the procurement of additional x-ray imaging equipment and/or expansion of existing diagnostic radiological staff so that the diagnostic radiological processing time was reduced by 20 and 50%, respectively, relative to the baseline performance were also evaluated with the aid of our DES model.

With the aid of the DES model, it was projected that the two aforementioned scenarios would result in a reduction in the median and 95th percentile TT of P2 and P3 from 1 to 14% (see Table 7.4 for details). These magnitudes of improvement were clearly less compared to those of similar scenarios when all laboratory test processing times were reduced by 20 and 50%. Essentially, this could be attributed to the fact the baseline average laboratory test processing time was three to five times longer than diagnostic radiological processing times.

7.4 Conclusions

DES modeling offers a scientific and evidence-based framework that allows decision makers to evaluate the effectiveness of different strategies without committing excessive resources and/or disrupting the workflow within a system of interest. Scientific evaluation of the impact of changes in processes is useful in health-care service delivery, particularly for those changes that entail costly interventions or measures and there is significant uncertainty over their effectiveness in achieving the desired outcomes. As health-care professionals work continuously to enhance health-care operational efficiency in response to rising health-care costs and patient expectations, DES models are excellent scientific frameworks that enable decision makers to realize health-care process enhancement via cost-effective strategies. This is evident in our study

where several measures that could reduce TFC and TT performance of ED were identified with aid of DES model. On the basis of the model's projection, it was then affirmed that matching manpower profile with that of patient arrival pattern is the most cost-effective way of managing both TFC and TT performance. Although reduction of laboratory test processing times could yield marginally better TT performance compared to manpower reallocation, the former strategies require significant capital investments of both hardware and staff in order to achieve the required reduction of laboratory test processing times. Overall, the aforementioned findings not only enabled the Singapore General Hospital (SGH) ED management team to gain practical insights into the relative cost effectiveness of various process improvement strategies which had previously been discussed with their respective impact on TFC and TT performance being constantly questioned. In addition, these findings had also given SGH ED management team a heads up on how the new ED facility in 2020–2025 should be operationally managed to realize the best possible patient experience.

Acknowledgements

This work was supported by Singapore Ministry of Health Healthcare Quality Improvement and Innovation Fund (HQI2F 2011/02). The authors are also indebted to the students from School of Business & Accountancy, Ngee Ann Polytechnic and all staff of Singapore General Hospital Emergency Department for their assistance in the data collection effort of this study.

References

Abo-Hamad W and Arisha A (2013) Simulation-based framework to improve patient experience in an emergency department. *European Journal of Operational Research* **224**(**1**), 154–166.

Allan W and Rainer D (2003) Using simulation in the architectural concept phase of an emergency department design. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1912–1916.

Ashour OM and Kremer GEO (2013) A simulation analysis of the impact of FAHP-MAUT triage algorithm on the emergency department performance measures. *Expert Systems with Applications* **40**(**1**), 177–187.

Bair AE, Song WT, Chen YC and Morris BA (2010) The impact of inpatient boarding on ED efficiency: a discrete-event simulation study. *Journal of Medical Systems* **34**(**5**), 919–929.

Bowers J, Ghattas M and Mould G (2012) Exploring alternative routes to realising the benefits of simulation in healthcare. *Journal of the Operational Research Society* **63**(**10**), 1457–1466.

Brenner S, Zeng Z, Liu Y, Wang JW, Li JS and Howard PK (2010) Modeling and analysis of the emergency department at University of Kentucky Chandler Hospital using simulations. *Journal of Emergency Nursing* **36**(**4**), 303–310.

Ceglowski R, Churilov L and Wasserthiel J (2007) Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society* **58**(**2**), 246–254.

Committee on Pediatric Emergency Medicine. (2004) Overcrowding crisis in our nation's emergency departments: is our safety net unraveling? *Pediatrics* **114**(**3**), 878–88.

Day TE, Al-Roubaie AR and Goldlust EJ (2013) Decreased length of stay after addition of healthcare provider in emergency department triage: a comparison between computer-simulated and real-world interventions. *Emergency Medicine Journal* **30**(**2**), 134–138.

Derlet RW (2002) Overcrowding in emergency departments: increased demand and decreased capacity. *Annals of Emergency Medicine* **39**(**4**), 430–432.

Duguay C and Chetouane F (2007) Modeling and improving emergency department systems using discrete event simulation. *Simulation-Transactions of the Society for Modeling and Simulation International* **83**(**4**), 311–320.

Felipe B, Hector J and Mahal D (2003) The use of simulation and design of experiments for estimating maximum capacity in an emergency room. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1903–1906.

Foo CL, Chan KC, Goh HK and Seow E (2009) Profiling acute presenting symptoms of geriatric patients attending an urban hospital emergency department. *Annals Academy Medicine Singapore* **38**(**6**), 515–520.

Gordon JA, Billings J, Asplin BR and Rhodes KV (2001) Safety net research in emergency medicine: proceedings of the Academic Emergency Medicine Consensus Conference on 'The Unraveling Safety Net'. *Academic Emergency Medicine* **8**(**11**), 1024–1029.

Gul M and Guneri AF (2012) A computer simulation model to reduce patient length of stay and to improve resource utilization rate in an emergency department service system. *International Journal of Industrial Engineering – Theory Applications and Practice* **19**(**5**), 221–231.

Hamrock E, Paige K, Parks J, Scheulen J and Levin S (2013) Discrete event simulation for healthcare organizations: a tool for decision making. *Journal of Healthcare Management* **58**(**2**), 110–124.

Hoot NR *et al* (2008) Forecasting emergency department crowding: a discrete event simulation. *Annals of Emergency Medicine* **52**(**2**), 116–125.

Hung GR and Kissoon N (2009) Impact of an observation unit and an emergency department-admitted patient transfer mandate in decreasing overcrowding in a pediatric emergency department: a discrete event simulation exercise. *Pediatric Emergency Care* **25**(**3**), 160–163.

Hung GR, Whitehouse SR, ONeill C, Gray AP and Kissoon N (2007) Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. *Pediatric Emergency Care* **23**(**1**), 5–10.

Hupert N, Hollingsworth E and Xiong W (2007) Is overtriage associated with increased mortality? Insights from a simulation model of mass casualty trauma care. *Disaster Medicine and Public Health Preparedness* **1**(**Suppl 1**), S14–S24.

Joshi AJ and Rys MJ (2011) Study on the effect of different arrival patterns on an emergency department's capacity using discrete event simulation. *International Journal of Industrial Engineering – Theory Applications and Practice* **18**(**1**), 40–50.

Khaw BW (2010) New Khoo Teck Puat Hospital helps ease bed crunch at hospitals: Khaw [WWW document] http://www.channelnewsasia.com/stories/singaporelocalnews/view/1071284/1/.html, accessed 25 July 2010.

Klein MG and Reinhardt G (2012) Emergency department patient flow simulations using spreadsheets. *Simulation in Healthcare* **7**(**1**), 40–47.

Levin SR *et al* (2008) Optimizing cardiology capacity to reduce emergency department boarding: a systems engineering approach. *American Heart Journal* **156**(**6**), 1202–1209.

Lim ME, Worster A, Goeree R and Tarride JE (2013) Simulating an emergency department: the importance of modeling the interactions between physicians and delegates in a discrete event simulation. *BMC Medical Informatics and Decision Making* **13**(**1**), doi:10.1186/1472-6947-13-59.

Mahapatra S, Koelling C, Patvivasiri L, Fraticelli B, Eitel D and Grove L (2003) Pairing emergency severity index5-level triage data with computer aided system design to improve emergency department access and throughput. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1917–1925.

Martha C, Ronald G, Richard L and Abdullah I (2003) A simulation-ILP based tool for scheduling ER staff. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1930–1938.

Martin M and David F (2003) Simulating six sigma improvement ideas for a hospital emergency department. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1926–1929.

Maull RS, Smart PA, Harris A and Karasneh AAF (2009) An evaluation of 'fast track' in AE: a discrete event simulation approach. *Service Industries Journal* **29**(**7**), 923–941.

The National Population and Talent Division. (2013) Population White Paper: A sustainable population for a dynamic Singapore [WWW document] http://202.157.171.46/whitepaper/downloads/population-white-paper.pdf, accessed 1 February 2013.

Ng G (2010) Home accidents befall more elderly victims [WWW document] http://www.asiaone.com/News/AsiaOne+News/Singapore/Story/A1Story20100924-238848.html, accessed 1 October 2010.

Paul JA and Lin L (2012) Models for improving patient throughput and waiting at hospital emergency departments. *Journal of Emergency Medicine* **43**(**6**), 1119–1126.

Paul P, Heng BH, Seow E, Molina J and Tay SY (2010) Predictors of frequent attenders of emergency department at an acute general hospital in Singapore. *Emergency Medicine Journal* **27**(**11**), 843–848.

Rasheed F, Lee YH, Kim SH and Park IC (2012) Development of emergency department load relief area-gauging benefits in empirical terms. *Simulation in Healthcare – Journal of the Society for Simulation in Healthcare* **7**(**6**), 343–352.

Rosmulder RW, Krabbendam JJ, Kerkhoff AHM, Houser CM and Luitse JSK (2011) Computer simulation within action research: a promising combination for improving healthcare delivery? *Systemic Practice and Action Research* **24**(**5**), 397–412.

Ruby B, Darell S, Wendy A and Mary H (2003) The use of simulation to evaluate hospital operations between the emergency department and a medical telemetry. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp 1887–1893.

Sargent R. (1996) Some subjective validation methods using graphical displays of data. *Proceedings of the Winter Simulation Conference*, pp 345–351.

Sargent R. (1999) Validation and verification of simulation models. *Proceedings of the Winter Simulation Conference*, pp. 39–48.

Simon S, Wendy A and Darrell S (2003) The use of simulation to reduce the length of stay in an emergency department. In *Proceedings of the 2003 Winter Simulation Conference* (Chick S, Sanchez P, Ferrin D and Morrice D, Eds), pp. 1907–1911.

Soemon T and Hiroko S (2004) Functional analysis for operating emergency department of a general hospital. In *Proceedings of the 2004 Winter Simulation Conference* (Ingall R, Smith M and Peters B, Eds), pp 2003–2011.

Wang JW, Li JS, Tussey K and Ross K (2012) Reducing length of stay in emergency department: a simulation study at a community hospital. *IEEE Transactions on Systems Man and Cybernetics Part A – Systems and Humans* **42**(**6**), 1314–1322.

Wang T, Guinet A, Belaidi A and Besombes B (2009) Modelling and simulation of emergency services with ARIS and Arena. Case study: the emergency department of Saint Joseph and Saint Luc Hospital. *Production Planning & Control* **20**(**6**), 484–495.

Zeng Z, Ma XJ, Hu Y, Li JS and Bryant D (2012) A simulation study to improve quality of care in the emergency department of a community hospital. *Journal of Emergency Nursing* **38**(**4**), 322–328.

8

Improving the Design and Operation of an Integrated Emergency Post via Simulation

N. J. Borgman, M. R. K. Mes, I. M. H. Vliegen and E. W. Hans
University of Twente, Enschede, The Netherlands

In the Netherlands, patients with an acute care demand after office hours often wrongly choose to visit the emergency department (ED), while they could have visited the general practitioners' post (GPP). This may lead to overcrowding and increased costs. In this paper, we focus on an Integrated Emergency Post (IEP) at a Dutch hospital, where the ED and the GPP have been merged into a single point of access for patients. To find the optimal process design for this new IEP, we use computer simulation incorporating patient preferences. We define many potential interventions, and compare these by categorizing and grouping them, and sequentially withdrawing ineffective interventions, while accounting for possible interaction effects. Results show a sustainable solution for all stakeholders involved, reducing patients' length of stay up to 17%. Based on these results, an intervention has been trialled in practice, showing a decrease in patient LOS.

8.1 Introduction

The delivery of acute care to patients is highly complex and constrained by limited resources. In addition, many of the involved processes are stochastic and there are interactions both within and external to the acute care providers. Consequently, the effects of organizational changes tend to be unpredictable. Therefore, operations research methods, such as computer simulation, are suitable for prospectively evaluating changes such as alternative resource allocations. Indeed, simulation is used often

Reprinted from *Journal of Simulation*, 9: 99–110, 2015, 'Improving the Design and Operation of an Integrated Emergency Post via Simulation', by N. J. Borgman, M. R. K. Mes, I. M. H. Vliegen and E. W. Hans. With kind permission from Operational Research Society Ltd. All rights reserved.

in emergency department modelling (Gunal and Pidd, 2010). In this paper we focus on the delivery of acute care outside office hours, when general practitioners' practices are closed. In the Netherlands, when people are confronted with an acute care demand outside normal office hours, they choose whether to go to an Emergency Department (ED) or go to a General Practitioners Post (GPP). In the Netherlands, as well as in several other countries, the organization of primary care delivery is shifting towards an increasing integration of triage and advice by phone, as well as larger care provider cooperatives (Grol *et al,* 2006). The self-referring patients—or 'walk-ins'—who arrive at the ED could often have been seen and treated by a general practitioner (GP), at a GPP, with significant cost savings (Hoot and Aronsky, 2008).

A new concept in the delivering of acute care after office hours is the integrated emergency post (IEP). Similarly to the earlier creation of GPPs, an IEP organizes the provision of after-hours care in a larger cooperative grouping of health-care providers by integrating an ED with a GPP. The main intention of the IEP is to alleviate ED overcrowding by shifting primary care demands from the secondary care provider to the primary care provider, while providing the necessary treatment for patients with an acute care need. The IEP thus offers a sustainable solution to ED overcrowding, resulting in cost savings for the hospital, increased quality of care for the patients and better usage of resources. In this way, the IEP not only has economic benefits, but also on a social and environmental level, and as such contributes to all three pillars of the Triple Bottom Line (TBL). The integration into the IEP gives patients more clarity on where to go with their acute care need, avoids travel times between the GPP and the ED, and provides more opportunities to treat high urgency patients at the ED. In addition, as the IEP is a new concept for health-care providers, this integration of acute health care offers new opportunities for efficiency and efficacy gains.

The objective of this study is to prospectively evaluate organizational interventions and design improvements for an IEP. To this end, we utilize insights from both the screening and the optimization literature as well as domain knowledge of the emergency care environment. A case study is carried out at the IEP in Almelo, the Netherlands, where the ED of the hospital ZiekenhuisGroep Twente merged with the GPP Centrale Huisartsen Post Almelo.

Sustainable development can be defined as the 'development that meets the needs of the present without compromising the ability of future generations to meet their own needs' (WCED, 1987). Similarly, sustainable health care ensures that in the future, resources are available

to address possible future needs. This study contributes to this in two different ways. First, we study an IEP, a new concept that in itself is sustainable, since it uses fewer resources to provide better care to urgent patients. Second, we describe a method, consisting of simulation modelling and a systematic approach, that even leads to better usage of the resources within the IEP. In this way, we contribute to all three pillars of the TBL: people, planet and profit. In this study we explicitly consider the effects of interventions on both people (patients as well as health-care professionals) and profit (cost savings for both the ED and GPP). Efficiently organizing the provision of acute care within an IEP allows for an equitable distribution of care, where patients receive care appropriate to their care needs. Within this study, patient preferences are used to define appropriate performance indicators, accounting for the people aspect in sustainability. Besides effectiveness, the costs of interventions are evaluated such that (in total) a cost effective organization of the IEP is designed (profit). This allows for an increased delivery in care while maintaining current costs. As such, the future ability to offer care to potential future needs is ensured. Implicitly, a better use of resources, for example, through a reduction of diagnostic equipment (previously placed at both GPP and ED), as well as less travel time between organizations for both patients and health-care providers may have a positive environmental impact.

The theoretical contribution of our work is twofold. First, we put forward a method to systematically evaluate a large number of interventions simultaneously, using discrete event simulation, to improve the processes in an acute care chain. Second, we validate and apply this approach to the IEP in Almelo, incorporating patient preferences, and provide insights into the efficient organization of an IEP in general, based on this case study. The practical contribution of our work is the delivery of a sustainable design of the new IEP, taking into account patient preferences and hospital costs, as well as the implementation of a flexible simulation model at the hospital. This implementation includes training of IEP staff in using the model, enabling them to adapt the model when necessary and to evaluate new interventions in the future.

The remainder of this paper is organized as follows. In Section 2, we review the screening and the optimization literature and applications of the reported findings in health care. In addition, we position the IEP concept in literature. Next, we describe the problem and the simulation model (Section 3). In Section 4, we present the approach used to define and evaluate interventions. Following this, we give the results from this

approach as well as describe the implementation in practice based on the results (Section 5). We end this paper with the conclusions drawn from this study (Section 6).

8.2 Literature

An often noted problem facing EDs is overcrowding through self-referring patients that could have been treated at a GP post. The gate-keeper function of a GP post ideally ensures that patients receive the appropriate care for their demand (Dale *et al*, 1995; Kulu-Glasgow *et al*, 1998). However, the cooperation between acute care providers is limited (Thijssen *et al*, 2012), and the effects of creating an IEP, and its potential efficiency gains are uncertain (Kool *et al*, 2008). Within the IEP evaluated in this paper, the collaboration between a GPP and an ED is investigated with regards to the effect of sharing resources, expertise, and organizational strengths, and the efficiency gains these may contain.

Computer simulation is an often used tool in health-care studies, with several comprehensive literature reviews detailing its uses (Jun *et al*, 1999; Fone *et al*, 2003; Brailsford *et al*, 2009; Gunal and Pidd, 2010; Mielczarek and Uziałko-Mydlikowska, 2012). Within the health-care domain, much attention has been given to the modelling of EDs, which are compared with other departments (Gunal and Pidd, 2010). Much of the simulation of EDs is focused on developing and creating valid models. For example, Sinreich and Marmor (2005) detail a basic approach to creating and using a simulation model, focusing on creating a flexible and easy-to-use model, as well as stakeholder involvement. Similarly, Jurishica (2005) discusses proven practices used in developing ED simulation models.

Following model building and validation, usually several interventions are evaluated. For example, Duguay and Chetouane (2007) present a detailed simulation model to evaluate different interventions focusing on staff and room availability with the aim to reduce ED waiting times, and Komashie and Mousavi (2005) evaluate the change in bed availability in an ED. Paul *et al* (2010) conducted a literature review on the use of simulation to investigate overcrowding of EDs and detailed the different evaluated changes. The interventions evaluated were divided over resource-related, process-related, and environmental-related scenarios. Most of the reported articles focus on the reduction of waiting times for patients (Paul *et al*, 2010).

Our goal is to prospectively evaluate many different interventions, using a simulation model, and to find a close to optimal process design

of the IEP. Related methodologies to support this are factor screening and simulation optimization. Factor screening is used to screen for influential factors, and aims to reduce model complexity and computation time, while still achieving good outcomes (Kelton, 2000; Kleijnen, 2008). Simulation optimization attempts to find the combination of controllable factor settings that lead to the best outcome (April *et al*, 2003). However, it must also deal with noise, as simulation outcomes are approximations of true performance indicators.

A straightforward screening method is a full factorial or 2^k design, such that both main effects and interactions may be evaluated (Law, 2007). A downside of this method is that, as the number of factors increase, many runs are required. To overcome this, fractional factorial designs may be used, which require fewer runs, with a loss of some interaction effects (Law, 2007). However, care should be taken not to combine potentially important effects (Montgomery, 2008). Other screening approaches for a large number of factors are two-stage group screening (Mauro, 1984; Ivanova *et al*, 1999; Trocine and Malone, 2001), sequential bifurcation (Bettonvil and Kleijnen, 1997; Cheng, 1997; Yaesoubi *et al*, 2010), iterated fractional factorial designs and supersaturated designs. We see that among the different screening techniques, many are specific in regard to the assumptions made, the maximum number of factors, applicability regarding qualitative and (continuous) quantitative factors, and the use of information already known before screening, such as the signs of interaction.

Different approaches to simulation optimization are ranking & selection (Boesel *et al*, 2003; Fu *et al*, 2005), response surface methodology (Fu *et al*, 2005), gradient-based procedures (Fu *et al*, 2005), random search (Andradóttir, 2006), and sample path optimization and metaheuristics (April *et al*, 2003). An example of simulation optimization applied to an ED is given by Ahmed and Alkhamis (2009), who determine the number of staff required to optimize patient throughput and waiting times. Similar to the screening methods, the application of optimization techniques relies on characteristics of the underlying simulation model. Depending on input variables and constraints, certain optimization methods are more applicable than others. For more details we refer to Andradóttir (1998) and Barton and Meckesheimer (2006). An additional challenge with optimization is that the methods may propose a solution which requires many organizational changes within the ED, which might prevent a successful implementation. Much of the literature on simulation in health care focuses on model construction and validation, with less emphasis on defining and evaluating possible interventions and their interaction effects. This has motivated us (i) to propose

a way to evaluate and improve over a large number of interventions using a discrete event simulation model and (ii) to consider the optimal organization of a new innovative concept in health care, that is, the IEP.

Since it is unlikely that the management of the IEP will accept drastic organizational changes in which many interventions are combined, we put forward a structured approach in which we identify only the most effective interventions for an area of change (eg, rostering alternatives). This results in a smaller set of interventions that increase the effectiveness of the IEP without making drastic organizational changes.

8.3 Problem description

In this section, we give a general description of an IEP and describe the implemented IEP in Almelo. Following this, we briefly address the simulation model that is used to evaluate the IEP and the way in which we incorporate patient and community preferences in formulating performance indicators. For a complete description of this simulation model, we refer to Mes and Bruens (2012).

8.3.1 Integrated emergency post design

In an IEP, the GPP and ED work together to provide acute care to patients outside of office hours. The largest change with the introduction of the IEP is that self-referring patients, or walk-ins, are now first seen at the GPP, instead of possibly going to the ED. There are several ways in which patients may enter the IEP: by calling the IEP, going to the IEP as a self-referral, and by being referred to the IEP by an external care provider. When a patient calls the IEP, a telephonic triage takes place, and depending on the urgency of care demand, the patient either gets a consultation at the IEP, a doctor visits the patient, the patient is referred straight to the ED (by ambulance), or the patient receives medical advice by phone. Self-referring patients first undergo physical triage by a GP assistant, after which they are sent home with medical advice, scheduled for a GP consultation or referred to the ED. Finally, external referrals are sent directly to the ED.

Patients that receive an appointment via telephone or self-refer and enter the IEP are first seen at the GPP. In most cases, this treatment is sufficient after which a patient goes home. Possibly, the patient may require an X-ray, after which, depending on the results, a patient goes home, or in case of a fracture, is referred to the ED. Similarly, other patients that cannot be treated at the GPP, or require additional treatment, are further referred to the ED.

Patients that enter the ED are first triaged again, as the triage system differs from the one used at the GPP, and then the patient history is

registered by an ED nurse or physician assistant. Afterwards, a patient might undergo multiple diagnostic tests and receives treatment. After all treatment is finalized, the patient leaves the system and continues his care path outside the ED or goes home. Additionally, some patients require multiple diagnostic tests with treatment after each test. Figure 8.1 shows a conceptual view of the care pathways defined of the IEP in Almelo.

As the IEP is a new concept, many organizational changes may improve the IEP's efficiency, for example, by pooling resources between the GPP and the ED, which can be prospectively evaluated using simulation. The IEP in Almelo has been modelled and validated in a detailed simulation model.

8.3.2 Simulation model of the integrated emergency post

The simulation model is characterized by patients, resources, and processes. Patients move through the IEP and require resources in the form of health-care professionals, diagnostic test equipment, and beds/rooms. Patients and resources are characterized by various attributes, such as urgency, diagnostics needed, availability, and capacity. The processes define the care path—the sequence of steps—that a patient must undergo.

Within the simulation model, a task list contains all the treatment steps that are to be taken next for all patients in the system. When a patient is created, a module is triggered that, depending on patient attributes and care pathway, adds the first task (treatment step) for that patient to the task list, and checks and prioritizes all open tasks. If a task is started, patient and resource handling is started, which involves updating resource availability, performance registration and animation. When a task ends, patients and resources are handled (ie, released), the next task in the care pathway is added to the task list, and the module again checks all open tasks. Besides the arrival of patients and the ending of a task, the passing of time triggers the module. For example, when a shift ends, there is a change in staff availability, requiring resource handling.

Three different types of processes (tasks) are modelled, being regular tasks, parallel tasks, and delay tasks. Regular tasks are the treatment steps required by patients. These are mostly processes where a patient and staff member come together in a room, and potentially require additional resources such as diagnostic equipment (eg, ultrasound). Parallel tasks are those where a patient need not be present, such as the reviewing of diagnostic tests. As such, this task may also be carried out while a patient undergoes another regular task. The delay tasks are processes that must be carried out before a regular task may start, causing waiting time. For example, when a patient requires treatment from

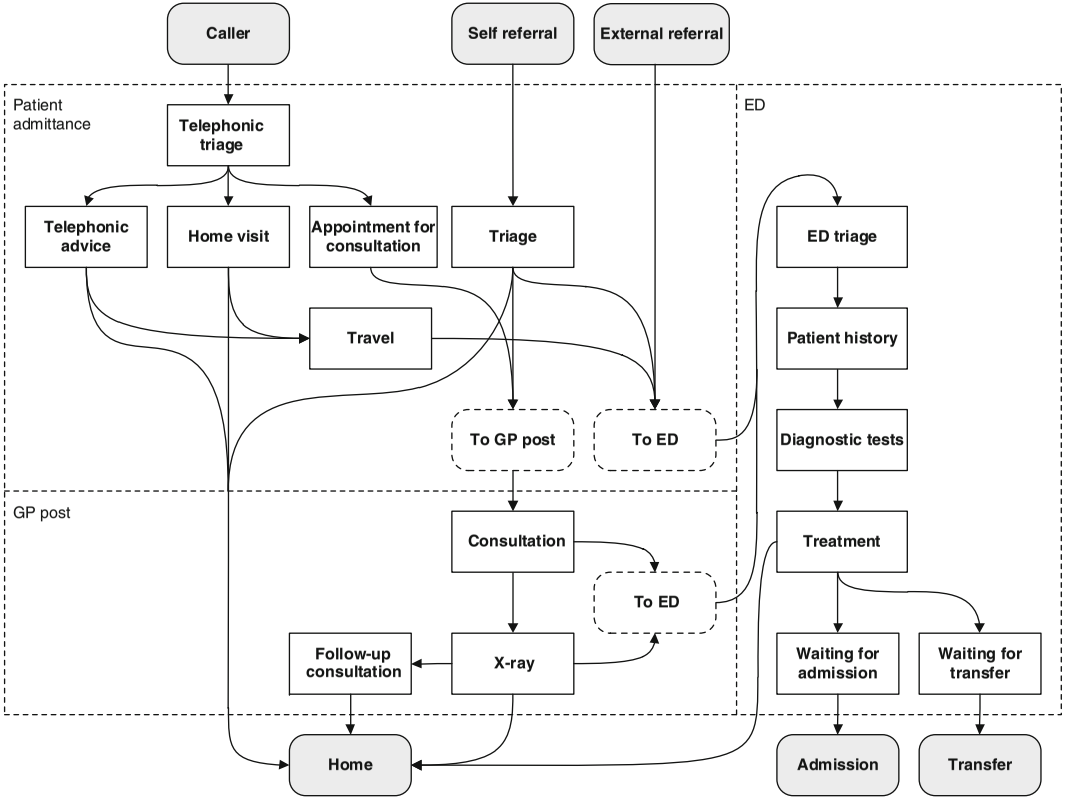


Figure 8.1 Patient flow and processes at the IEP Almelo (Mes and Bruens, 2012)

a specialist, this staff type is called in from elsewhere in the hospital. This waiting time for the specialist to finish his current job and travel to the ED is modelled as a delay task. Within the simulation model, the arrival of patients, as well as the occurrence of treatments and diagnostic tests is modelled using probability distributions based on four years of historical data, as well as on expert opinion.

Arrivals of patients entering the IEP are described using a non-stationary Poisson process, depending on the time of day, day of the week, and week of the year. As input we have 24 arrival rates (one per hour) for weekdays (Monday till Friday) and 24 arrival rates for days in the weekend. In addition, we have 7 day factors and 52 week factors that we multiply with the arrival rate to determine the arrival rate within a given hour at a specific day. These factor distributions are determined using multiple years of historic data and represent the daily and weekly fluctuations in patient arrivals. For the non-stationary Poisson process, we generate arrivals based on the highest arrival rate per hour in combination with the thinning procedure (Law, 2007).

The resources in the model are distinguished as staff, equipment and rooms. The following staff is included: GP, GP assistant, triage assistant, ED nurse, nurse practitioner, physician assistant, medical resident and surgical resident. In addition, medical specialists and diagnostic nurses are included as external staff, who may be called to the ED when needed for treatment and diagnostic tests respectively. Since these staff types are not usually present at the ED, but requested frequently, we take into account the waiting time for their availability, and travel time to the ED. Regarding equipment and rooms, we include one triage room, six GP rooms, eight ED rooms, a CT room, two plaster rooms, and two X-ray rooms. In addition there is portable ECG and ultrasound equipment which are used in combination with an ED room. For more details on the components and the construction of the simulation model, we refer to Mes and Bruens (2012).

8.3.3 Modelling for sustainability

Organizations place an increasing focus not only on cost-related performance indicators, but also on social and environmental factors. These factors, and especially the social factors, are crucially important to health-care organizations. This paper is part of a research project studying the optimal process design of an IEP, such that patients are seen by the right care provider, without unnecessary delays, while accounting for patient preferences (Doggen *et al*, 2010). This research scope accounts not only for production measures, but for social and environmental measures as well. In addition to a simulation study and

advice aimed to improve the current situation, the simulation model is designed flexibly, to enable health-care professionals (policy advisors) to adapt the model to the changing environment and to evaluate further improvements in the future.

To account for patient preferences, and to construct social and equitable performance indicators for this study, Fransman (2011) conducted a patient and community preference analysis using best-worst scaling. These outcomes showed that patients primarily value (lower) waiting times, followed by direct access to care provider (self-referral over appointment based), and type of care provider (physician over nurse practitioner). These indicators are incorporated in the simulation model, which is now used by the IEP. To keep our presentation concise, we focus on the key performance indicator (KPI), that is, the patient length of stay (LOS).

To improve the IEP, we aim to evaluate the effects of all potential interventions, and also account for the interaction that can take place between interventions. However, as the number of potential interventions increases, an evaluation of all combinations becomes intractable. Therefore we want to improve the IEP without evaluating every intervention alternative. The next section formulates the approach to first identify potentially effective interventions and then further evaluate these interventions.

8.4 Approach

In this section, we formulate the approach used to identify and evaluate potential organizational interventions. As many interventions are identified, we first formulate and categorize the interventions into several categories and eliminate ineffective interventions. Following this, we evaluate the remaining interventions and combine these into several combinations.

Given the simulation run time, a full factorial design involving many interventions quickly becomes intractable. In addition, some interventions can have many variations, such as changing staff numbers, creating many potential staffing schedules. We, therefore, first assess for which of the interventions further evaluation seems promising. We do this by evaluating interventions in groups, which reduces the number of possible alternatives. Furthermore, it is unlikely that the management of the IEP will accept drastic organizational changes in which many interventions are combined. We use a structured approach in which up to three influential interventions per group are included. This

increases the effectiveness of the IEP without drastically changing its organization. Therefore, we use the approach given below.

1. List interventions
2. Categorize interventions based on type and specificity
   a. Evaluate (interaction) effects per group
   b. Select the most effective interventions per group
3. Formulate intervention sets
4. Compare intervention sets
   a. Evaluate absolute intervention outcomes
   b. Scenario analysis

Paul *et al* (2010) give an overview of interventions evaluated in healthcare simulation, and make the distinction between process, resource, and environmental (eg, increasing patient arrivals) changes. Following this approach, we group interventions into process and resource changes. In addition, we further divide the resource interventions into subgroups consisting of interventions of similar nature, namely staff, diagnostics, allocation, and a roster options group.

Each 'roster options' intervention is a combination of several smaller interventions. For example, an earlier or later start of the main shifts gives insights into the fit between patient arrivals and number of staff, which is a combination of varying several staffing levels and starting times.

The environmental changes are similar to simulations carried out in a scenario analysis, evaluating the effect future demographic changes may have.

Based on the patient and community preferences we compare interventions using the patient LOS as KPI, which incorporates both treatment and waiting times (cf use in Sinreich *et al* (2012) and Ashour and Kremer (2013)). We use total LOS as a KPI, because potential interventions may affect waiting times as well as treatment times (eg, due to the intervention to use a single triage system). Some interventions target the ED or GPP specifically, so we split the LOS into GPP LOS and ED LOS. However, some interventions may target specific patient groups (eg, high urgency patients). Therefore, in the second part of our analysis, we also study the impact of the interventions for urgent and non-urgent patient groups. Also, we make a distinction between the impact on week and weekend days.

New interventions are formulated from the groups, combining the effective interventions. This is done not only based on simulation outcomes,

but also based on expert opinion about the feasibility of the interventions (eg, staff availability and costs). Ideally, an equitable distribution of health care places more emphasis of those patients that need more care (ie, high urgency patients); therefore, the intervention set alternatives are compared by evaluating LOS for both high and low urgency patients individually. The categorization of high and low urgency patients is based on the triage systems used at the GPP and ED. The GPP uses the Dutch GP society (NHG) (Giesen *et al*, 2006), and the ED uses the Manchester Triage System (Mackway-Jones *et al*, 2006). At the GPP, a patient's urgency ranks from U4 (not urgent), to U1 (very urgent), and at the ED, a patient's urgency is ranked blue (not urgent), green, yellow, orange, or red (very urgent). We categorized U1 and U2 patients, as well as red and orange patients as high urgency patients, and all remaining patients as low urgency patients.

8.5 Results

In this section, we conduct the approach described before and evaluate the (interaction) effects within each of the intervention groups. Following this, we select the interventions that show significant improvements, and use these to formulate intervention sets, which we then compare using the simulation model. Finally, a detailed evaluation and scenario analysis is carried out on the most promising intervention sets.

8.5.1 List interventions

To come up with interventions to evaluate, we use both simulation literature and interviews with stakeholder from the GPP and ED to define as many interventions as seem feasible to simulate and implement. In total 24 interventions are defined. For example, a stakeholder at the intervention hospital noted the possibility of utilizing a single triage system, such that patients that are referred from the GPP to the ED no longer have to be triaged a second time. Another example is the allocation of staff to patients. Currently, when a patient arrives, first ED specialists or residents are assigned to that patient, and if none of them is available, a physician assistant is assigned. As an intervention this prioritization is reversed, so that lower urgency patients are treated first by a physician assistant, keeping ED specialists and residents available for higher urgency patients that may arrive. A third example is the staffing of medical specialists at the ED. Currently, specialists are called when needed at the ED; by having specialists placed at the ED, the waiting time once a specialist is called is removed. An overview of all defined

interventions is given in Table 8.1, the numbers in parenthesis indicate the intervention number.

8.5.2 Categorize interventions

Following the listing of interventions, a categorization is made based on the type of intervention as well as on the resulting number of interventions per category. Table 8.1 lists the division of interventions over the possible categories. The interventions in the category 'process' and in the group 'allocation' are all binary (on/off), while the diagnostics and staff changes may have several alternatives created by adding or subtracting more than one resource. In addition, staffing changes can also be temporal in nature by changing the starting times. To evaluate

Table 8.1 Identified and categorized interventions

Category	Group	Intervention
Process		Use a single triage system (1)
		Change triage protocol to let ED nurse order diagnostics (2)
		Give priority only to high urgency patients (3)
		Assign physician assistants to patients before physicians (4)
		Initiate request of hospital admission earlier (5)
Resource	Staff	Vary the number of ED nurses (6)
		Vary the number of surgical residents (7)
		Vary the number of internal medicine residents (8)
		Vary the number of physician assistants (9)
		Vary the number of general practitioners (10)
		Vary the number of ED specialists (11)
Resource	Diagnostics	Vary the number of X-ray rooms (12)
		Vary the number of CT scan rooms (13)
		Vary the number of ECG equipment (14)
		Vary the number of Ultrasound equipment (15)
Resource	Allocation	Treat (low urgency) ED patients in GPP rooms (16)
		Let physician assistants treat both ED and GPP patients (17)
		Let ED nurse treat GPP patients (18)
		Let ED specialists/residents treat GPP patients (19)
Resource	Roster options	Use medical specialists at IEP at all times (20)
		Use future hospital roster (21)
		Replace internal medicine resident with ED specialist (22)
		Earlier main shift (GP/GP assistant/ED nurse) (23)
		Later main shift (GP/GP assistant/ED nurse) (24)

the potential of staff interventions, we first only add an additional staff member during the busiest hours of the IEP. We use the outcomes and insights of the staff interventions in the next step, where the intervention sets are created, by further specifying feasible staffing alternatives (ie, staff allocations). Similarly for diagnostics, an additional diagnostics machine is added in the simulation model. The roster options group consists of larger organizational interventions, which are a combination of smaller interventions.

For the first four groups, the insights into the main effects as well as interaction effects between interventions are evaluated per group using full factorial designs, and used to define the (up to) three most promising interventions per group. As the interventions in the roster options group are mutually exclusive staffing alternatives, these are evaluated individually.

During the categorical evaluation, we set the run length such that the specified precision of the most variable intervention has at most a relative error of 5% with a confidence level of 95% (Law, 2007). Using this method, the simulation run length is 32 weeks per experiment for the process, staff, and diagnostics designs; 48 weeks for the pooling and allocation design; and 31 weeks for the roster design. Table 8.2 lists the number of experiments per factorial design that are evaluated initially, as well as the number of simulation runs carried out per experiment.

In total, 24 different interventions are evaluated, of which the first 19 (Table 8.1) are evaluated using four full factorial designs. The last five interventions are evaluated individually. This results in 133 different intervention combinations that are evaluated, with a total number of 4507 runs. Based on the outcomes of these experimental designs and individual comparisons, we select the interventions that both have a significant (positive) effect on the IEP, and are deemed feasible based on expert opinion.

Table 8.2 Experiments per intervention category

Experimental design	**#interventions**	**#experiments**	**#runs (#exp *x* #repl)**
Process interventions	5	32 ($=2^5$)	1024 (32×32)
Staff interventions	6	64 ($=2^6$)	2048 (64×32)
Diagnostics	4	16 ($=2^4$)	512 (16×32)
Pooling and allocation	4	16 ($=2^4$)	768 (16×48)
roster interventions	5	5 ($=5^1$)	155 (5×31)
Total	24	133	4507

8.5.3 Category outcomes

Of the 19 interventions evaluated with full factorial designs, 12 are found to have a significant main effect within their respective groups (ie, full factorial design), with an additional five significant interaction effects. Figure 8.2 lists all the main and two-way interactions that are found to be significant ($\alpha = 0.05$).

Several process changes show an improvement at the ED. Both a single triage system (1), as well as a direct ordering of more diagnostic tests (2) reduce the ED LOS by approximately 300 seconds (5%). Furthermore, the direct admission requests (5) of specific patient groups reduce ED LOS by approximately 120 s (2%). Reprioritizing patients (3) reduces average GPP LOS by 20 s (1.3%); however, from results not shown here, low urgency (U4) LOS decreases by 10%, while LOS increases by 4 and 18% for U2 and U3 patients, respectively, redistributing the time spent at the GPP over the different patient types.

From the interaction effect in Figure 8.2, it is apparent that the biggest influence on reducing the GPP LOS is achieved by adding a physician assistant (9), followed by adding a GP (10). In addition, we see that the two-way interaction effect between these staff types is positive; interpreting this, the combination of these staff types has a diminished effect on the reduction of the LOS. For the ED, we see that the biggest influence is achieved by adding an ED specialist (11), followed by adding a surgical resident (7) and adding a physician assistant (9). By adding these staff types, the average LOS is reduced by 115 (11), 60 (7) and 36 (9) seconds, respectively. Looking at the interaction effects on the GPP and ED LOS, we see that the biggest reduction on GPP LOS is achieved by allowing ED staff to treat GPP patients (18 and 19).

Additionally, there is a substitution effect where the combined interventions have a dampening effect on each other. By letting ED staff treat GPP patients, the bottleneck of staff being unavailable decreases, and shifts to another resource type. In this situation, while staff may be available, treatment room shortages may occur. The roster alternatives show a large reduction in ED LOS when medical specialists instead of residents are scheduled at the IEP. By staffing the ED with medical specialists, they no longer need to be called from within the hospital, removing the waiting- and travel time when requested for consultation. This staffing of medical specialists results in a LOS reduction of over 10 min (10%). However, this would also be a costly intervention, as specialist salaries are considerably higher than those of the residents'.

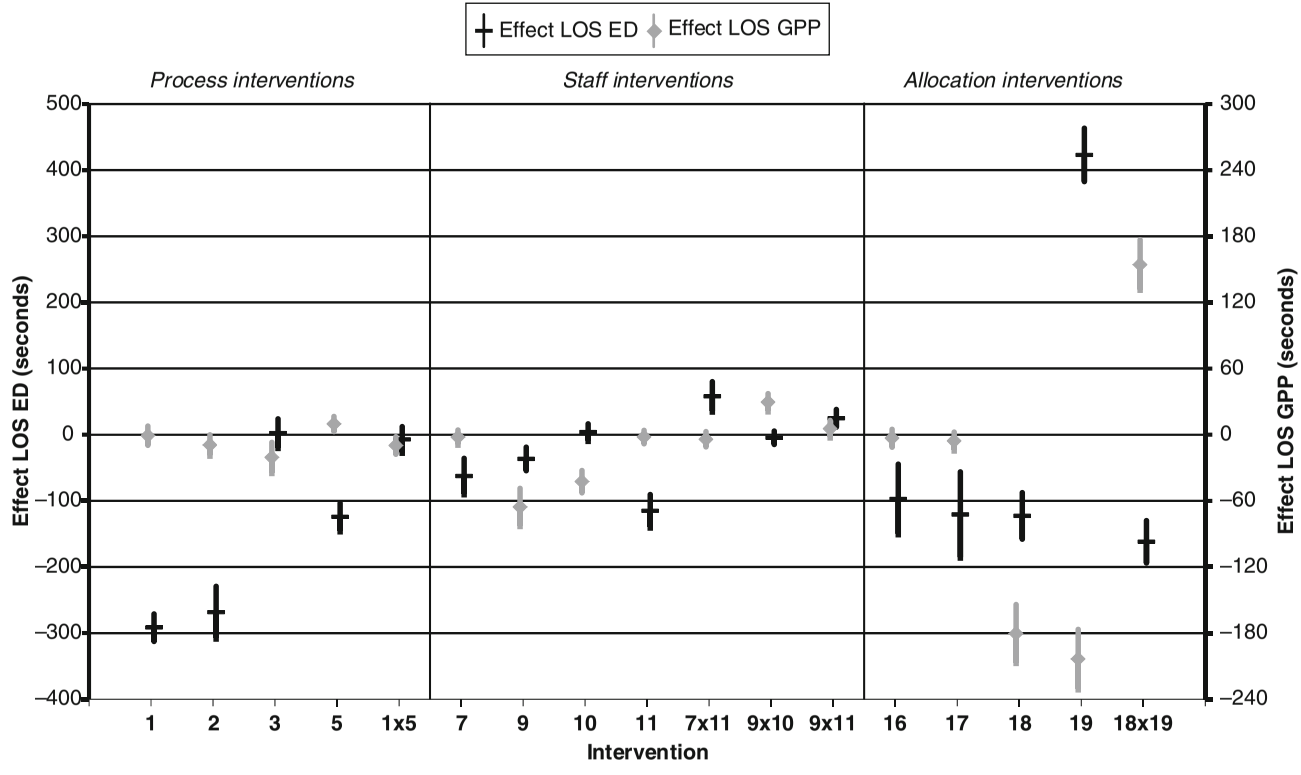


Figure 8.2 Significant ED and GPP main and two-way interaction effect confidence intervals ($\alpha = 0.05$) (A × B denotes the interaction effect between interventions A and B)

Based on these outcomes, combined with insights on associated costs and feasibility, we identify the following effective interventions.

1. Use a single triage system (1)
2. Change triage protocol to let ED nurse order diagnostics (2)
3. Initiate request of hospital admission earlier (5)
4. Vary the number of surgical residents (7)
5. Vary the number of physician assistants (9)
6. Vary the number of ED specialists (11)
7. Treat (low urgency) ED patients in GPP rooms (16)
8. Let physician assistants treat both ED and GPP patients (17)

8.5.4 Formulate intervention sets

As the variations of staff allow for many different schedules, we further specify interventions which are not of a binary nature, and formulate several cost equivalent alternatives. These alternatives combined with the binary process and pooling interventions result in the interventions shown below. In total, we define 10 interventions, which we combine into five sets (combinations of interventions).

1. Use a single triage system
2. Change triage protocol to let ED nurse order diagnostics
3. Initiate request of hospital admission earlier
4. Treat (low urgency) ED patients in GPP rooms
5. Let physician assistants treat both ED and GPP patients
6. Roster alternatives
   a. Replace surgical resident with ED specialist and add a physician assistant during the weekends busy hours (similar to intervention 21, future hospital staffing schedule)
   b. Schedule two physician assistants during the Saturday and Sunday busy hours instead of a GP
   c. Schedule a surgical and internal medicine resident instead of the ED specialist during the Saturday and Sunday busy hours
   d. Schedule a physician assistant instead of an ED nurse during the Saturday and Sunday busy hours
   e. Schedule a physician assistant instead of an ED nurse during the first opening hours of the IEP (5 pm–8 pm)

The different process and pooling interventions can be combined with each other to form different intervention sets. However, none of the five process and pooling interventions showed interaction within their

respective groups, and it seems likely that there will be little interaction between these interventions. Therefore, we expect that a combination of these interventions, together with a roster alternative, will show the greatest LOS reduction for the IEP. To evaluate the effect of the roster alternatives, we combine each of them with the selected process and pooling interventions as shown in Table 8.3. Specifically, from the listed interventions we create five intervention sets. In each set interventions 1–5 are evaluated. In addition, each of the five sets also incorporates one of the mutually exclusive roster alternatives (6a–6e). Note that intervention 6e only affects the weekday evening schedule, instead of the weekend days.

8.5.5 Compare intervention sets

To evaluate the effect of the intervention sets, we make a distinction between LOS at the ED and GPP, as well as in LOS during weekend and weekdays. Additionally, we look at the LOS for high (U1 and U2 at the GPP, and red and orange at the ED) and low (U3 and U4 at the GPP, and yellow-blue at the ED) urgency patients. Figure 8.3 shows the effect of the various intervention sets with 95% confidence intervals. We set the run length of the experiments such that the specified precision of the most variable intervention and LOS measurement has at most a relative error of 5% with a confidence level of 95% (Law, 2007). This results in a run length of 222 weeks per experiment.

All intervention sets improve the ED LOS compared with the original situation, as not only the procedural interventions are implemented, there is also an additional staff member working during the busy hours. The results between intervention sets in Figure 8.3 are significantly different ($\alpha = 0.05$) for all comparisons except the weekday LOS for High urgency GPP patients, and set 2 and 3 regarding weekend high urgency GPP LOS. The highest decrease is shown when two residents are scheduled instead of an ED specialist (set 3). However, for the cases where the

Table 8.3 Intervention sets

Set\Intervention	**1**	**2**	**3**	**4**	**5**	**6a**	**6b**	**6c**	**6d**	**6e**
Set 1	x	x	x	x	x	x				
Set 2	x	x	x	x	x		x			
Set 3	x	x	x	x	x			x		
Set 4	x	x	x	x	x				x	
Set 5	x	x	x	x	x					x

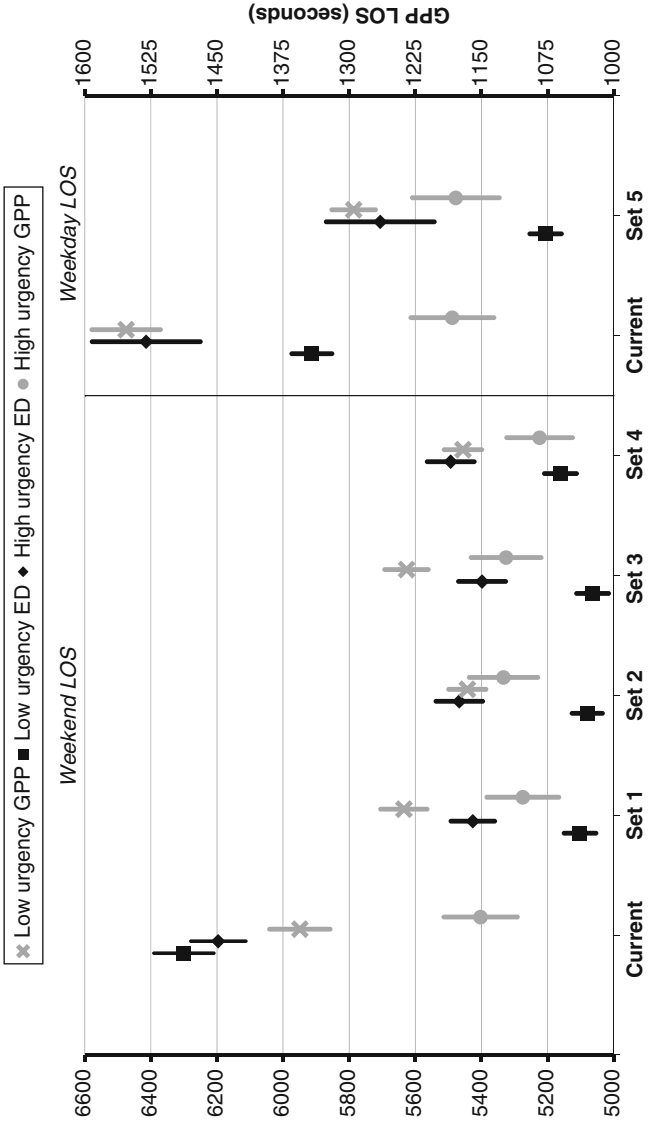


Figure 8.3 Average LOS outcome confidence intervals (95%) per intervention set

differences are not significant, the differences are small compared with the total LOS times.

Similarly, we see a decrease in GPP LOS. However, the effect on high urgency patients is remarkably lower, with the confidence intervals overlapping with those of the current situation. Replacing a GP with two physician assistants (set 2) has the greatest effect on lowering the low urgency patient LOS. However, the differences between combinations are all minor. The differences in ED LOS between the best and worst intervention set differ approximately 100 s, and at the GPP the differences are around 60 s.

As no alternative intervention set decreases LOS significantly more than any other intervention set, the most promising weekend alternative seems to be 6a (set 1), given that this set contains a roster alternative preferred by the IEP stakeholders. In addition, the weekday replacement of an ED nurse with a physician assistant (6e) that treats both GPP and ED patients shows a decrease for all low urgency patients, making this a promising alternative as well.

We simulate the interventions individually to assess the absolute effects of each individual intervention on the IEP, and compare them with each other. Table 8.4 shows the LOS for both the GPP and the ED over both type of day (weekend or weekday) as well as high or low urgency patients. The bolded outcomes are found to be significant ($\alpha = 0.05$).

Overall, all selected interventions show a significant improvement over the current situation. Of these, the roster alternatives show the greatest effect on GPP LOS, and process interventions on ED LOS. In addition, the pooling of resources, such as staff, shows that both the GPP and ED can benefit, while the overall staffing costs remain virtually unchanged. In addition, the interventions have no significant negative effects on any subset of patients. This is especially important for the interventions that target the low urgency patient groups, as an (overall) decrease in LOS may still allow for an unequal distribution of care over patient groups.

With a sensitivity analysis we evaluate the effects of potential environmental changes (cf Paul *et al*, 2010). By varying the number of patient arrivals, as well as the urgency of patients, the LOS is measured for both the current organization of the IEP, and all interventions from Table 8.4 combined. The sensitivity analysis results show that with the selected interventions, both ED and GPP are able to treat both more patient arrivals and more acute patients. For example, when patient arrivals increase, the difference in average GPP LOS between selected interventions and the current organization of the IEP increases. At a

Table 8.4 Absolute outcomes (seconds) per intervention with significant outcomes ($\alpha = 0.05$) made bold

	Weekday high urgency	**Weekday low urgency**	**Weekend high urgency**	**Weekend low urgency**
GPP performance				
Current situation	1186	1557	1151	1356
1	+4 (+0.4%)	+9 (+0.6%)	−28 (−2.5%)	−16 (−1.2%)
2	+11 (+0.9%)	+14 (+0.9%)	−32 (−2.8%)	−5 (−0.4%)
3	+33 (+2.8%)	+19 (+1.2%)	−28 (−2.5%)	−14 (−1%)
4	+29 (+2.4%)	+3 (+0.2%)	−8 (−0.7%)	+4 (+0.3%)
5	+8 (+0.7%)	−38 (−2.4%)	−32 (−2.8%)	+ 12 (+0.9%)
6a	+19 (+1.6%)	−34 (−2.2%)	−21 (−1.8%)	***−122*** **(*−9%*)**
6e	−8 (−0.7%)	**−274 (−17.6%)**	−7 (−0.6%)	−7 (−0.5%)
ED performance				
Current situation	6421	5901	6201	6299
1	−193 (−3%)	**−287 (−4.9%)**	**−183 (−2.9%)**	**−349 (−5.5%)**
2	**−366 (−5.7%)**	**−203 (−3.4%)**	**−381 (−6.1%)**	**−290 (−4.6%)**
3	−215 (−3.3%)	**−126 (−2.1%)**	**−141 (−2.3%)**	−109 (−1.7%)
4	+36 (+0.6%)	**−119 (−2%)**	−18 (−0.3%)	−108 (−1.7%)
5	+5 (+0.1%)	−15 (−0.3%)	−71 (−1.2%)	**−236 (−3.7%)**
6a	+17 (+0.3%)	−46 (−0.8%)	**−238 (−3.8%)**	**−476 (−7.5%)**
6e	+26 (+0.4%)	−11 (−0.2%)	−28 (0.4%)	−7 (−0.1%)

20% patient increase, the average GPP LOS increases by approximately 65% in the current situation, and 40% with the selected interventions. At a 50% patient increase, the average GPP LOS with the selected interventions increases with 80%, while the GPP LOS in the current situation increases with 250%, showing that an optimized IEP is better equipped to treat an increasing number of patients.

8.5.6 Implementation

Based on the simulation study outcomes, a physician assistant that treats both GPP and ED patients during the starting hours of the IEP has been trialled in a 3-week pilot study. The aim of this study was to investigate the effects of adding a physician assistant in the actual situation, treating both ED and GPP patients. During a 3-week measurement period (January 2013), every week day from 5 pm until 8 pm, a physician assistant consulted and treated both low urgency GPP patients, as well as low urgency ED patients. During this pilot study, in addition to

other performance indicators, GPP waiting times (n = 273) were measured and compared with measurements carried out before the pilot (n = 237). This pilot study showed a positive effect for GPP patients, reducing the average waiting time with 124 s while having no effect on ED patient LOS. These outcomes are similar to those of the simulation model, where no significant effect on the ED LOS is visible and the GPP LOS is reduced by the same order of magnitude. This similarity has been recognized by the IEP staff. From this we conclude that the simulation model is a valid representation of the actual situation.

8.5.7 Sustainability

By using community and patient preferences in the KPIs, we evaluated both economic and social effects of the interventions. From our results, we see that all patients benefit from using the IEP, and those patients that need the most care, receive the most benefit (ie, ED patients). An additional side effect of the integration into an IEP is environmental in nature. With the repositioning of care providers such that they are placed at a single location, both travel time and costs that patients would have incurred when travelling between the GPP and the ED are removed. Furthermore, resources in the IEP are better used thanks to the integration, also leading a more sustainable solution.

The end product of this research is twofold: first we give advice to the health-care providers on an improved process design for the IEP in the current situation, and second we developed a reusable simulation tool and systematic approach the IEP can use to evaluate potential interventions that arise in future situations. To this end, simulation tutorials have been given to the IEP stakeholders, including physicians, ED managers, and the GPP director. During these tutorials, simple health-care processes have been modelled to gain understanding and acceptance of simulation modelling, and by using the simulation model of the IEP, the participants could evaluate various interventions by themselves. In addition, the hospital appointed one employee, responsible for process improvements in the hospital, to keep the simulation model up to date, and use it to both help and evaluate the IEP in continuous and sustainable process improvements.

8.6 Conclusion

We used a systematic approach in defining and evaluating many organizational interventions for an IEP. We did this by first grouping interventions and selecting the most effective interventions per group.

Following this, we formulated intervention sets which were compared and further evaluated. Using this approach, we evaluated both the effects of potential interventions, as well as the interaction between these interventions. This enabled us to compare and evaluate many changes while keeping the required simulation time feasible. The use of the simulation model, as well as the structured approach was essential to evaluate the IEP in Almelo, without intervening in the actual processes of the IEP.

We identified various interventions, divided by process interventions, staffing interventions, and resource allocation interventions, showing a reduction in patient LOS at both the GPP and ED. Process and allocation interventions show a decrease of 2–6% in ED LOS, and staffing alternatives reduce GPP and ED LOS with 3–17%, depending on the patient group. The IEP offers a sustainable solution to the problem of ED overcrowding, where all stakeholders benefit: better care is provided to patients by increasing clarity (ie, for patients to decide where to go) and effectiveness, at reduced costs. By accounting for patient and community preferences, various interventions have been identified resulting in positive effects on both social and environmental factors without causing drastic changes to the organization and without introducing additional costs for the ED or the GPP. In addition, a reusable simulation tool is embedded within the hospital with staff trained to use this, allowing for continuous improvements of the IEP.

Finally, based on the results of this study, a pilot project of one of the interventions, letting physician assistants treat patients at both the GPP and ED, has been trialled in practice, showing an improvement that closely matches the prediction resulting from our simulation model. Future work will involve modelling the interaction between the IEP and other departments in the hospital, such that the entire care pathway chain can be optimized. In addition, a follow-up study would be of interest evaluating the implementation and continued use of the simulation model and approach, as well as the hospital's experiences using this model. This would help to further understand the barriers found in practice when using simulation models, and indicate how such problems may be overcome.

References

Ahmed MA and Alkhamis TM (2009). Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research* **198**(3): 936–942.

Andradottir S (1998). A review of simulation optimization techniques. In*: Proceedings of the 30th conference on Winter simulation*, Washington, DC, USA, IEEE Computer Society Press, pp 151–158.

Andradóttir S (2006). Chapter 20 An overview of simulation optimization via random search. In: Henderson SG and Nelson BL (eds). *Handbooks in Operations Research and Management Science*. Vol. 13, Elsevier: Amsterdam, pp 617–631.

April J, Glover F, Kelly JP and Laguna M (2003). Simulation-based optimization: practical introduction to simulation optimization. In: *Proceedings of the 35th conference on Winter simulation: driving innovation*, New Orleans, Louisiana, Winter Simulation Conference, pp 71–78.

Ashour OM and Kremer GEO (2013). A simulation analysis of the impact of FAHP-MAUT triage algorithm on the emergency department performance measures. *Expert Systems with Applications* **40**(1): 177–187.

Barton RR and Meckesheimer M (2006). Chapter 18: Metamodel-based simulation optimization. In: Henderson SG and Nelson BL (eds). *Handbooks in Operations Research and Management Science*. Vol. 13, Elsevier: Amsterdam, pp 535–574.

Bettonvil B and Kleijnen JPC (1997). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research* **96**(1): 180–194.

Boesel J, Nelson BL and Kim S-H (2003). Using ranking and selection to "clean up" after simulation optimization. *Operations Research* **51**(5): 814–825.

Brailsford SC, Harper PR, Patel B and Pitt M (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation* **3**(3): 130–140.

Cheng RCH (1997). Searching for important factors: Sequential bifurcation under uncertainty. In: *Proceedings of the 29th conference on Winter simulation,* Atlanta, Georgia, USA, IEEE Computer Society, pp 275–280.

Dale J, Green J, Reid F and Glucksman E (1995). Primary care in the accident and emergency department: I. Prospective identification of patients. *Bmj* **311**(7002): 423–426.

Doggen CJM, Hans EW, Snel JE, Velde DV and Verheij HJW (2010). *Subsidieaanvraag: "Optimale logistiek en patienten voorkeuren in de acute zorgketen; de huisartsenpost en spoedeisende hulpin een geintegreerde spoedpost."* Research Proposal, University of Twente.

Duguay C and Chetouane F (2007). Modeling and improving emergency department systems using discrete event simulation. *Simulation* **83**(4): 311–320.

Fone D *et al* (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health* **25**(4): 325–335.

Fransman C. (2011). *Patient and community preferences for out-ofhours emergency care using best-worst scaling.* MSc thesis, University of Twente.

Fu MC, Glover FW and April J (2005). Simulation optimization: a review, new developments, and applications. In: *Proceedings of the 37th conference on Winter simulation,* Orlando, Florida, Winter Simulation Conference, pp 83–95.

Giesen P, Franssen E, Mokkink H, Bosch W, van Vugt A and Grol R (2006). Patients either contacting a general practice cooperative or accident and emergency department out of hours: A comparison. *Emergency Medical Journal* **23**(9): 731–734.

Grol R, Giesen P and van Uden C (2006). After-hours care in the United Kingdom, Denmark, and the Netherlands: New models. *Health Affairs* 25(6): 1733–1737.

Gunal MM and Pidd M (2010). Discrete event simulation for performance modelling in health care: A review of the literature. *Journal of Simulation* **4**(1): 42–51.

Hoot NR and Aronsky D (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2): 126–136.

Ivanova T, Malone L and Mollaghasemi M (1999). Comparison of a two-stage group-screening design to a standard 2k-p design for a whole-line semiconductor manufacturing simulation model. In: *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future—Volume 1*, Phoenix, Arizona, USA, ACM, pp 640–646.

Jun JB, Jacobson SH and Swisher JR (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society* **50**(2): 109–123.

Jurishica CJ (2005). Emergency department simulations: medicine for building effective models. In: *Proceedings of the 37th conference on Winter simulation,* Orlando, Florida, Winter Simulation Conference, pp 2674–2680.

Kelton WD (2000). Design of experiments: experimental design for simulation. In: *Proceedings of the 32nd conference on Winter simulation,* Orlando, Florida, Society for Computer Simulation International, pp 32–38.

Kleijnen JPC (2008). Design of experiments: overview. In: *Proceedings of the 40th Conference on Winter Simulation,* Miami, Florida, Winter Simulation Conference, pp 479–488.

Komashie A and Mousavi A (2005). Modeling emergency departments using discrete event simulation techniques. In: *Proceedings of the 37th conference on Winter simulation*, Orlando, Florida, Winter Simulation Conference, pp 2681–2685.

Kool R, Homberg D and Kamphuis H (2008). Towards integration of general practitioner posts and accident and emergency departments: A case study of two integrated emergency posts in the Netherlands. *BMC Health Services Research* **8**(1): 225.

Kulu-Glasgow I, Delnoij D and de Bakker D (1998). Self-referral in a gatekeeping system: Patients' reasons for skipping the general-practitioner. *Health Policy* **45**(3): 221–238.

Law AM (2007). *Simulation Modeling and Analysis*. McGraw-Hill: Boston, MA.

Mackway-Jones K, Marsden J and Windle J Machester Triage Group (2006). *Emergency Triage*. Wiley: New York.

Mauro CA (1984). On the performance of two-stage group screening experiments. *Technometrics* **26**(3): 255–264.

Mes M and Bruens M (2012). A generalized simulation model of an integrated emergency post. In: *Proceedings of the Winter Simulation Conference*, Berlin, Germany, Winter Simulation Conference, pp 1–11.

Mielczarek B and Uziałko-Mydlikowska J (2012). Application of computer simulation modeling in the health care sector: A survey. *Simulation* **88**(2): 197–216.

Montgomery DC (2008). *Design and Analysis of Experiments*. Wiley: New York.

Paul SA, Reddy MC and DeFlitch CJ (2010). A systematic review of simulation studies investigating emergency department overcrowding. *Simulation-Transactions of the Society for Modeling and Simulation International* **86**(8–9): 559–571.

Sinreich D, Jabali O and Dellaert NP (2012). Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *IIE Transactions (Institute of Industrial Engineers)* **44**(3): 163–180.

Sinreich D and Marmor Y (2005). Emergency department operations: The basis for developing a simulation tool. *IIE Transactions (Institute of Industrial Engineers)* **37**(3): 233–245.

Thijssen WA, Giesen PH and Wensing M (2012). Emergency departments in the Netherlands. *Emergency Medicine Journal* **29**(1): 6–9.

Trocine L and Malone LC (2001). Experimental design and analysis: An overview of newer, advanced screening methods for the initial phase in an experimental design. In: *Proceedings of the 33rd conference on Winter simulation*, Arlington, Virginia, IEEE Computer Society, pp 169–178.

WCED (1987). *Our Common Future*. Oxfourd University Press: New York.

Yaesoubi R, Roberts SD and Klein RW (2010). Modification of cheng's method: An alternative factor screening method for stochastic simulation models. In: *Proceedings of the Winter Simulation Conference*, Baltimore, Maryland, Winter Simulation Conference, pp 1034–1047.

9

A Simulation Case Study to Improve Staffing Decisions at Mass Immunization Clinics for Pandemic Influenza

M. F. Beeler, D. M. Aleman and M. W. Carter
University of Toronto, Toronto, ON, Canada

Mass immunization clinics (MICs) are an important component of pandemic influenza control strategies in many jurisdictions. Decisions about staffing levels at MICs affect several factors of concern to public health authorities: total vaccination volume, patient wait-times, operating costs, and intra-facility influenza transmission risk. We present a discrete-event simulation of an MIC to assess how strongly staffing changes affect these factors. The simulation is based on data from Canadian clinics responding to pandemic H1N1 in 2009. This study is the first to model flu transmission risk at an MIC, and the first to relate such risk to staffing decisions. We show that the marginal benefit of adding staff is greatly underestimated if indirect waiting costs and intra-facility infections are not considered.

9.1 Introduction

Vaccination is an important component of pandemic influenza plans in many jurisdictions (Centers for Disease Control and Prevention, 2011; Public Health Agency of Canada, 2011; United Kingdom Department of Health, 2011). During the 2009 H1N1 pandemic, mass immunization clinics allowed for the rapid, large-scale administration of vaccines without overwhelming hospitals, regular walk-in clinics, and physician offices. Since vaccine development and distribution can take between 4 and 6 months from the onset of the pandemic (United Kingdom Department

Reprinted from *Journal of the Operational Research Society*, 65: 497–511, 2014, 'A Simulation Case Study to Improve Staffing Decisions at Mass Immunization Clinics for Pandemic Influenza', by M. F. Beeler, D. M. Aleman and M. W. Carter. With kind permission from Operational Research Society Ltd. All rights reserved.

of Health, 2011), vaccines may become available only in the pandemic's later stages, as was the case with the 2009 H1N1 vaccine, which started being administered in North America a couple weeks before the peak of the pandemic's second wave in November 2009.

The delayed availability of the vaccine has two important consequences. First, the effectiveness of the vaccination programme depends not only on the total percentage of the population vaccinated, but also on the programme's speed, as vaccinations given later in the pandemic have less value than those delivered earlier (Mylius *et al,* 2008). High-volume mass immunization clinics (MICs) offer such speed, but require high levels of staffing to meet public demand. Second, the vaccination programme will take place in the context of a healthcare system burdened by influenza cases. Pandemics increase demand for healthcare workers (HCWs), and may also decrease supply due to flu cases among HCWs (Mitchell *et al,* 2012). Consequently, public health authorities must make difficult decisions about how many staff to allocate to MICs, possibly at the expense of other areas of the healthcare system.

Through discussions with Niagara Region Public Health (NRPH) and Toronto Public Health, two public health units in Ontario, Canada, we found that decision-makers operating MICs want to balance a variety of objectives, including total vaccinations, patient wait times, overall operating cost and cost per vaccine administered, and infection prevention within their facilities. The first two objectives were of greatest concern. Wait times during the period of peak demand for vaccines sometimes exceeded 2 h in many parts of Canada (Yang, 2009) and the United States (Las Vegas Review-Journal, 2009; NY1 News, 2009).

Cost considerations are also important from both a short-term and long-term perspective. In the short term, MIC managers would benefit from knowing the marginal productivity of labour (MPL)—that is, the number of additional vaccinations achieved from adding or subtracting one HCW—and whether the MPL is the same in both directions, and how much it might vary when other events occur, such as surges in demand. The MPL is important for making economic decisions about whether to hire additional staff or allocate existing staff to other facilities.

The average cost to vaccinate a patient also factors into the ongoing debate about the cost-effectiveness of mass vaccination as a response to pandemic influenza. Mass immunization programmes in Canada and the United States during the 2009 H1N1 pandemic were criticized as expensive and excessive (Amico, 2009; Waldie and Alphonso, 2009; Blackwell, 2010), with total costs in Canada and the US borne by all levels of government being estimated at CAD$1 billion (Blackwell, 2010) and

USD$6.15 billion (Amico, 2009), respectively, though subsequent cost-effectiveness studies have attempted to address these concerns (Sander *et al*, 2010; Durbin *et al*, 2011). This study does not examine the overall cost-effectiveness of mass vaccination, which depends on many factors other than the effective operation of MICs.

Infection prevention at healthcare facilities is a priority throughout the sector (Centers for Disease Control and Prevention, 2010; World Health Organization, 2012). During a pandemic, infections can occur at any venue where people gather—hospitals, schools, workplaces, shopping centres, etc. The same holds true for MICs. Public health units in Ontario took care to communicate to the public that people experiencing influenza-like symptoms should not attend MICs, and the two public health units consulted in this study believed that the risk of influenza transmission at MICs from symptomatic patients was very low. However, patients with asymptomatic or very mild infections, or who are in the pre-symptom incubation period, which may last 1–4 days (Centers for Disease Control and Prevention, 2010), are believed to be infectious, albeit less so than patients with strong symptoms (Influenza Team, European Centre for Disease Prevention and Control, 2007). Reducing patient wait times has the additional benefit of further reducing the risk of influenza transmission within MICs.

This study uses a discrete-event simulation (DES) of an MIC to provide public health units with information about the likely impact of changes in staffing levels on these various measures of MIC performance. A range of scenarios are considered through a full-factorial designed experiment, including surges in demand, restrictions on vaccine eligibility, and extension of operating hours. Sensitivity analysis (SA) is conducted on several parameters that were not tested in the designed experiment.

The MIC model was built using data from MICs run by NRPH in Ontario, Canada, during the 2009 H1N1 pandemic. NRPH's MICs were run similarly to many of the large urban MICs run elsewhere in Ontario. Although some aspects of the MIC model may not overlap completely with MICs from other jurisdictions, the direction and approximate magnitude of several important effects reported in this study are likely generalizable.

9.2 Literature review

Several researchers have used DES to help plan the mass distribution of antibiotics and vaccines. Hupert *et al* (2002) developed a model to

inform staffing decisions at a mass antibiotic distribution centre to respond to bioterrorism threats, and Washington *et al* (2005) developed a model to optimize staffing assignments at a hypothetical smallpox mass vaccination clinic. Both studies had to rely on hypothetical inputs and expert opinion since no such clinics have been operated.

The potential value of simulation studies to emergency vaccination planning is further discussed by Asllani *et al* (2007). Although the authors do not present simulation results, they do discuss the differences between how responses to diseases such as seasonal flu, pandemic flu, smallpox, and anthrax should be modelled. Because pandemic influenza can be expected to affect a larger proportion of the population than an anthrax bioterrorism event, and often spreads faster than seasonal flu, simulation studies for other vaccine or antibiotic distribution problems might not adequately reflect the high-volume and high-speed operations that MICs for pandemic influenza need to achieve (Asllani *et al,* 2007).

Washington (2009) examined the cost and capabilities of a mass influenza and pneumococcal vaccination clinic using simulation, and made efforts to base the model off data from a real clinic's operations. Although the model's goal was to simulate a full-capacity response—what might be expected during a pandemic—the clinic that the model was based on was operated during the seasonal flu season and had turnout that was under 50% of the targeted capacity. Andress (2003) and Phillips and Williamson (2005) discuss several insights on MIC management arising from actual MIC operations, but do not specifically address how to optimize staffing decisions.

Simulations have been used extensively to understand the dynamics of influenza spread at the international (Khan *et al,* 2011), regional (Los Alamos National Laboratory, 2006; Stroud *et al,* 2007; Savachkin and Andres Uribe-Sanchez, 2012), and municipal levels (Elveback *et al,* 1976; Haber *et al,* 2007; Das *et al,* 2008; Gojovic *et al,* 2009; Sander *et al,* 2010; Lee *et al,* 2010; Aleman *et al,* 2011). However, no studies, simulation-based or otherwise, have been conducted on influenza transmission risks specifically within MICs.

There have been many studies examining seasonal and pandemic influenza transmission risk in various settings, including households (Viboud *et al,* 2004; Sikora *et al,* 2010), trains (Cui *et al,* 2011), airplanes (Moser *et al,* 1979; Klontz *et al,* 1989; Foxwell *et al,* 2011; Khan *et al,* 2011), cars (Knibbs *et al,* 2012), schools (Lessler *et al,* 2009), universities (Mei *et al,* 2010; Araz *et al,* 2011), and hospitals (Salgado *et al,* 2002; Noakes and Sleigh, 2009; Barnes *et al,* 2010; Wong *et al,* 2010; Bearden *et al,* 2012). For the most part, these studies document particular

outbreaks and discuss context-specific risk factors and infection control measures, and do not generalize well to MICs. For example, whereas patients at a busy MIC may spend up to 2 h at the facility and may come into proximity with a dozen or so other patients and 2–3 HCWs, a hospital inpatient may spend days in a hospital, making contact with a greater number of HCWs, possibly less contact per day with other patients, but also more contact with physical objects and surfaces in the facility. Furthermore, whereas MICs are for patients without symptoms of influenza, hospitals may house patients with acute influenza, as well as patients who are more susceptible to the virus.

The modes and magnitude of pandemic influenza transmission risk at MICs likely differ from those of other healthcare facilities. The techniques this study uses to estimate this risk are adapted from widely used methods in agent-based pandemic modelling at the regional and municipal levels (Elveback *et al*, 1976; Los Alamos National Laboratory, 2006; Haber *et al*, 2007; Stroud *et al*, 2007; Das *et al*, 2008; Aleman *et al*, 2011; Savachkin and Andres Uribe-Sanchez, 2012). These methods compute individual infection probabilities based on a susceptible person's exposure time to infected persons in different sub-environments (eg households, workplaces, schools), taking into account heterogeneous mixing patterns, and sometimes adjusted for factors that are believed to increase an individual's susceptibility or infectivity. All of these simulations work in hourly or daily time-steps. To offer a more granular picture of potential transmission risk at MICs, this study simulates and records susceptible patient interactions with infectious patients by the second.

This study is the first to simulate a pandemic influenza MIC using data from real clinics operated during a pandemic. The study builds upon the MIC management literature by considering a more complete range of facility performance measures and by testing the impact of staffing changes under different scenarios. It is also the first study to model the risk of influenza infection at an MIC during a pandemic, and to use such risk measures to help inform staffing decisions.

9.3 Model overview

The flow of patients through an MIC involves a modest number of decision points, queues, and processes, as illustrated in Figure 9.1. NRPH pioneered an electronic patient record system that was used throughout Ontario in 2009 to manage patients and record service-time data at MICs during the H1N1 pandemic. The system recorded a timestamp for every instant that a patient's record was accessed or closed at a service

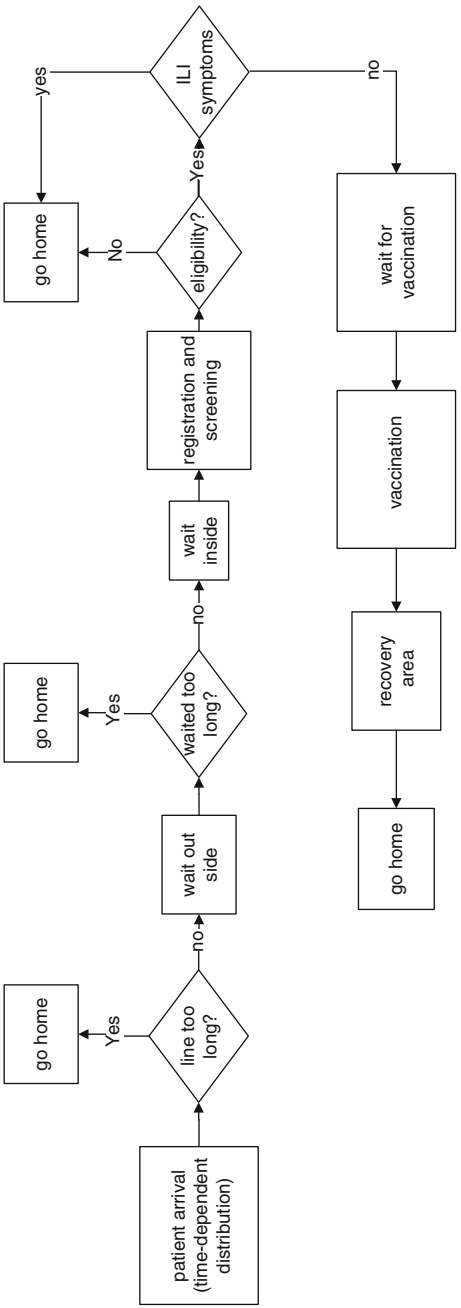


Figure 9.1 Flowchart of patient movement through MIC in Niagara Region, ON

station, corresponding in theory with their physical arrival and departure from the service station.

Data from this electronic system were obtained for 26–28 October 2009 at one facility, coinciding with the first three days of the public immunization programme. Data for 3, 6, and 12 November at a separate facility were also used, coinciding with a provincial policy that restricted vaccine access to prioritized groups. The combined data sets provided arrival and service time records for 9598 patients. The pandemic in Canada is considered to have peaked in early November (Public Health Agency of Canada, 2010). In addition, hourly staffing schedules and the number of patients seen per hour per staff member were also provided, along with staffing costs, floor plans, and facility operation instructions.

The input data from NRPH revealed several complex patterns that could not be captured with analytical queuing models. Patient arrival rates and service times at registration and injection stations varied according to the hour of the day, the age of the patient, and the size of the family group with whom patients arrived. It was also found that service times decreased as wait times increased. We conjecture that MIC staff could have increased their productivity in response to seeing long queues build up. Furthermore, the presence of prioritization policies that restricted vaccine access to certain groups based on age and health status had an impact not only on the total patient turnout, but also on the size of arriving groups and the patient age distribution.

The parameters used in the baseline scenario for the model are based on the staffing levels and patient demand experienced on 26 October 2009 at a particular MIC which was described by NRPH as operating at full capacity. The remainder of this section describes the model's components in greater detail.

9.3.1 Patient arrival rates

Patients who arrive at the MIC must wait outside if the facility is not yet open or at capacity. Patients arrive in batches ranging in size from one individual to families of five, starting 30 min before the official facility opening time at 8:30 am. The distribution of family batch sizes was provided by NRPH (Gentry, 2011). No incidences of families with six or more members arriving together were recorded. The arrival of family groups was modelled as a non-homogeneous Poisson process, with interarrival times being exponentially distributed. Arrivals were set to occur until 6:30 pm in a full-capacity scenario and 7:00 pm in a hypothetical overloaded scenario, marking the time that facility doors would close to new patients.

Patient arrival rates had to be estimated indirectly because patients were only recorded in the electronic record system once they reached the registration desk, after arriving and waiting in line. The arrival rates were calibrated to ensure consistency with several other known parameters: the total number of patients vaccinated, the maximum pre-registration wait time noted by staff, and the wait times and patient queue lengths within the facility between service stations. Differences existed in patient turnout between days and between facilities, requiring separate arrival rate calibrations to replicate each day's turnout. The baseline scenario used in this study reflects turnout and staffing levels at an NRPH clinic on 26 October 2009.

NRPH reported that patient arrival rates varied by the hour. Although patient arrival times were not recorded directly, NRPH's data show hourly variations in staff productivity that seemed consistent with the manager's description of the hours that were typically busiest (mid-afternoon and early evening). The interarrival times used for the 26 October baseline scenario are presented in Figure 9.2.

9.3.2 Patient characteristics

The patients are assigned labels indicating their health status, their age group, whether they belong to a government-defined priority group, the size of the group with which they arrived, and whether they

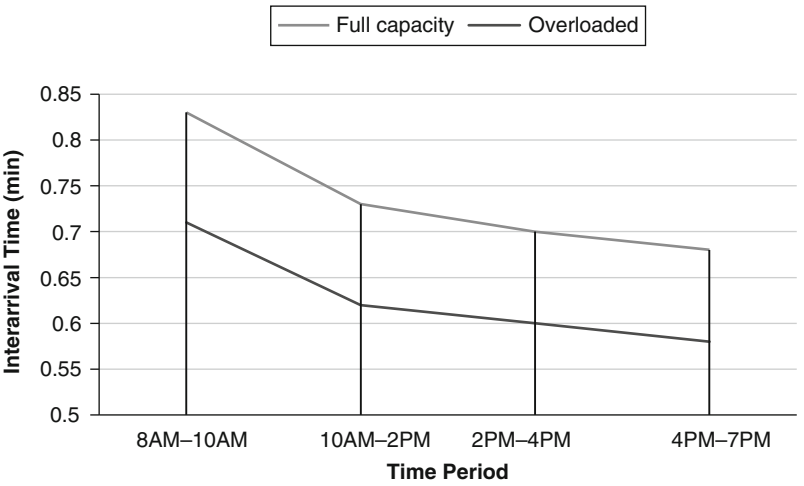


Figure 9.2 Interarrival times in minutes for patient batches in two scenarios

have eligible provincial health insurance. Data collected by NRPH were used to determine the proportion of patients assigned to each category.

Each patient is assigned one of two health states: susceptible or infectious. Public health officials in the Niagara Region reported that public communications efforts informing people not to come to MICs if they had influenza-like symptoms were largely successful. Consequently, the model assumes that infectious patients at the MIC were asymptomatic or had mild enough symptoms to prevent effective identification and screening. The simulation was run with 0, 2, and 3% of arriving patients being infectious, with the remainder of patients being susceptible.

Since transmission risk is believed to be age-dependent (Haber *et al*, 2007), patients were assigned to one of four age groups: 0–4, 5–18, 19–64, and 65+ years of age. The service times at registration and injection stations were also found to be age-dependent. The percentage of patients belonging to each age group depended on the government-mandated prioritization restrictions in place. Prior to 31 October 2009, no prioritization restrictions were in place. From 1–13 November, access to vaccines was restricted to the following groups: people under 65 with chronic medical conditions; pregnant women; children six months to five years in age; healthcare workers involved in pandemic response or the delivery of essential healthcare services; and care providers or household contacts of people at high risk who cannot be immunized. From 14–16 November, the restrictions were relaxed to include youth up to the age of 18, household members of pregnant women, first responders, and provincial corrections officers; however, data on the distribution of ages and group arrival sizes for this period were not available. From 17 November onwards, vaccine access was unrestricted. Table 9.1 contains age-group patient percentages for the initial unrestricted period and the first stage of restricted access.

NRPH's data indicate that expected service times at registration and injection stations also depend on the size of the patient's family group.

Table 9.1 Percentage of patients per age group, by vaccine priority restriction scenario (Statistics Canada, 2006; Gentry, 2010)

	Age range (years)			
	0–4	**5–18**	**19–64**	**65+**
Canadian population	5.84	18.06	63.38	12.72
Unrestricted access, 26–31 October	14.61	17.35	59.58	8.46
Restricted access, 1–13 November	16.58	16.32	61.16	5.94

Table 9.2 Percentage of patients per family batch size, by vaccine priority restriction scenario (Gentry, 2010)

	Family size				
	1	**2**	**3**	**4**	**5**
Niagara population	26.0	31.0	20.0	14.0	9.0
Unrestricted access, 26–31 October	35.0	28.0	20.0	13.0	4.0
Restricted access, 1–13 November	52.5	27.5	13.0	6.0	1.0

The likelihood of a patient belonging to a family group of a particular size receiving treatment on the same day is summarized in Table 9.2.

9.3.3 Patient wait-time tolerance, line tolerance, and facility closure

It is possible that some patients may balk at the size of the MIC's line and choose to leave, while other patients may reneg on their decision to get vaccinated if they have waited too long. Balking and reneging are well-established as phenomena in many queuing systems, including healthcare settings (Zenios, 1999; Fomundam and Herrmann, 2007; Asaro *et al*, 2008). However, the factors that can influence balking and reneging rates, such as the value and substitutability of the service, the possibility of postponement, and the customer's cost of waiting, vary highly across contexts. No empirical estimates for reneging and balking rates for mass immunization or sufficiently comparable health service settings have been published. Unfortunately, it was not possible to observe reneging and balking behaviour directly because the MICs could only record patients who made it to the registration desk.

The possibility of reneging and balking was built into the model to allow realistic bounds to be placed on outdoor wait times. These bounds were deemed particularly important for simulated scenarios in which there would be staff shortages and significant queue build-up. It seemed necessary to generate a bound of sorts since patients would have had a limited willingness to wait. Although reneging may be less common in healthcare settings where there are steep health consequences for postponing/cancelling service, or significant fines for missed appointments, MIC reneging does not impose such costs on patients. Since the particular reneging parameters used were informed guesses and not grounded in data, detailed SA on the reneging and balking parameters is presented

in the Results Section. Varying the reneging and balking parameters did not have a significant effect on total vaccinations, though outdoor wait times were affected.

For reneging, a patient's maximum tolerance for waiting outside the facility before leaving altogether was generated using a normal distribution with a mean of 90 min and a standard deviation of 20 min. For balking behaviour, the simulation gives patients a line length tolerance: if a newly arrived patient finds the number of people in line waiting to enter the facility is greater than their line tolerance, they will simply go home. Intuitively, a person's line tolerance and wait-time tolerance are likely correlated. Under this assumption, the line tolerance was set to be normally distributed with a mean of twice the wait-time tolerance level and a standard deviation of 20 min. An hour before the facility closes at 6:30 pm, the MIC stops taking new patients and closes its doors. Patients still outside at this time go home while the patients inside proceed through the MIC.

9.3.4 Patient screening

It was assumed that at least 0.5% of patients arriving at the facility would not be eligible to receive a vaccination by virtue of not holding provincial health insurance or not belonging to an eligible priority group. In the model, such patients are sent home after being screened at the registration desks. Patients with a health state involving symptoms of influenza-like-illness (ILI) are also screened at this stage. The effectiveness of screening can be set to less than 100% to reflect the possibility that some patients with mild ILI symptoms may want vaccination and consequently claim to be feeling fine to avoid being sent away. The number of patients turned away for these reasons at actual MICs in the Niagara Region is not known.

9.3.5 Registration and vaccination stations

Once inside the MIC, patients enter the registration queue, which has a 50-person capacity. At capacity, patients are unable to enter the facility and wait outside. At registration, eligible patients provide clerical staff with personal data and answer questions about co-morbidities and possible contra-indications. Following registration, patients enter a new queue to receive the vaccine from nursing staff. The queue for vaccine administration has a maximum capacity of 250 to reflect space constraints. When this queue reaches capacity, the registration desks are programmed to stop processing patients until space becomes available. After being vaccinated, patients wait in a recovery area for 10 min in

case they have an adverse reaction to the vaccine. Then, patients leave the facility.

Each point of service—registration, flu assessment, and vaccination—is assigned a distribution for service times, as well as a certain number of staff, each of whom serves one patient at a time. Two data sets were used to generate the service time distributions. The first data set consisted of entries from NRPH's electronic patient record system that captured the length of time that a patient's electronic file was open at each service station. The second data set contained the number of patients served by each staff member over each hour of the day. The service times recorded in the electronic system were found to be significantly lower than those implied by the hourly throughput and the overall number of patients vaccinated by the facility, even when taking into account break time and minor inefficiencies. Consequently, the mean service time was estimated using the number of patients seen per hour per staff member. The service times recorded in the electronic patient system were still used to obtain standard deviation estimates, to infer a likely shape of service time distributions, and to estimate the effect of patient age on service time.

The sample standard deviation of the registration times from 26 to 28 October 2009 was 1.16 min, and 1.25 min for vaccination times. The distribution of service times in the electronic record system at both registration and vaccination stations was positively skewed (Figure 9.3), and, of course, non-negative. Lognormal distributions were chosen to generate both registration and vaccination times so as to replicate this positive skew and non-negativity.

The effects of patient age group on service time was found through fitting an ANOVA model, and are summarized in Table 9.3. All effects are relative to the 0–4 age group. The data suggest that adults and seniors take roughly 32 and 45 s longer to register than small children, respectively, perhaps because they are more likely to have more complicated medication and health histories to document at the registration desk. The data also show that children under 5 years require slightly more time to vaccinate than older patients. These findings mean that the optimal mix of registration staff and vaccination staff could change depending on the demographic profile of the patient population. Based on these numbers, if 1000 single adults attending a large MIC were substituted with additional children in the family batches, 8.9 fewer registration clerk hours and 6.7 additional nurse hours would be required to meet demand, all else held equal. Since vaccine prioritization rules are often based on age, major changes in patient demographics are quite likely and may affect staff planning.

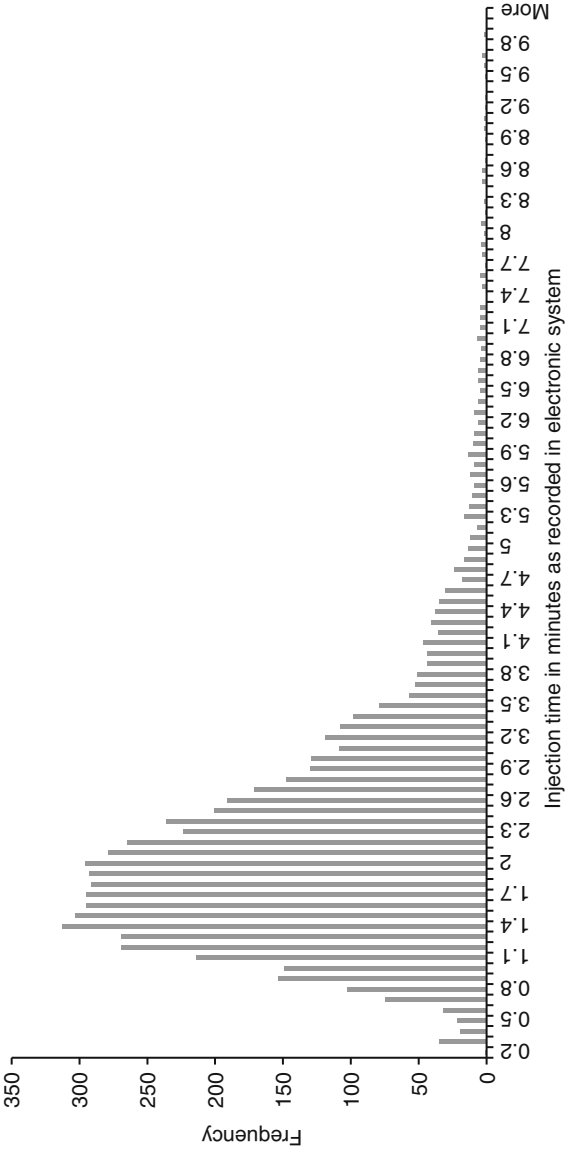


Figure 9.3 Histogram showing positive skew for injection times recorded in electronic system, 26–28 October 2009

Table 9.3 Effects of patient age on service time (minutes), relative to service times for the 0–4 age group

Age range	Registration		Vaccination	
	Effect	p-value	Effect	p-value
5–18	0.047	0.043	–0.22	0.050
19–64	0.550	<0.001	–0.40	<0.001
65 +	0.760	<0.001	–0.30	<0.001

Table 9.4 Effects of family group arrival size on registration time per-patient

Family batch size	1	2	3	4	5
Expected registration time (min)	3.60	3.48	3.25	3.18	3.28

In addition to age-based adjustments, the expected value of a patient's registration time also depends on the size of the family group with whom she/he arrived (Gentry, 2011), since address information and family name spelling might not need to be repeated for additional family members. Initial estimates for these service time differences were based on the registration time-stamps from the electronic patient record system. However, the service times recorded electronically seemed to represent only a quarter of the time registration staff spent at the registration desk registering patients based on our analysis of NRPH's staffing data. Consequently, the expected registration times used in the simulation were set to four times the average values calculated using the electronic records, and are presented in Table 9.4. The weighted average of these registration times ranges from 3.42 to 3.49 min, depending on which prioritization policy is in place, and is close to the rate of 3.45 min calculated indirectly by dividing total minutes believed to have been worked by the total patients registered over those days.

Vaccination times did not vary with the patient's family size. However, the data sets did reveal a difference in productivity between nurses with different professional specialties. In particular, vaccine preventable disease nurses and infectious disease nurses, who represented just under half the nurses at the MICs, served patients about 12% faster than other nurses ($p = 0.0109$), though this effect was only statistically significant for the earliest two of the six days for which data were provided.

9.3.6 Changes in staff productivity by hour

Although the MIC managers believed there to be a consistent stream of patients served by staff working at full capacity, the data sets revealed significant variation in staff productivity over different hours of the day, across different days, and among staff members. Although productivity on 26 October for both nurses and registration staff was lowest in the morning and highest in the 2 h before closure, the variation on 27 October did not exhibit a clear trend, while productivity on 28 October was highest in the early afternoon, as shown in Figure 9.4. The extent of the variation is significant, but its causes are unclear.

The differences among staff were also quite high, even though the MIC managers described the 26–28 October 2009 period as especially busy, with staff working non-stop and very hard. For example, the standard deviation of each registration staff member's average hourly productivity during the unrestricted vaccine access stage was 5.5 patients per hour, or just over 25% of the average rate of 20.0 patients registered per hour. Variation between nurses at vaccination stations was just as large. On most days, the most productive nurses and registration staff served twice as many patients per hour as their counterparts with the lowest productivity. When compared with the most productive hour worked by any staff member on a given day, average staff productivity was only

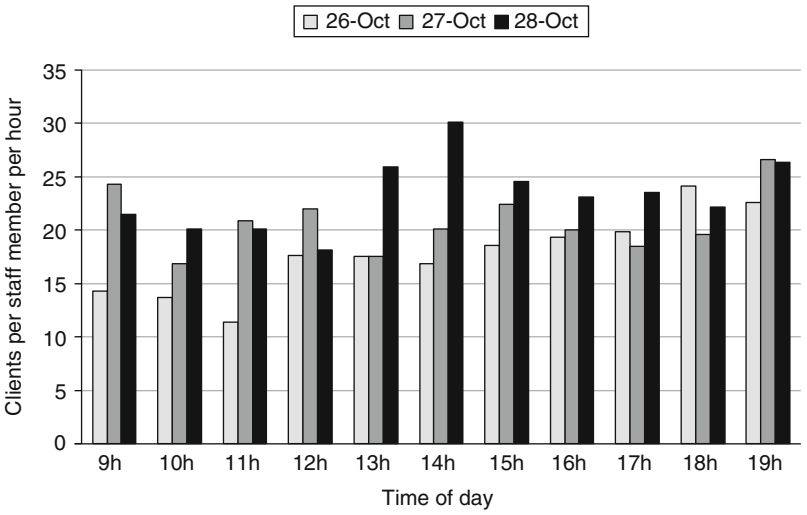


Figure 9.4 Registration staff productivity by time of day

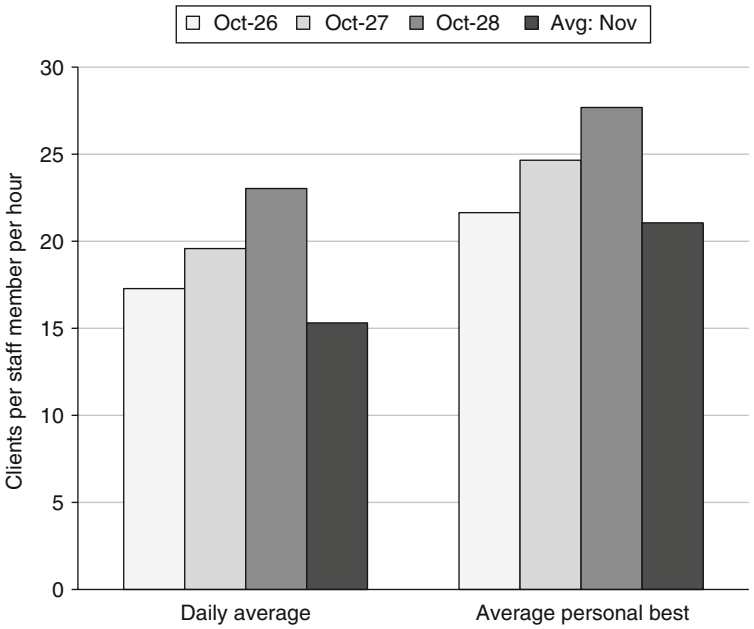


Figure 9.5 Comparison of hourly productivity, 26–28 October and November

46–56% of the top performer's best hour. This gap has two components: staff members' average hourly productivity was 81–85% of their own personal bests for the day; and the averages of the daily personal bests were only 58–70% the level of the top performer's personal best. Figure 9.5 shows how average hourly productivity and average personal best productivity of registration staff varied over three consecutive days during the busiest period, while falling in the month of November, in part due to lower MIC utilization.

The number of people visibly waiting could have put greater pressure on staff members, leading to a faster work pace. However, the data sets did not contain information on outdoor queue lengths at different times of the day, so this hypothesis could not be verified. It is not known whether small errands may have disrupted some nurses assigned to vaccination stations. Such disruptions could bias these productivity estimates downwards, while upwardly biasing the productivity variance.

Since an effect of queue length on service time seemed plausible, the service times for a patient in the simulation were reduced by 4.8 s for every 10 people waiting behind them in line, inside or outside the facility. This adjustment helped the simulation replicate some of the

productivity changes observed over each day, with faster service times coinciding with periods when the lines were longest. It is not known whether this positive relationship between line length and service times was present at other facilities.

9.3.7 Operating costs

To calculate the average cost per vaccine administered, labour costs for registration clerks and nurses, vaccine material costs, and space rental costs were included in the model (Table 9.5). The cost of administration, information technology, and greeters was not included, so the overall cost per vaccine in the model is conservative. The space rental costs is an estimate of what a public health unit would have to pay in the event that a free public facility is not available.

9.3.8 Risk of disease transmission

Influenza transmission is believed to be unlikely at distances greater than 2 m from an infected person (Brankston *et al*, 2007). The risk of a contagious patient infecting a susceptible patient is modelled using the function $1 - e^{-\lambda_{ij}t}$ (the cumulative density function of an exponential distribution), where t is the amount of time the two patients spend within 2 m of each other, and where λ_{ij} is the age-dependent hazard rate (probability of infection) for a minute of exposure between susceptible person i and infectious person j, given their respective age groups. The difference in transmission risk for asymptomatic, pre-symptomatic, and mild-symptom cases compared with fully symptomatic cases is not known, but believed to be less severe (Influenza Team, European Centre for Disease Prevention and Control, 2007). To err on the side of caution, the transmission risk rates used in the MIC simulation represent average transmissibility calibrated in the population-level influenza simulation study by Haber *et al* (2007). These rates are given in Table 9.6.

Susceptible patients can be exposed to multiple infectious persons simultaneously. The patient's hazard rate for infection at a given point

Table 9.5 MIC operating costs (CAD)

Item	Cost	Comments
Vaccine cost per unit	$10	Government purchase price (CBC News, 2009) + shipping
Daily space rental	$500	N/A for NRPH's MICs
Hourly nurse wage plus benefits	$48	Actual (Gentry, 2011)
Hourly clerk wage plus benefits	$27	Actual (Gentry, 2011)

Table 9.6 Pandemic influenza transmission hazard rates (λ) per minute (Haber *et al*, 2007)

Age group of infectious person	Age group of susceptible person			
	0–4	5–18	19–64	64+
0–4	0.00059	0.00062	0.00033	0.00080
5–18	0.00058	0.00061	0.00033	0.00080
19–64	0.00057	0.00053	0.00032	0.00080
64+	0.00057	0.00054	0.00029	0.00102

in time is set to the sum of the individual hazard rates for all of the infectious persons with whom they are making contact at that point in time. Since the number and type of infectious persons a susceptible person may be exposed to will change as she/he progresses through the MIC, the simulation re-assesses with whom the susceptible person has contact every time his/her position in the MIC changes.

To track total exposure time to contagious patients, each susceptible person is assigned an exposure variable that gets updated at every time-step k. To update the exposure variable, the simulation must determine which individuals are within 2 m of a given susceptible individual, and then check their health status. For computational speed, the 2-m radius was assumed to include individuals within M ranks on either side of the individual's position in a queue during time t_k. It was assumed that $M=3$ for queues inside the MIC and $M=4$ for queues outside the MIC entrance. The maximum number of people with whom a patient can have contact at a given time is therefore $2M$. To represent greater congestion or adjacent single-file queues, M can be increased appropriately.

Letting $\mathcal{K}_i$ represent all time intervals k over which susceptible person i occupies a distinct position in the MIC system, and letting the set $\mathcal{I}_{ik}$ represent all infectious patients with whom the susceptible person has contact during time period t_k, the probability of a susceptible person i becoming infected in the MIC can be represented as $1-e^{-\sum_{k\in\kappa_i}\sum_{j\in\mathcal{I}_{ik}}\lambda_{ij}t_k}$.

The expected number of infections per day is the sum of the probabilities of infection for each susceptible individual in the system that day. The simulation does not consider the probability of transmission to HCWs on duty since the transmission risk to individuals wearing protective masks is likely low, and it is likely that most MIC workers would be vaccinated.

In very long queues, the people to whom a susceptible person is exposed is re-assessed every time the person's rank in the queue changes, even though this set of people may remain constant for up to 60 min. This superfluous calculation can be rectified by simply multiplying the sum of the hazard rates from those same individuals by the time interval over which the patient is in queue with them. However, even without this reduction in computations, the model's one-day run time is under 90 s.

9.4 Model validation

It is not possible to validate the estimates of the number of patients becoming infected at an MIC against empirical data on infection risks in MICs or in other public venues because no such empirical data exist.

The infection risk parameters used in this simulation were originally generated for use in a population-level influenza spread model (Haber *et al*, 2007) so as to ensure that the proportions of the simulated population in different age groups that became infected would resemble the proportions from past pandemics. It is important to recognize that all infection risk parameters calculated in this way are limited for a number of reasons: the set of parameter values that could lead to the desired simulated pandemic outcomes is not necessarily unique; the infection risk parameters depend on the assumptions made about social contact patterns, which change by culture, geography, and socioeconomic conditions; infection risk is affected by ambient conditions such as relative humidity and temperature (Tellier, 2006); and there is no consensus on which physical mechanisms of transmission are most important (Tellier, 2006). In light of these limitations, the infection counts produced by this simulation model should be viewed as ballpark estimates, subject to revision as more becomes known about the probability of influenza transmission at the individual level.

Validation for the overall design of the model was done by setting staffing and arrival rate levels to match those of the actual MICs for which NRPH provided data, and comparing the empirical patient count, post-registration wait times, and indoor queue lengths with those generated by the simulation. In order to ensure that the simulation's outputs were within 10% of the empirical outputs, it was necessary to reduce nursing staff availability by 5%, that is, nurses would be at their vaccination stations only 95% of the time, not counting official breaks. This change may reflect the fact that nurses sometimes had to serve other functions in the MIC, such as collecting vaccination supplies from storage and distributing them to other nurses.

9.5 Results

The simulation model was built using Simul8 (Visual8, Mississauga, ON) and run on a Windows 7 computer with a 2.1 GHz processor and 4GB of RAM. The peak-capacity (baseline) scenario and several hypothetical scenarios were simulated to understand the relationship of infection risk, throughput, and costs.

From an economic point of view, the results offer strong reasons to invest additional resources in MICs in order to reduce wait times. Shorter wait times also mean less exposure to contagious patients—an additional public health benefit. A designed experiment was performed to examine how changes in a variety of factors affect several measures of MIC performance.

9.5.1 Baseline scenario

An 'as-is' baseline configuration of the MIC model was created based on the staffing levels, patient turnout, and hours of operation of an MIC operated by NRPH on 26 October 2009, described by management as operating at full capacity. On this day, no restrictions were in place on who could access vaccines. Nine registration staff and 11 vaccination nurses (four VPD/ID RNs and seven regular RNs) were on duty throughout the day. The facility opened its doors at 8:30 am and closed them at 6:30 pm, at which point it would only serve patients already inside the facility. Two per cent of patients were assumed to be infectious and without symptoms. The response variables tracked in the simulation and their results are presented in Table 9.7. Patients with influenza but no symptoms were not included in the number of people vaccinated or in the cost per vaccination calculation since their vaccinations would not benefit them.

The trials show that the expected number of infections occurring in the MIC is roughly 0.55% of the number of people vaccinated. The count is still high enough to merit attention to transmission risk in MICs, but does not detract from the benefits of mass immunization.

Table 9.7 Baseline scenario results, 21 replications

Response variable	Value	95% confidence interval
Expected infections	9.08	(8.54, 9.63)
Number vaccinated	1658.8	(1645.5, 1672.2)
Direct cost per vaccination (CAD)	16.76	(16.67, 16.85)
Societal cost per vaccination (CAD)	32.32	(31.94, 32.70)
ATS—Average time in system (min)	90.89	(88.87, 92.91)

The average number of patients registered per full-day clerk was 184.3, and the average number vaccinated per full-day nurse was 150.8.

Although patients spent close to 91 min on average at the MIC, less than 10% of that time was spent receiving services from MIC staff. Most of the time was spent waiting, and a large part of that wait time took place outside. The total number of person-hours spent by the public at the MIC, including people who did not get into the facility in time or who were screened on the basis of eligibility, was 2644.9, with a 95% confidence interval of (2516.9, 2772.9). Patients who arrived and then left immediately because the line length exceeded their line tolerance were excluded from the average time in system (ATS) calculations.

It would be unrealistic not to assign some value to this lost time, which ultimately does impose a cost on the economy and society, directly by taking up formal work time, or indirectly taking away time from domestic activities. Wait times were penalized with a cost of $10 per hour, roughly Ontario's Minimum Wage in late 2009, to allow us to estimate the total social cost per vaccination, comprising of both lost time and MIC expenditures. Taking wait times into consideration in this way roughly doubled the cost per vaccination from $16.76 to $32.32.

9.5.2 Sensitivity analysis

Since the number of distinct scenarios to test in a full-factorial designed experiment grows exponentially in the number of factors, it was not possible to test the simulation for all possible sources of variation via the experiment. To compensate, oneway SA was conducted on several uncertain parameters not tested in the experiment. The results for response variables of interest are presented.

9.5.2.1 Patient interarrival times

The distribution of patient arrival times in the baseline scenario was chosen to ensure that the number of persons receiving vaccination and the ATS were consistent with the values for the reference clinic that the baseline scenario is supposed to imitate. The reference clinic served 1699 patients on 26 October 2009. Since the model only counts vaccines given to the 98% of patients who are uninfected as useful vaccinations, the arrival rates were calibrated to yield an average vaccination count close to $0.98 \times 1699 = 1665$. Patient interarrival times also impacted ATS. The target ATS value range for the baseline clinic was 85–95 min, including time spent in line outside the facility prior to registration. The interarrival times for all periods of the day were varied by a multiplicative factor applied to the baseline values, with 10 replications per case. The results are presented in Figure 9.6.

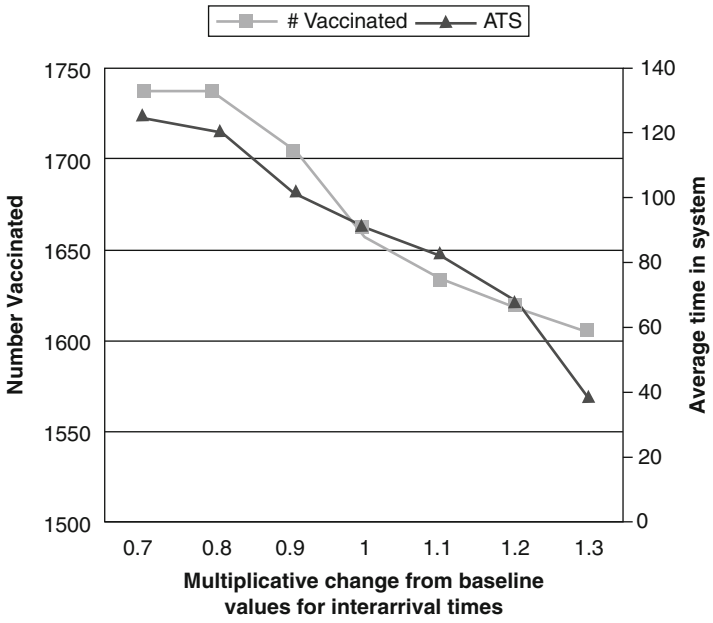


Figure 9.6 Vaccinations and average time in system *versus* patient interarrival time adjustment

The range of variability for total vaccinations is bounded above by the physical capacity of the facility, with small increases happening alongside faster arrival rates as a consequence of the slight relationship between visible queue lengths and staff service times assumed in the model. The total number vaccinated did not fall proportionally with the lower rates of arrival because a large portion of patients in the baseline scenario were going home prior to entering the facility because of the long outdoor queue. When the rate of arrivals was slower, the queue was shorter, resulting in a higher proportion of the arrivals staying in the queue and entering the facility—this partially compensated for the slower arrival rate. In addition, while the ATS increased with higher turnout, patient wait time tolerance limited ATS growth by causing some patients to go home without entering the MIC after having waited in the outdoor line.

9.5.2.2 Patient wait time tolerance

Patient wait time tolerance was also varied by using multiplicative factors, with 10 replications per case. The results are presented in Figure 9.7. The effect on total vaccinations was limited, but more noticeable for ATS.

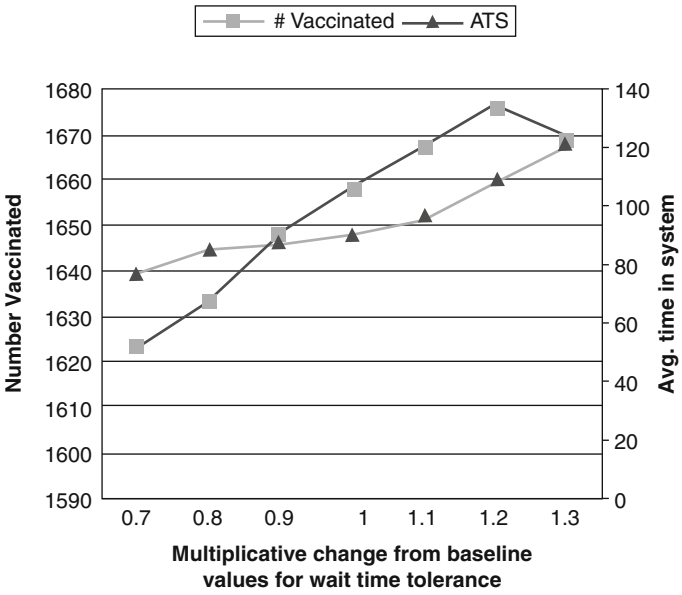


Figure 9.7 Vaccinations and average time in system *versus* patient wait time tolerance adjustment

9.5.2.3 Transmission parameters

Many factors are believed to influence the risk of influenza transmission, such as relative humidity and temperature (Tellier, 2006), ventilation (Moser *et al*, 1979), and intensity of symptoms (Influenza Team, European Centre for Disease Prevention and Control, 2007), though these effects have not been quantified in studies involving realistic transmission between humans. To account for this uncertainty, the hazard rate parameters for influenza transmission were varied by 50%, holding all other parameters at their baseline values. Twenty replications per case were used. The mean number of infections in the lower and higher risk scenarios was 4.23 and 13.51, with 95% confidence intervals of (3.24, 5.22) and (11.80, 15.22), respectively, compared with 9.08 infections in the baseline. The percentage change in the number of infections is similar to the percentage change in the transmission rates.

9.5.2.4 Arrival of symptomatic patients

In the baseline scenario it was assumed that patients with symptoms of influenza would not come to the MIC. Relaxing this assumption, we

let an additional 1% of the population be infectious and symptomatic, and allowed for screening at the registration desk with 100, 75, and 50% efficacy. Twenty replications per case were used. The mean number of infections in these three cases was 11.20 (9.82, 12.58), 11.98 (10.83, 13.13), and 12.56 (11.57, 13.57), respectively. The increase in the number of infections, although small, was statistically significant in all cases. As the screening effectiveness decreased, infections increased slightly, but the increases were not statistically significant.

9.5.3 Designed simulation experiment

A simulation experiment was conducted to determine the effects of the six factors presented in Table 9.8 on total vaccines administered, ATS, expected infections, and costs per vaccination. Data from a full-factorial design with one run per scenario were combined with 20 additional replications of the baseline scenario to fit a linear ANOVA model with main effects and several interactions of interest. A total of 344 simulation runs

Table 9.8 Factors examined in experiment

Factor	Levels
Factor 1: Priority access restrictions	Level 1: No restriction; patient demographic is similar to actual patient in NRPH during H1N1 Level 2: No restriction; patient demographic is representative of Canadian population Level 3: Restricted to priority groups
Factor 2: Nurse staffing	Level 1: Maintain 11 nurses from baseline scenario Level 2: Reduce the number of specialist RNs by one Level 3: Increase the number of specialist RNs by one
Factor 3: Registration clerk staffing	Level 1: Maintain nine registration clerks from baseline scenario Level 2: Reduce the number of clerks by one Level 3: Increase the number of clerks by one
Factor 4: Hours of operation	Level 1: Maintain doors open at 8:30 am and close at 6:30 pm Level 2: Doors open at 8:00 am and close at 7:00 pm
Factor 5: Infection rate	Level 1: Maintain at 2% of patients as asymptomatic infectious Level 2: Increase to 3% of patients asymptomatic infectious Level 3: No patients are infectious
Factor 6: Surge in patient demand	Level 1: Patient arrival rate based on 26 October 2009 turnout Level 2: 15% surge in patient arrival rate

were used in the data analysis. The results for each response variable are summarized in Table 9.9. Only effects that are both economically and statistically significant effects are presented. The values in the tables represent the relative difference from the baseline scenario described earlier.

Table 9.9 Factors affecting number vaccinated, ATS, expected infections, direct costs per vaccination, and total societal costs per vaccination

Significant factors	**Change from baseline**	**p-value**
Number vaccinated		
Turnout representative of population	31.48	<0.001
Restriction to priority groups	–144.55	<0.001
Increment percent infectious	–18.44	<0.001
No patients infectious	31.85	<0.001
Extra hour	105.05	<0.001
Remove one nurse	–126.06	<0.001
Add one nurse	50.737	<0.001
Remove one clerk	–39.25	0.0045
15% surge in patient demand	69.28	<0.001
Interaction: priority restriction AND remove one nurse	77.75	<0.001
Interaction: priority restriction AND add one nurse	–68.31	<0.001
Interaction: representative turnout AND add one nurse	29.53	0.025
Interaction: add one nurse AND add one clerk	28.25	0.032
Interaction: surge AND remove one nurse	–35.76	0.0011
Interaction: surge AND add one nurse	45.44	<0.001
Average time in system		
Turnout representative of population	5.18	0.0021
Restriction to priority groups	–37.7	<0.001
Extra hour	–7.48	<0.001
Remove one nurse	14.97	<0.001
Add one nurse	–14.86	<0.001
Remove one clerk	3.88	0.021
15% surge in patient demand	18.68	<0.001
Interaction: priority restriction AND add one nurse	4.99	0.012
Interaction: priority restriction AND remove one clerk	8.01	<0.001
Interaction: remove one nurse AND remove one clerk	–4.96	0.013

Table 9.9 Continued

Significant factors	Change from baseline	p-value
Expected infections		
Restriction to priority groups	–5.08	<0.001
Increment percent infectious	3.56	<0.001
No patients infectious	No infections	automatic
Add one nurse	–1.05	0.013
15% surge in patient demand	2.53	<0.001
Interaction: priority restriction AND increment % infectious	–1.12	0.040
Interaction: priority restriction AND remove one nurse	1.27	0.020
Interaction: surge AND increment percent infectious	0.97	0.032
Direct cost per vaccination		
Increment percent infectious	0.592	<0.001
No patients infectious	–1.23	<0.001
Add one clerk	0.156	0.0029
Interaction: priority restriction AND none infectious	0.372	<0.001
Interaction: priority restriction AND add one nurse	0.382	<0.001
Interaction: surge AND none infectious	–0.162	0.031
Interaction: surge AND add one nurse	–0.223	0.0030
Interaction: surge AND remove one clerk	–0.206	0.0063
Total societal cost per vaccination		
Turnout representative of population	0.67	0.027
Restriction to priority groups	–6.35	<0.001
Increment percent infectious	0.58	<0.001
No patients infectious	–1.62	<0.001
Remove one nurse	2.81	<0.001
Add one nurse	–2.21	<0.001
Remove one clerk	0.61	0.046
15% surge in patient demand	3.39	<0.001
Extra hour	–1.36	<0.001
Interaction: priority restriction AND add one nurse	1.24	<0.001
Interaction: priority restriction AND remove one clerk	1.43	<0.001
Interaction: remove one nurse AND remove one clerk	–0.85	0.020
Interaction: surge AND add one nurse	–0.70	0.020

9.5.3.1 Number of vaccinations

All factors have a statistically significant effect on the number of susceptible patients receiving vaccinations. The effect of the priority restriction policy in lowering turnout reflects the fact that these policies make it difficult for large families to arrive and get vaccinated all at once, meaning more of the arrivals will consist of individuals or couples. This decline in turnout is consistent with the decline observed in November when the priority restrictions were actually in place (Gentry, 2011).

When patient turnout was representative of the population, all response variables were affected. This factor level was included to determine whether it is reasonable to assume that the patient population is representative of the actual population without impacting the simulation's output. The differences in response variable outputs between scenarios using a representative population *versus* the actual population profile show that this is not the case.

Marginal changes in the number of nurses had significant effects on the number of patients vaccinated, but the size of the effect depended on other factors. In the baseline scenario, adding one nurse led to a small gain of only 50.7 new vaccinations, while removing one nurse led to a loss of 126.1 vaccinations. However, the benefit of adding a nurse is 28.2 vaccinations greater if an extra clerk is also added, though adding a clerk alone had no significant effect on total vaccinations, and removing a clerk led to a meagre decline in vaccinations of 39.2. The benefit of adding a nurse completely disappeared under the priority restriction scenario, when turnout was about 8.7% lower. Likewise, the consequence of losing one nurse—126.1 forgone vaccinations in the baseline—was 77.7 vaccinations less severe under the priority restriction scenario. These results indicate that changes in patient demand have a very strong effect on the MPL, and that the marginal impact of a unit change in staffing is not symmetric with respect to the direction of change.

The results in the surge scenario show a similar interaction between nursing levels and patient demand. The 15% surge in demand alone led to an extra 69 patients being treated. The increase in treated patients likely results from the model's assumption that staff work slightly faster in response to seeing longer lines. If this assumption is incorrect, a surge might not lead to any additional patients being treated compared with a clinic that is already close to full capacity. The surge also amplified the effect of changes in nursing levels: Adding a nurse led to 45.4 more vaccinations than would have been added by an extra nurse in the baseline scenario, and removing a nurse led to 35.8 more vaccinations lost than the 126.1 lost under regular demand. Although average hourly

clerk and nurse productivity was close to 190 and 150 patients served, respectively, in no scenarios is the effect of changing clerk levels close to the average clerk productivity, and only in some scenarios is the effect of adjusting nurse levels comparable.

The vaccination count declined by 18.44 when the percentage of patients who are contagious was incremented; this drop reflects the effect of not counting vaccinations given to the additional people who are infected or who become infected. Extending operating hours by 60 min, which is almost a 10% increase in the amount of time the facility's doors are open, led to a 6.3% increase in vaccinations.

9.5.3.2 Average time in system

The decline in patient turnout under the priority restriction scenario led patients' ATS to fall by 37.7 min, or 41.5%. In contrast, the 15% surge in patient demand only increased ATS by 18.68 min, or 20.6%. Extending operating hours by 1 h further reduced average wait times by 7.48 min, or 8.2%. Removing and adding one nurse had a symmetrical effect of roughly 15 min in either direction. Removing a clerk increased ATS by 3.88 min, while adding a clerk brought no benefit.

Under the priority restriction scenario, the benefit of adding a nurse was 5 min lower, and the effect of removing a clerk was 8.0 min stronger. The interaction between removing a clerk and the priority restriction sceario may be due to the shift in family sizes towards singles and couples, whose higher registration times could have increased the sensitivity of ATS to changes in clerk staffing levels. The coefficient for the interaction between removing a nurse and clerk simultaneously is negative, suggesting that while removing nurses and clerks one at a time increases ATS, the combined effect of removing both is less than the sum of the effects of removing them individually.

9.5.3.3 Expected number of infections and infection risk

Several factors affected the expected number of infections, in part by reducing the number of patients arriving at the MIC—not necessarily a good thing—but also by reducing patients' ATS. The baseline expected number of infections was 9.08, representing 0.55% of those vaccinated. The priority restriction policy cut expected infections by 55.9%, and resulted in a relative decline in infection risk per patient of 51.7%. Incrementing the percentage of patients infected from 2 to 3% led to 3.56 new infections on average, an increase of 39.2%. The 15% surge in patient demand led to a 27.9% increase in infections and a 22.7% increase in per patient infection risk, due largely to longer ATS.

Adding one nurse averted 1.05 infections, which seems small but is an 11.6% improvement. Furthermore, from a societal perspective, the extra infected person's lost productivity over several days is likely of greater economic value than the cost of the additional nurse.

Several factor interactions were also present. The effects of the surge and of incrementing the percentage of infectious patients amplified each other slightly. The consequence of increasing the percentage of infectious patients was less severe in the priority restriction scenario. The reduction in expected infections from 9.08 to 4.00 under the priority restriction scenario was 1.27 infections less when one nurse was removed, though the effect of removing a nurse was not statistically significant in the baseline scenario.

9.5.3.4 Direct and societal costs per vaccination

Although four factors had statistically significant effects on the cost per vaccination, the effects were fairly modest. The cost per vaccination is 7.4% higher in the baseline scenario than when no infectious patients are in the model, a fact which indicates that the small percentage of vaccines given to people who will not benefit from them has economically meaningful implications. Adding a clerk to the baseline scenario increased the cost per vaccination, as did adding one nurse in the priority restriction scenario since demand was lower. Meanwhile, the cost per vaccination decreased in the surge scenario when a nurse was added and when a clerk was removed, indicating that the MIC may have best coped with a surge, given limited resources, by exchanging a clerk for a nurse, bringing the nurse:clerk ratio to 3:2.

If the economic loss due to patients waiting is taken into consideration, adding a clerk no longer hurts the cost per vaccination, and the benefit of adding a nurse is larger: 6.8% improvement in the baseline cost level and a 9.0% improvement in the surge cost level as opposed to no improvement in the baseline and a 1% improvement in the surge scenario. The factors that affected ATS also affected the societal cost per vaccine, and in similar ways. This cost per vaccine measure strikes a balance between higher patient throughput per worker per hour and keeping wait times short.

9.6 Conclusions

The simulation experiment's results show that making MIC staffing decisions on the basis of reaching vaccination targets, without considering the societal/economic losses associated with long waits

or the possibility of intra-facility infection risk, could lead to under-investment in MICs. If an hour of a patient's time is valued at the minimum wage, then the cost of wait times of the length seen during the busiest days of the 2009 H1N1 pandemic becomes comparable with the direct financial cost of MIC staff, space rental, and vaccine materials combined. In addition to reducing wait-times, adding staff in the right circumstances also modestly reduced the expected number of intra-facility infections.

The analysis of staff productivity data revealed that there could be potential to increase patient throughput at MICs without hiring additional staff. Even though staff appeared to be working at full speed throughout the day, staff members' productivity varied considerably throughout the day, across different days, and between individuals. However, the reasons for these productivity variations remain unclear.

This simulation study is the first to estimate the risk of influenza transmission in a mass immunization clinic. As with all simulation studies, the results presented are contingent on the model's parameters and simplifying assumptions. Significant potential exists for gains in model accuracy and precision on several fronts. Transmission hazard rates within a 2-m radius are likely not well represented because virus-laden droplet and/or aerosol density in the air will become more diffuse as the planar surface area, or volume, of space surrounding an infectious person increases with distance. The possibility of different infection risks for persons standing in front of an infected person *versus* behind was also not considered.

Possible improvements to the simulation model include the incorporation of direction-specific disease transmission (ie, disease transmission is more likely to occur in face-to-face orientations rather than facing the same direction in a queue), increased accuracy in the estimates of outdoor waiting times, and empirical verification of the relationship between perceived future workload, as represented by queue length, and hourly staff productivity. Data-based reneging and balking parameter estimates could also improve the bounds on queue growth, and the accuracy of the public's estimated time loss. That patients are recorded only after waiting outside is a major obstacle to assessing reneging behaviour. The use of appointment-based vaccination scheduling in conjunction with walk-ins could also be examined. An appointment plus walk-in approach might reduce variability in patient turnout, reduce wait times, and also generate reneging data.

Although the validity of the infection risk estimates produced by this simulation is subject to uncertainty, the estimates remain useful for

comparing relative infection risk changes under different scenarios, and offer a reasonable approximation for actual infection counts using the best influenza transmission risk parameters available. It was found that no more than 1% of MIC patients are likely to become infected with influenza, even in a worst-case scenario in which 3% of other patients are infectious and without symptoms in addition to very long wait times. The risk of infection at an MIC does not outweigh the benefits of vaccination, even in a worst-case scenario. It is difficult to conceive of a plausible situation that would make the MICs counter-productive to pandemic control efforts. However, the risk of infection is still large enough to deserve attention from infection control experts, and is one of many factors to consider in hiring additional staff to reduce wait times. By ensuring adequate spacing between patients in waiting areas, good ventilation, sanitation of touch surfaces, etc, MIC managers may be able to increase the net positive effect of their facilities on pandemic control efforts by reducing intra-facility infection risk.

References

Aleman DM, Wibisono TG and Schwartz B (2011). A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak. *Interfaces* **41**(3): 301–315.

Amico C (2009). How much will the H1N1 flu cost the U.S.? *PBS News*, 8 October http://www.pbs.org/newshour/updates/health/july-dec09/flu-costs_10-08.html, accessed 5 February 2012.

Andress K (2003). A postevent smallpox mass vaccination clinic exercise. *Disaster Management Response* **1**(2): 54–58.

Araz OM, Lant T, Fowler JW and Jehn M (2011). A simulation model for policy decision analysis: A case of pandemic influenza on a university campus. *Journal of Simulation* **5**(2): 89–100.

Asaro PV, Lewis LM and Boxerman SB (2008). Emergency department overcrowding: Analysis of the factors of renege rate. *Academic Emergency Medicine* **14**(2): 157–162.

Asllani A, Dileepan P and Ettkin L (2007). A methodology for using simulation to optimize emergency mass vaccination parameters. *Journal of Medical Systems* **31**(6): 453–459.

Barnes S, Golden B and Wasil E (2010). A dynamic patient network model of hospital-acquired infections. In: Johansson B, Jain S and Montoya-Torres J (eds). *Proceedings of the 2010 Winter Simulation Conference (WSC 2010)*. Winter Simulation Conference, Baltimore, MD, pp 2249–2260.

Bearden A *et al* (2012). An outbreak of the 2009 influenza A (H1N1) virus in a children's hospital. *Influenza and Other Respiratory Viruses* **6**(5): 374–379.

Blackwell T (2010). Ottawa H1N1 spending was too high, ineffective: critics. *National Post*, 6 April. http://www.nationalpost.com/news/story.html?id=2770570, accessed 5 February 2012.

Brankston G, Gitterman L, Hirji Z, Lemieux C and Gardam M (2007). Transmission of influenza A in human beings. *Lancet Infectious Disease* **7**(4): 257–265.

CBC News (2009). Canada to order 50.4 million H1N1 vaccine doses; $400m contract goes to GlaxoSmithKline factory in Quebec City. http://www.cbc.ca/news/health/story/2009/08/06/swine-flu-vaccine.html, accessed 3 January 2012.

Centers for Disease Control and Prevention (2010). Interim guidance on infection control measures for 2009 H1N1 influenza in healthcare settings, including protection of healthcare personnel. http://www.cdc.gov/h1n1flu/guidelines_infection_control.htm, accessed 8 November 2012.

Centers for Disease Control and Prevention (2011). CDC guidelines for large-scale influenza vaccination clinic planning. http://www.cdc.gov/flu/professionals/vaccination/vax_clinic.htm, accessed 6 November 2012.

Cui F *et al* (2011). Transmission of pandemic influenza A (H1N1) virus in a train in China. *Journal of Epidemiology* **21**(4): 271–277.

Das TK, Savachkin AA and Zhu Y (2008). A large scale simulation model of pandemic influenza outbreaks for development of dynamic mitigation strategies. *IIE Transactions* **40**(9): 893–905.

Durbin A, Corallo A, Wibisono T, Aleman D, Schwartz B and Coyte P (2011). A cost effectiveness analysis of the H1N1 vaccine strategy for Ontario, Canada. *Journal of Infectious Diseases and Immunity* **3**(3): 40–49.

Elveback LR, Fox JP, Ackerman E, Langworthy A, Boyd M and Gatewood L (1976). An influenza simulation model for immunization studies. *American Journal of Epidemiology* **103**(2): 152–165.

Fomundam S and Herrmann JW (2007). *A survey of queuing theory applications in healthcare*. Technical Report TR 2007–24, Institute for Systems Research, University of Maryland. URL http://drum.lib.umd.edu/handle/1903/7222, accessed 28 August 2013.

Foxwell RA, Roberts L, Lokuge K and Kelly PM (2011). Transmission of influenza on international flights, May 2009. *Clinical Infectious Diseases* **17**(7): 1188.

Gentry A (2010). PECS data monitoring system. Internal Document, Niagara Region Public Health.

Gentry A (2011). Protocol for electronic clinic systems version 3.0—data analysis. Internal Document, Niagara Region Public Health.

Gojovic MZ, Sander B, Fisman D, Krahn MD and Bauch CT (2009). Modelling mitigation strategies for pandemic (H1N1) 2009. *Canadian Medical Association Journal* **181**(10): 673–680.

Haber MJ *et al* (2007). Effectiveness of interventions to reduce contact rates during a simulated influenza pandemic. *Emerging Infectious Disease* **13**(4): 581–589.

Hupert N, Mushlin A and Callahan M (2002). Modeling the public health response to bioterrorism: using discrete event simulation to design antibiotic distribution centers. *Medical Decision Making* **22**(5): s17–25.

Influenza Team, European Centre for Disease Prevention and Control (2007). Influenza transmission: re-search needs for informing infection control policies and practice. *Eurosurveillance* **12**: 3189.

Khan K *et al* (2011). *An analysis of canada's vulnerability to emerging infectious disease threats via the global airline transportation network*. Scientific report, St. Michael's Hospital.

Klontz KC, Hynes NA, Gunn RA, Wilder MH, Harmon MW and Kendal AP (1989). An outbreak of influenza A/Tiawan/1/86 (H1N1) infections at a naval

base and its association with airplane travel. *American Journal of Epidemiology* **129**(2): 341–349.

Knibbs D, Morawska L and Bell S (2012). The risk of airborne influenza transmission in passenger cars. *Epidemiology and Infection* **140**(3): 474–475.

Las Vegas Review-Journal (2009). Long lines mark first day of H1N1 vaccine shots. http://www.lvrj.com/news/breaking_news/Long-lines-mark-first-day-of-H1N1-vaccine-shots-64673797.html, accessed 7 November 2012.

Lee B Y *et al* (2010). A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 H1N1 influenza pandemic. *Vaccine* **28**(31): 4875–4879.

Lessler J, Reich NG and Cummings DA the New York City Department of Health and Mental Hygiene Swine Influenza Investigation Team (2009). Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. *The New England Journal of Medicine* **361**(27): 2628–2636.

Los Alamos National Laboratory (2006). EpiSimS Los Angeles case study. http://public.lanl.gov/stroud/LACaseStudy5.pdf, accessed 15 March 2011.

Mei S, Van De Vijver D, Xuan L, Zhu Y and Sloot PMA (2010). Quantitatively evaluating interventions in the influenza A (H1N1) epidemic on China campus grounded on individual-based simulations. In: Sloot PMA, van A Albada GD and Dongarra J (eds). *Proceedings of the International Conference on Computational Science (ICCS 2010)*. International Conference on Computational Science, Amsterdam, pp 1675–1682.

Mitchell R *et al* (2012). Impact of the 2009 influenza A (H1N1) pandemic on Canadian health care workers: A survey on vaccination, illness, absenteeism, and personal protective equipment. *American Journal of Infection Control* **40**(7): 611–616.

Moser MR, Bender TR, Margolis HS, Noble GR, Kendal AP and Ritter DG (1979). An outbreak of influenza aboard a commercial airliner. *American Journal of Epidemiology* **110**(1): 1–6.

Mylius S, Hagenaars T, Lugner A and Wallinga J (2008). Optimal allocation of pandemic influenza vaccine depends on age, risk, and timing. *Vaccine* **26**(30): 3742–3749.

Noakes CJ and Sleigh PA (2009). Mathematical models for assessing the role of airflow on the risk of airborne infection in hospital wards. *Journal of the Royal Society, Interface* **6**(S6): S791–S800.

NY1 News (2009). Long lines wait for H1N1 vaccines at weekend clinics. http://queens.ny1.com/content/top_stories/108933/long-lines-wait-for-h1n1-vaccines-at-weekend-clinics, accessed 7 November 2012.

Phillips FB and Williamson JP (2005). Local health department applies incident management system for successful mass influenza clinics. *Journal of Public Health Management Practice* **11**(4): 269–273.

Public Health Agency of Canada (2010). Lessons learned review: Public Health Agency of Canada and Health Canada Response to the 2009 H1N1 pandemic. http://www.phac-aspc.gc.ca/about_apropos/evaluation/reports-rapports/2010-2011/h1n1/context-contexte-eng.php #figure-2-2, accessed 11 November 2012.

Public Health Agency of Canada (2011). The Canadian pandemic influenza plan for the health sector. http://www.phac-aspc.gc.ca/cpip-pclcpi/index-eng.php, accessed 6 November 2012.

Salgado CD, Farr BM, Hall KK and Hayden FG (2002). Influenza in the acute hospital setting. *The Lancet Infectious Diseases* **2**(3): 145–155.

Sander B *et al* (2010). Is a mass immunization program for pandemic (H1N1) 2009 good value for money? Evidence from the Canadian experience. *Vaccine* **28**(38): 6210–6220.

Savachkin A and Uribe-Sanchez Andres (2012). Dynamic redistribution of mitigation resources during influenza pandemics. *Socio-economic Planning Sciences* **46**(1): 33–45.

Sikora C *et al* (2010). Transmission of pandemic influenza A (H1N1) 2009 within households: Edmonton, Canada. *Journal of Clinical Virology* **49**(2): 90–93.

Statistics Canada (2006). Census 2006. http://www12.statcan.ca/census-recensement/2006/rt-td/ index-eng.cfm, accessed 15 March 2011.

Stroud P, Del Valle S, Sydoriak S, Riese J and Mniszewski S (2007). Spatial dynamics of pandemic influenza in a massive artificial society. *Journal of Artificial Societies and Social Simulation* **10**(4): 9.

Tellier R (2006). A review of aerosol transmission of influenza A virus. *Emerging Infectious Diseases* **12**(11): 1657–1662.

United Kingdom Department of Health (2011). UK Influenza Pandemic Preparedness Strategy 2011. http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_131040.pdf, accessed 6 November 2012.

Viboud C *et al* (2004). Risk factors of influenza transmission in households. *The British Journal of General Practice* **54**(506): 684–689.

Waldie P and Alphonso C (2009). Cost of vaccinating nation hits $1.5 billion and climbing. *Globe and Mail*, 12 November 2009. http://www.theglobeandmail.com/life/health-and-fitness/health/conditions/cost-of-vaccinating-the-nation-hits-15-billion-and-climbing/article1318824/, accessed 5 February 2012.

Washington M, Mason J and Meltzer M (2005). Maxi-Vac: planning mass smallpox vaccination clinics. *Journal of Public Health Management Practice* **11**(6): 542–549.

Washington ML (2009). Evaluating the capability and cost of a mass influenza and pneumococcal vaccination clinic via computer simulation. *Medical Decision Making* **29**: 414–423.

Wong BC *et al* (2010). Possible role of aerosol transmission in a hospital outbreak of influenza. *Clinical Infectious Diseases* **51**: 1176–1183.

World Health Organization (2012). Infection prevention and control in health care. http://www.who.int/csr/bioriskreduction/infection_control/en/index.html, accessed 8 November 2012.

Yang J (2009). First-day of H1N1 shots meant 3 hours of waiting. http://www.healthzone.ca/health/newsfeatures/swineflu/article/716547, accessed 7 November 2012.

Zenios SA (1999). Modeling the transplant waiting list: A queueing model with reneging. *Queuing Systems* **31**(3): 239–251.

10

Modelling Treatment Effects in the HIV/AIDS Epidemic

J. D. Griffiths, Z. F. Lawson and J. E. Williams
Cardiff University, South Glamorgan, UK

Advances in recent treatments for HIV/AIDS patients have shown dramatic outcomes in extending the incubation period and AIDS survival time, while also providing significant improvements in the quality of patients' lives. A compartmental model is proposed to analyse the effects of the various treatment regimens which have been introduced. The results produced are in good agreement with routinely collected data relating to levels of HIV/AIDS incidence and prevalence in the UK homosexual population. Some parameter values within the model are obtained from surveys, census results, etc, but others are derived using a maximum likelihood estimation procedure. Finally, the model is used to project levels of incidence and prevalence over the next few years, and to investigate several possible scenarios.

10.1 Introduction

AIDS is officially diagnosed by the presence of one or more defining conditions. In order to monitor the progress of HIV within the body a marker path is used to describe the incubation period, where markers are a consequence of disease. One of the most commonly used markers for this purpose is CD4 cell count. In general, the lower the CD4 cell count, the worse the condition of the immune system and hence the severity of the illness. Markers have proved to be of great importance in monitoring the effect of therapy; in particular, measurements of CD4 can give an indication as to whether HIV has become resistant to

Reprinted from *Journal of the Operational Research Society*, 57: 1413–1424, 2006, 'Modelling Treatment Effects in the HIV/AIDS Epidemic', by J. D. Griffiths, Z. F. Lawson and J. E. Williams. With kind permission from Operational Research Society Ltd. All rights reserved.

the current treatments available. The year 1987 saw the development of a premier anti-HIV drug, zidovudine (AZT). Treating an individual with just one such antiretroviral drug is known as mono-therapy and can result in significantly greater CD4 cell recovery compared with no therapy. The suggestion that continuous triple or quadruple therapy in acute HIV infection would give an even higher probability of short- to medium-term viral suppression (Smith *et al*, 2004) and the fact that HIV can become resistant to any one drug, has meant that, in 1995, a combination (cocktail) of potent and highly active antiretroviral drugs (HAART), which the patient is required to take on a regular basis and indefinitely, was introduced. Treatment regimens now involve a combination of two or three different drugs, with an increasing number of individuals receiving quadruple therapy (Beck *et al*, 1999). Prior to the introduction of these new combination therapies, nearly 50% of young men who became infected with HIV died within 12 years of initial infection (NIH, 2002). Following the introduction of these advanced treatments, there has been a well-documented reduction in mortality and risk of AIDS-defining illnesses (Van Sighem *et al*, 2003), including a reduction in the hazard ratio for death which fell from an already reduced value of 0.47 in 1997 to 0.16 in 2001 (CASCADE Collaboration, 2003). In fact, progression to an AIDS-defining illness, in those who can tolerate treatment and do not develop drug resistance, has been effectively stopped as long as HIV is diagnosed, and treated properly, early enough in the course of clinical progression (McHenry *et al*, 2002). Part of the decline in AIDS cases and deaths from AIDS might also be attributed to changes in the infected and treated population during the years 1996–2000, and not just the initial impact of cocktail therapies, as the population shifted towards one in a less advanced stage of HIV-1 infection at the start of HAART with a growing fraction of therapy-naïve patients (Van Sighem *et al*, 2003). This argument helps justify the extended transition time of changing parameter values, to allow for a treatment effect, as indicated later in this article.

The aim of this paper is to model the effect that treatments and therapies have had on the HIV/AIDS epidemic, particularly since 1995 when combination antiretroviral drugs were first established. Once a good-fitting, realistic, model is founded, it is implemented to predict short-term estimates of the HIV/AIDS epidemic, assuming there are no immediate changes in the parameter values, extrapolating up until the year 2008. Also, this paper looks at possible future scenarios when parameters are subject to modification, for instance, a change in behaviour due to public awareness or government-sponsored campaigns.

10.1.1 Mathematical model

Generally, infectious disease is modelled using SIR, or SIS, models. These involve dividing the population into sub-populations, those who are: susceptible; infected and recovered (or susceptible again). For HIV/AIDS, since individuals do not recover from the disease, an extension of a SI model is used, with an AIDS stage and a Death from AIDS stage included. These four sub-populations can then be split further to create various other sub-populations, as seen in publications by Hethcote and Van Ark (1980), Griffiths *et al* (1990), Hethcote *et al* (1991), Bailey (1993, 1994a, 1994b, 1997), Griffiths and Williams (1994), Pasqualucci *et al* (1998) and Rossi and Schinaia (1998). For the model implemented here, the sub-populations are split into those individuals who indulge in high-risk activities and those who adopt low-risk behaviour. Roberts and Dangerfield (1990b) divide the subpopulations even further than this, into three risk behaviour groups: high, moderate and low. Also, the infected sub-population can be sub-divided into a number of smaller sections in order to demonstrate the various stages of disease progression that an individual may pass through before being diagnosed with AIDS. The model used in this paper is illustrated in Figure 10.1; an individual in a particular state will occupy that state for a period of time, and movements between compartments are governed by transfer rates. The sub-populations are denoted by

$X(t)$ = the number of Susceptibles at time t,
$Y(t)$ = the number of High Risk Infectives at time t,
$V(t)$ = the number of Low Risk Infectives at time t,
$A(t)$ = the number of High Risk AIDS cases at time t,
$Z(t)$ = the number of Low Risk AIDS cases at time t

Following an individual through the Cardiff model, that individual can stay within the sub-population they are in, transfer to another sub-population with an adjoining arrow, exit the model from causes other than AIDS (eg dying from causes other than AIDS at any stage in the model as well as, if not infected by the virus, exiting the model by no longer participating in high-risk behaviour or homosexual sexual activities), or exit the model due to AIDS, that is, death from AIDS. It is beneficial to maintain as simple a model as possible; the inclusion of too many unimportant parameters will result in greater ambiguity because of the increased levels of estimation involved. This work, like that of Dangerfield and Roberts (1989, 1994, 1996), and Roberts and Dangerfield (1990a), uses a compartmental model to emulate the transmission of HIV in the

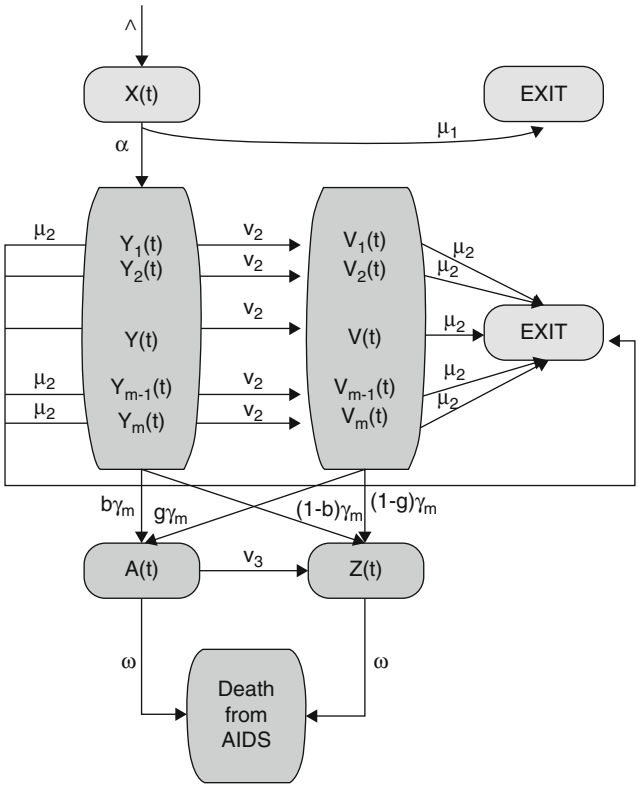


Figure 10.1 Cardiff HIV/AIDS epidemic model

male homosexual community. There are a number of basic assumptions implied in the creation of the Cardiff Model, including the assumption that an individual can only enter the model, and a susceptible can only become infected, if participating in high-risk activities. Also, an individual can move from high-risk to low-risk within the infective sub-populations, the change in behaviour potentially resulting from the impact of educational programmes, enhanced awareness and counselling; however, it is not expected that an individual will return to their prior high-risk behaviour at any point, although it is relatively easy to add this reverse movement to the model. Another simplification imposed is the exclusion of the low-risk susceptible sub-population. This ensures that the model is in its most basic form without affecting expected numbers of AIDS/HIV incidence, prevalence or deaths produced by the model. There is a

parameter related to each possible movement through the Cardiff model. These parameters are shown in Figure 10.1 and are defined as follows:

Λ = the arrival rate into the high-risk susceptible category,
μ_1 = the rate that susceptibles exit the model from causes other than AIDS,
μ_2 = the rate that infectives exit the model from causes other than AIDS,
ν_2 = the rate that high-risk infectives transfer to the low-risk category,
ν_3 = the rate that high-risk AIDS cases transfer to the low-risk category,
b = proportion of high-risk infectives who, on diagnosis with AIDS, continue their high-risk behaviour,
g = proportion of low-risk infectives who, on diagnosis with AIDS, change their behaviour to high-risk,
α = $\beta c(Y(t) + A(t))/(X(t) + Y(t) + A(t))$ = the rate at which homosexual men become infected, where βc is the infectivity rate, and $(Y(t) + A(t))/(X(t) + Y(t) + A(t))$ represents the proportion of high-risk HIV-infected individuals at time t,
γ_i = the transition rate from stage i of HIV infection to stage $i + 1$, $i = 1, \ldots, m$,
ω = the death rate from AIDS.

From Figure 10.1, we may set up the following equations:

$$X(t + \delta t) = X(t) + [\Lambda - \mu_1 X(t) - \alpha X(t)]\delta t \tag{1a}$$

$$\left.\begin{aligned} &Y(t+\delta t) = \sum_{i=1}^{m} Y_i(t+\delta t) \\ &\text{where } Y_1(t+\delta t) = Y_1(t) = [\alpha\, X(t) - (\gamma_1 + \nu_2 + \mu_2)Y_1(t)]\delta t \\ &\text{and} \\ &\qquad Y_i(t+\delta t) = Y_i(t) + [\gamma_{i-1}Y_{i-1}(t) - (\gamma_i + \mu_2 + \nu_2)Y_i(t)]\delta t, \\ &\qquad i = 2, \ldots, m. \end{aligned}\right\} \tag{1b}$$

$$A(t + \delta t) = A(t) + [\gamma_m b Y_m(t) + \gamma_m g V_m(t) - (\omega + \nu_3)\, A(t)]\delta t \tag{1c}$$

$$\left.\begin{aligned} &V(t+\delta t) = \sum_{i=1}^{m} V_i(t+\delta t) \\ &\text{where } V_1(t+\delta t) = V_1(t) = [\nu_2 Y_1(t) - (\gamma_1 + \mu_2)V_1(t)]\delta t \\ &\text{and} \\ &\qquad V_i(t+\delta t) = V_i(t) + [\gamma_{i-1}V_{i-1}(t) - \nu_2 Y_i(t) - (\gamma_i + \mu_2)V_i(t)]\delta t, \\ &\qquad i = 2, \ldots, m. \end{aligned}\right\} \tag{1d}$$

$$Z(t + \delta t) = Z(t) + [\gamma_m(1 - b)Y_m(t) + \gamma_m (1 - g)V_m(t) + \nu_3 A(t) - \omega Z(t)]\delta t \quad (1e)$$

The model incorporates m stages within the infective categories, in accordance with the natural history of HIV infection (Longini *et al*, 1996). We denote the transition rate per unit time from stage i to stage $i + 1$ by γ_i, where $i = 1, \ldots, m$ and stage $m + 1$ relates to the full-blown AIDS condition. This notation is applied due to the assumption that there are only '1-step' forward movements between stages. $Y_i(t)$ and $V_i(t)$ ($i = 1, \ldots, m$) denote the number of HIV infectives at time t in stage i of the high-risk and low risk sub-categories, respectively. We may use equations (1a)–(1e) to estimate such quantities as HIV incidence, HIV prevalence, AIDS incidence and AIDS prevalence as functions of time. If we take $\delta t = 1/52$ (years), then HIV incidence for that year is given by $\alpha \sum_{i=1}^{52} X(i)$, where $X(i)$ ($i = 1, 2, \ldots, 52$) denotes the number of high-risk susceptibles present in week i. Similarly, AIDS incidence is given by

$$\gamma_m \left[\sum_{i=1}^{52} Y_m(i) + \sum_{i=1}^{52} V_m(i) \right]$$

where $Y_m(i)$ and $V_m(i)$ denote the total number of high-risk and low-risk infectives present in the last stage of HIV infection, respectively, in week i. The term prevalence refers to the number of HIV (and AIDS) cases alive at any given time t. Thus, HIV prevalence is defined as $Y(t) + V(t) + A(t) + Z(t)$, whereas AIDS prevalence, alone, is defined as $A(t) + Z(t)$.

One of the many benefits of having individual equations, belonging to each subcategory of the HIV infective populations, is that by solving Equations (1b)–(1e), the HIV prevalence for each stage of HIV progression can be obtained. Sabin *et al* (1994) state that knowledge of the numbers within each stage of HIV infection can be valuable in areas such as health care planning; information on how many HIV infectives are there in the later stages of HIV disease will provide good estimates for future AIDS incidence which can then be used in budgeting for potential health care and treatment costs.

10.1.2 Modelling treatment effects

We now need to estimate the values of the various parameters used in the model. To this end, we use estimation techniques in comparing output values from the model for HIV and AIDS incidence with the

observed AIDS and HIV incidence data obtained from the European Non-Aggregate AIDS Data Set, or ENAADS (EuroHIV, 2003). The observed data utilized for this fit has formerly been adjusted for under-reporting and reporting delays. However, data over the most recent years may still be deemed unreliable; hence, our comparison covers the years 1979–2002 only. Using the ENAADS data set as the source of the observed AIDS incidence data, the first reported case of AIDS in the UK was in 1979. Thus, we take this as the first year of the HIV/AIDS epidemic and so 1978 is employed as the base year (ie $t = 0$). Several researchers, including Bailey (1991, 1992) and Griffiths and Williams (1995), have implemented the method of maximum likelihood to estimate parameters within an epidemic model. One limitation of the technique is the assumption that parameters remain static. For many of the parameters within the model this is not the case, predominantly due to the introduction of treatments and media campaigns, public awareness and attitudes affecting behavioural patterns over time (Griffiths and Williams, 1994). One way of overcoming this difficulty is to consider a number of maximum likelihood estimates (MLE's) based upon different time eras that are linked together, thus allowing new values for parameters to be estimated at certain time intervals based on behavioural changes and the introduction of various treatment regimens. The HIV/AIDS epidemic lends itself to three time eras: 1978–1986 (pretreatment); 1987–1994 (early treatment) and 1995–2002 (combination treatment). The Poisson distribution (as implemented by Bailey (1991, 1992) and Griffiths and Williams (1995)) offers itself as a suitable statistical distribution to model the number of new AIDS cases diagnosed over a fixed time interval, and has the probability distribution function (p.d.f) below (Equation (2a)).

$$f(a_i) = \frac{\alpha_i^{a_i} e^{-\alpha_i}}{a_i!}, \quad i = 1,2,\ldots,T \tag{2a}$$

where a_1, a_2, …, a_T are the observed AIDS cases for T successive time periods, and α_1, α_2, …, α_T are the expected AIDS cases for T successive time periods.

Next, the expression for the log likelihood of this Poisson model (Equation (2b)) is maximized, subject to the expected AIDS cases, α_i.

$$L = \log\left\{\prod_{i=1}^{T} \frac{\alpha_i^{a_i} e^{-\alpha_i}}{a_i!}\right\} = \sum_{i=1}^{T} \{a_i \log \alpha_i - \alpha_i - \log(a_i!)\} \tag{2b}$$

The expected AIDS cases, a_i, are the terms created by the model itself and from the fact that the model is adjusted in accordance with the n parameters to be estimated, θ_j, $j = 1, \ldots, n$, it is such that a_i are dependent upon θ_j. Accordingly, the expression L (Equation (2b)) is differentiated by each of the parameters to be estimated, θ_j, and the results set to zero to give a set of n simultaneous equations. Thus, $\underline{\theta}$ and $\underline{S}$ are created, where $\underline{\theta}$ is the vector of the n parameters to be estimated and $\underline{S}$ the vector of the associated partial derivatives.

$$\underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \quad \underline{S} = \begin{bmatrix} \delta L/\delta\theta_1 \\ \delta L/\delta\theta_2 \\ \vdots \\ \delta L/\delta\theta_n \end{bmatrix} = \underline{0}$$

By letting θ_j be the initial approximation to the MLE $\hat{\theta}_j$, for large T (ie a large number of successive time periods), $\hat{\theta}_j$ is asymptotically unbiased and efficient, that is

$$\hat{\theta}_j \sim N\{\theta_j, \underline{I}(\theta_j)^{-1}\}$$

Also, by definition, the variance–covariance matrix, $\underline{V}$, is equal to the inverse of the information matrix, $\underline{I}$, which is identified in the multivariate case as: $\underline{I} = -E[(\delta^2 L)/(\delta\theta_j \delta\theta_k)] = \underline{V}^{-1}$.

Subsequently, the iterative procedure (Equation (2c)), which is commonly known as the method of scoring for parameters, is obtained using theory on approximate solutions of likelihood equations (Kendall and Stuart, 1973). Equation (2c) can then be implemented to calculate the estimates of the parameters $\underline{\theta}_j$, $j = 1, 2, \ldots, n$.

$$\underline{\theta}_{n+1} = \underline{\theta}_n + \underline{I}^{-1}\underline{S} \tag{2c}$$

Before production of the computational model can begin, the estimation of the number of HIV positive homosexuals, and AIDS case homosexuals, in each stage, at the end of year 1978, that is $Y_i(1978)$ and $A(1978)$, respectively, needs to be determined. The proportions are as shown in Equation (3), based upon work by Bailey (1991), that is, a geometric distribution based on the initial number of AIDS cases in the base year (1978), where a_{1978} relates to the number of HIV-positive homosexuals at the end of year 1978 and $\rho = \beta c/\gamma_i$.

$$\left.\begin{aligned} Y_i &= \frac{a_{1978}\rho}{(1+\rho)^i}, \quad i = 1,\ldots,6 \\ A &= \frac{a_{1978}}{(1+\rho)^6} \end{aligned}\right\} \tag{3}$$

Now the model has been created, the initial conditions need to be set. Based upon a combination of work by Snary (2000), Lowrie (2000) and Griffiths *et al* (2000), or gained directly from population statistics, these are defined to be as follows:

- IP = Length of incubation period in years = 11.4 (Snary, 2000),
- N_{1978} = number of high-risk homosexuals in the population at the end of year 1978 = $X(1978) + Y(1978) + V(1978) = 10\,000$,
- μ_1 = rate at which homosexuals exit the susceptible category from causes other than AIDS = 0.003 (DH1, 2002),
- μ_2 = rate at which homosexuals exit the high and low-risk infective categories from causes other than AIDS = 0.03 (Snary (2000) and Lowrie (2000) to 2.d.p.),
- ν_2 = rate at which high-risk infectives become low-risk = 0.1 (Snary, 2000),
- ν_3 = rate at which high-risk AIDS cases become low-risk = 0.5 (Snary (2000) and Griffiths *et al* (2000)),
- b = proportion of high-risk infectives who stay high-risk once diagnosed with AIDS = 0.5 (Snary, 2000),
- g = proportion of low-risk infectives who become High Risk once diagnosed with AIDS = 0.1 (Snary, 2000),
- ω = death rate = 1 (Snary, 2000),
- m = number of stages in HIV infection = 6 (Snary, 2000).

The parameters to be estimated are

- Λ = the entry rate into the model,
- a_{1978} = the number of HIV positive homosexuals at the end of year 1978,
- βc = the infectivity parameter,
- r = the proportion the incubation period lengthens due to the effect of treatment.

These parameters have been chosen for estimation due to the difficulty in obtaining reliable values for them from other sources. The entry rate, Λ, can take a number of forms and is often deliberated within

models of the HIV/AIDS epidemic; the number of infected individuals at the start of the epidemic, a_{1978}, causes difficulty in estimation as there are no relevant data available; and the infectivity parameter, βc, a sociological parameter, is difficult to ascertain by its very nature. Finally, of the parameters chosen for estimation, the parameter r, the proportional increase in the incubation period (IP), is considered. It can be viewed as a measurement of the effect treatment has had on the epidemic, due to its ability to lengthen the time that HIV-positive individuals live with HIV, thus postponing the onset of AIDS and, consequently, death from AIDS. It is included within the model as a product of the IP, where it is fixed at a value of 1 before the introduction of treatments, and allowed to increase to a value greater than 1 when estimated in the MLE process. Figure 10.2 illustrates the parameters to be estimated and at which point in time this estimation occurs. The first and second columns on the left of Figure 10.2 depict the three time eras into which the model is split and the years they span over, respectively. The third column lists the two parameters, Λ, a_{1978}, which are estimated at the start of the

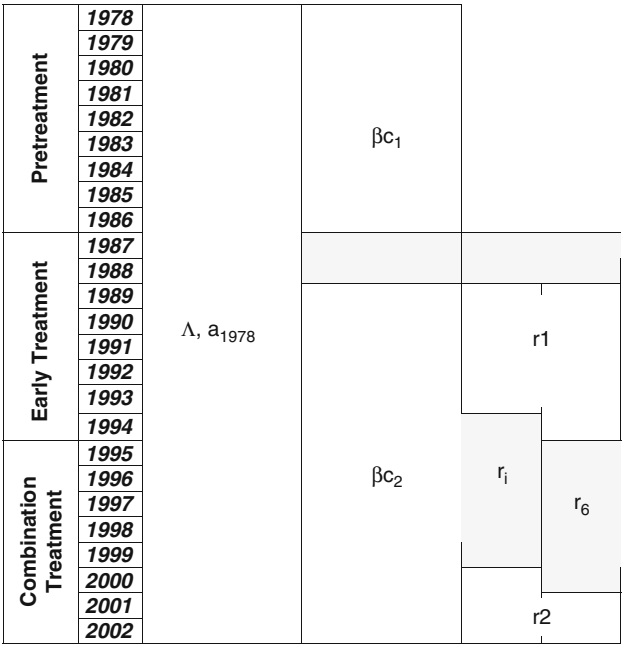


Figure 10.2 Sequence of parameter estimation

epidemic, using observed data, for the entire time period. That is, they are static parameters and, once estimated, do not change value. The next column shows the process by which the infectivity parameter is estimated. Two estimates of βc are made over the entire time period: the first, βc_1, is in relation to the start of the epidemic when people were unaware of the virus and its significance; the second, βc_2, relates to a change in behaviour and attitudes witnessed among homo/bisexual men due to a government campaign promoting safe sex in 1985/1986 (Johnson and Gill, 1989; Evans *et al*, 1993), that is a much reduced infectivity parameter. Of course, the value of this parameter would not change immediately, but would take time to transform from its original value to its new value. Thus, a smoothing technique is introduced into the model such that the change in parameter value, βc, is spread over 2 years (104 weeks) as follows: $|\beta c_1 - \beta c_2|/104$. This increment is then subtracted from the original value, βc_1, each time period of 1 week (δt), until it reaches its new value, βc_2. The treatment effect, r, holds three values over the entire time period due to the fact that introduction of treatments changing the face of the epidemic occurs twice; firstly, in 1987 when antiretrovirals were first established, then later in 1995 when combination therapy was introduced. Initially, before the introduction of treatments, r takes the value of 1 so as not to interfere with the previous model, that is, the IP is its original length. The parameter r is then re-evaluated in the second time period as a result of the first set of treatments being introduced. Again, a smoothing technique is applied for this change in value, to cover 2 years, 1987–1988, this time the increment being equal to $|r - r_1|/104$. For the introduction of cocktail therapies, such as HAART, in 1995, a more complicated transition takes place. The full results of these new and extremely effective treatment regimens have taken longer to occur than the initial impact of new treatments alone might be expected to, due to the changing population of infectives, to one which is more treatment naïve, and the increase in, and application of, experience and knowledge with respect to the use of combination therapies and drug resistance. Thus, for this change in value, from r_1 to r_2, a smoothing time of 6 years has been engaged. This means a smoothing increment of $|r_1 - r_2|/312$ added to r_1 for each time period of 1 week, until it reaches its new value, r_2. Furthermore, previously the assumption was one of a constant staged incubation period, with $\gamma_1 = \gamma_2 = \ldots = \gamma_6$; however, the introduction of new combination treatments has confirmed this notion to be unrealistic as those individuals in the later stages of HIV infection will not have the same reaction to the new therapies as those who are in the earlier stages of

HIV infection. Thus, r has been split into two values over the time of transition; r_i relates to those in stages 1–5 of HIV infection while r_6 relates to those in stage 6 of HIV infection. It is then proposed that the improvement in IP that occurs due to the introduction of combination treatments will be delayed by 1 year for those in stage 6, compared to those in stages 1–5. Thus, the smoothing of value r_1 to r_2 occurs between the years 1994 and 1999 for r_i, with the same values being used for r_6, but 1 year later, such that the transition from r_1 to r_2 occurs between 1995 and 2000 for r_6.

Finally, certain constraints need to be placed upon the parameters in order to maintain realism within the model. These are shown below as follows:

- Λ, $a_{1978} > 0$, non-negativity,
- βc_1, $\beta c_2 \geqslant 0.5$, non-negativity (since both $\beta > 0$ and $c > 0$) plus boundary for realism,
- $\beta c_1 \geqslant \beta c_2$, where βc_1 is the first estimate of βc and βc_2 is the second,
- $1 \leqslant r_1, r_2 \leqslant 5$, restricted to only *increase* the IP due to the introduction of treatments, plus an upper boundary for realism,
- $r_2 \geqslant r_1$, ensuring that as time progresses the value of r only increases the length of the IP further, due to the introduction of new treatments.

10.2 Results

The maximum likelihood estimation (MLE) technique suggests improvements in fit between observed and expected AIDS incidence, subject to the alteration of the parameters being estimated. However, to allow the inclusion of HIV incidence data into the fitting of the model, a measure of the goodness of fit is introduced for both AIDS incidence and HIV incidence. χ^2 values are summed over all time periods to measure the goodness-of-fit for AIDS incidence data (1979–2002) and HIV incidence data (1995–2002). The sum of these two χ^2 values is then minimized subject to changes in the parameters (Λ, a_{1978}, βc and r). The parameter values estimated that created the best fits seen in Figures 10.3 and 10.4 are as follows:

$$\Lambda = 1665$$

$$a_{1978} = 405$$

$$\beta c_1 = 1.28, \beta c_2 = 0.5$$

$$r_1 = 1.23, \; r_2 = 4.89$$

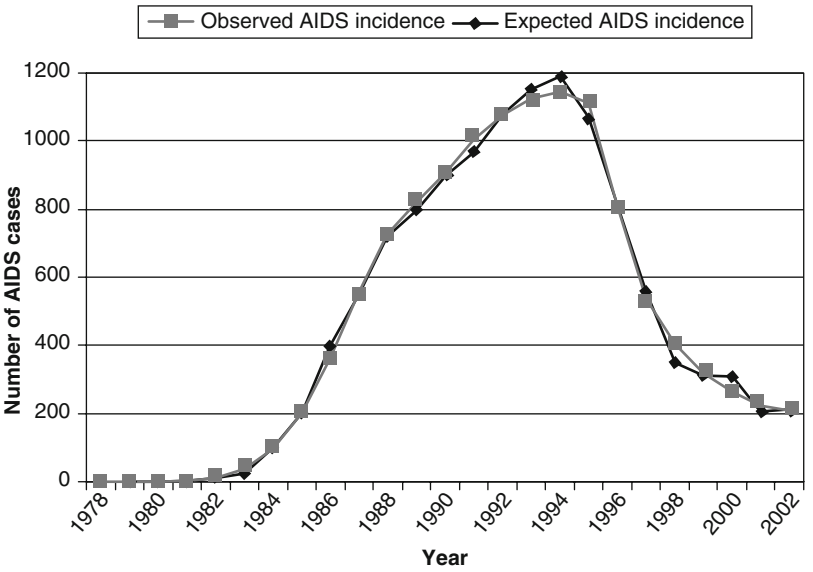


Figure 10.3 Fit between observed and expected AIDS incidence, 1979–2002

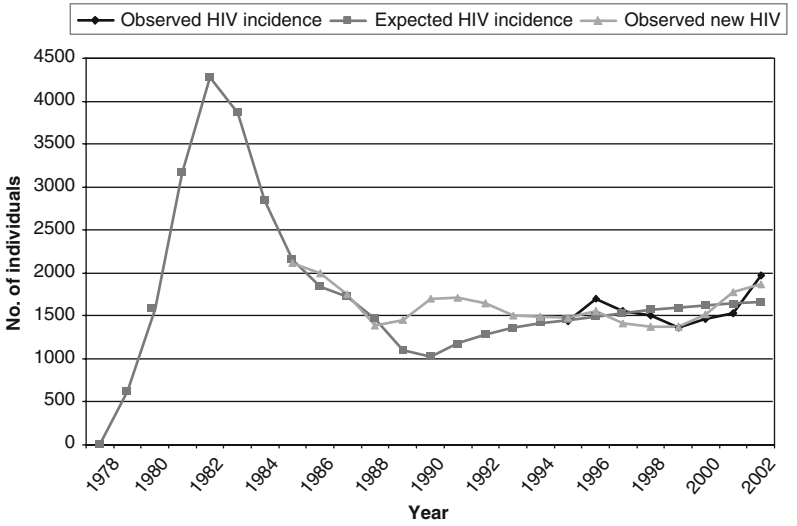


Figure 10.4 Fit between observed and expected HIV incidence, 1995–2002

These parameters result in a total minimized χ^2 value of 177.313. This is made up of a χ^2 value of 32.825 (3 dp), summed over the years 1979–2002, for the fit between observed and expected AIDS incidence and a χ^2 value of 144.488 (3 dp), summed over the years 1995–2002, for the fit between observed and expected HIV incidence. Within the total AIDS incidence fit, separate values for each treatment era can be withdrawn as follows: Pre-treatment era (1979–1986) AIDS χ^2 = 9.476 (3 dp); early treatment era (1987–1994) AIDS χ^2 = 4.843 and combination treatment era (1995–2002) AIDS χ^2 = 18.506. It is noteworthy that due to the inclusion of HIV data into the fitting process, the original possible overall fit for AIDS incidence is compromised somewhat. However, the overall benefits to the model observed when including the HIV data, far outweigh this slight cost in AIDS fit, which is still outstandingly good. As can be seen from the χ^2 values, the majority of the 'bad fit' stems from the HIV incidence data. This is partly because HIV incidence data is only available for the years 1995–2002 from the ENAADS data set (maintaining consistency with the data set from which observed AIDS cases were extracted). Included in Figure 10.4 is the count of new diagnosed cases of HIV, obtained from the Health Protection Agency (HPA, 2005) and by direct contact with the Communicable Disease Surveillance Centre (CDSC), which dates back to 1985. It is important to realize that new HIV diagnoses and HIV incidence are not equivalent due to the inconsistency of when individuals are diagnosed with HIV compared to the point in time when infection actually occurs. However, numbers of newly diagnosed HIV cases can still be a good approximation to HIV incidence figures; hence comparisons between the expected HIV incidence produced by the model and the observed new diagnosed HIV infections attained from HPA, may be considered useful. Prior to 1985, it is probable that a shape, like that seen in the expected HIV incidence curve depicted in Figure 10.4, occurs when compared with suggested trends investigated by Snary (2000), Downs and Houweling (1997), Day Report (1996) and De Angelis *et al* (1998), all of whom, in general, estimate that HIV incidence numbers peak in 1982/1983 at a level of about 3500–4500 new HIV cases per year. Post 1983, owing to a government campaign promoting safe sex in 1985/1986 (Johnson and Gill, 1989; Evans *et al*, 1993), levels of HIV incidence are estimated to have fallen quite rapidly and, in more recent years, have maintained their reasonably low, and stable, levels of about 1600 new cases of HIV each year.

Other results can also be extracted from the model, including the expected number of deaths from AIDS and the prevalence of HIV/AIDS. These can also be compared with observed data counts in order to

establish whether the model is fitting the HIV/AIDS epidemic realistically. Figure 10.5 shows the fit between expected and observed deaths from AIDS. Within the initial conditions of the model, the survival time, $1/\omega$, was defined to take a value of 1, based on work by Snary (2000) and Lowrie (2000). This value has not been altered or estimated within the model and, consequently, has kept its value as 1 throughout the fitting process. This may be deemed as unrealistic due to the introduction of treatments in 1987 and 1995, which have been associated with a much reduced death rate. Observed death rates from AIDS can be obtained from the CDSC and, consequently, a new variable value for ω can be investigated throughout the epidemic using a goodness-of-fit measurement for the observed and expected deaths data and minimizing the total χ^2 subject to changes in the value of the death rate, ω. This process of minimizing the χ^2 has been performed and the results exemplified in Figure 10.5. The total χ^2 value for the fit of deaths from AIDS data reduced from 817.106 (when ω was fixed at a value of 1 over the entire time period, 1979–2002) to nearly a sixth of the amount,

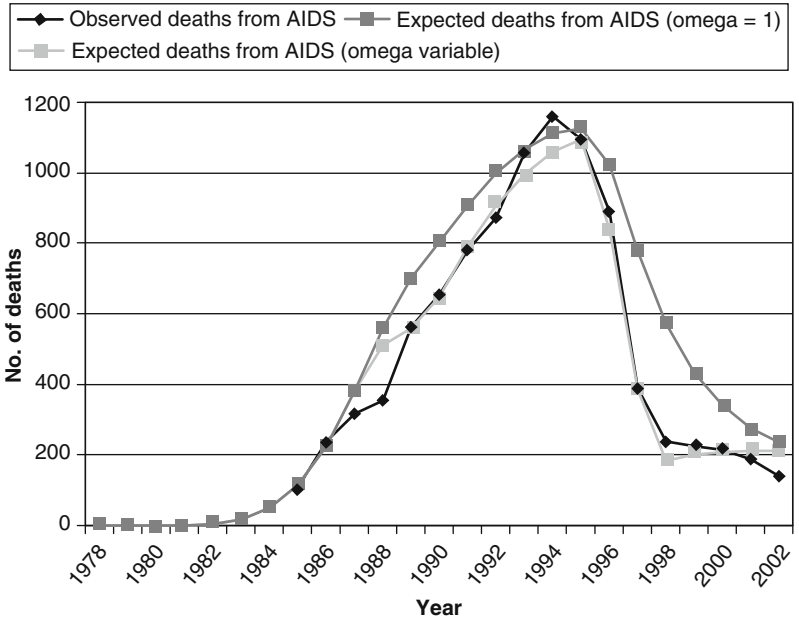


Figure 10.5 Comparison between observed and expected deaths from AIDS, 1985–2002

137.872 (3 dp). The improvement in fit was created by ω initializing at a fixed value of 1 for the pretreatment era (1979–1986), and reducing to a value of 0.6 in the early treatment era (1987–1994). This change in value of ω is consistent with work produced by Dangerfield *et al* (2001), who suggest that the mean survival time (where the mean survival time = 1/the mean death rate) changes from 0.9 year to 1.8 year as a result of AZT being introduced in 1988. The death rate, within the Cardiff model, then drops even further, to a value of 0.09 in the combination treatment era (1995–2002). As previously, smoothing techniques are employed; for each change in value of ω, a smoothing period of 2 years was utilized, (ie smoothing periods over the years 1987–1988 for the first change in value and 1996–1997 for the second).

Figure 10.6 illustrates the fit for HIV/AIDS prevalence, as produced by the model with variable values of ω. Data on observed HIV/AIDS prevalence only dates back to 1996, stemming from a variety of sources: The Public Health Laboratory Service (PHLS, 2002) provided diagnosed figures for years 1996–1998 and 2000; CDSC (2000) and the Department of Health (DH, 2000) provided the 1999 prevalence figure; DH (2002) provided amounts for 2001 and 2002. In addition to these

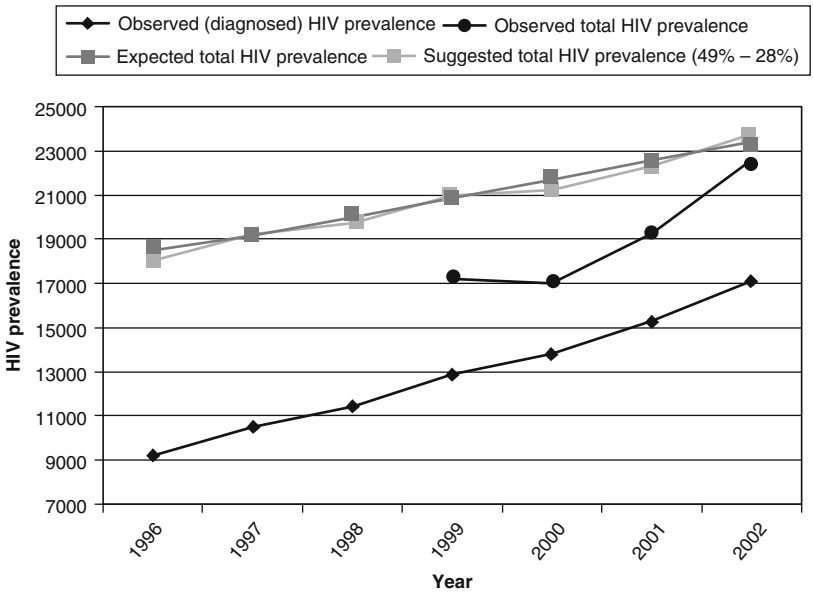


Figure 10.6 HIV prevalence investigation and comparison, 1996–2002

diagnosed prevalence figures, it is expected that a number of HIV positive individuals are unaware of their status; for the years 1999–2002, total prevalence estimates can be obtained from DH (2000) and CDSC (2000) for 1999; PHLS (2002) for 2000; and DH (2002) and CDSC (2002) for 2001 and 2002. These total prevalence estimates (diagnosed and undiagnosed), as obtained from the sources listed above for the years 1999–2002, assume the following percentages of homosexuals are unaware of their serostatus: 25, 19, 23 and 24%, respectively. Figure 10.6 also shows the expected HIV/AIDS prevalence as produced by the model. This is the sum of the numbers in the four sub-populations, $Y(t)$, $V(t)$, $A(t)$ and $Z(t)$ for each year 1996–2002. Comparison between observed and expected prevalence is difficult due to the few observed statistics available and their inconsistency in origin. However, as an example of a possible good fit, the percentage of homo/bisexual men unaware of their infection status has been fixed at a decreasing linear amount each year, starting at 49% unaware in 1996 and reducing to 28% unaware by 2002 (falling by 3.5% each year). The reduction in the proportions of individuals unaware of their serostatus each year can be argued partly due to the improvements in surveillance and testing over time, but also as a result of the introduction of treatments; the fear of testing and diagnosis of HIV may have diminished due to an increase in confidence of the new treatments available. However, a proportion of the population as large as 49% being unaware of their infection is difficult to validate.

10.2.1 Extrapolations and discussion

Now that a good-fitting, realistic, model has been formulated, it can be used to project the short-term future of the HIV/AIDS epidemic, with the assumption that all parameter values and conditions remain constant. Predicting HIV, AIDS and deaths from AIDS up until the year 2008 (with ω variable), as seen in Figure 10.7, shows that levels of incidences are maintaining the stability witnessed in the last few years of the modelled epidemic. This means that, assuming no changes, approximately 1680 people will become infected with HIV each year and AIDS incidence and deaths from AIDS will sustain 200 new cases each year from now until the final year extrapolated, 2008. This may initially seem like a positive trend creating a hopeful insight of the epidemic to come. However, when investigated deeper there are further predictions that may cause concern. Figure 10.8 shows the prevalence of individuals within the sub-populations, $X(t)$, $Y(t)$, $V(t)$, $A(t)$ and $Z(t)$, extrapolated to the year 2008. While the number of individuals susceptible to HIV/AIDS ($X(t)$) is

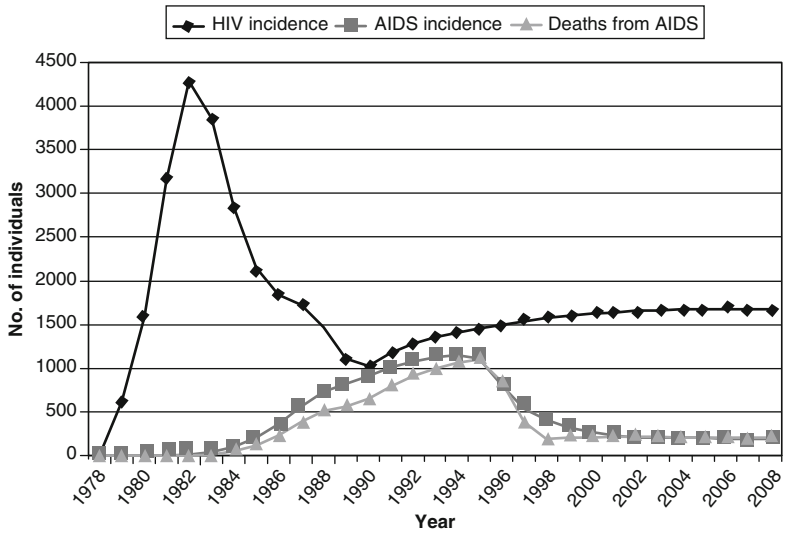


Figure 10.7 Extrapolated HIV, AIDS and deaths from AIDS up to 2008

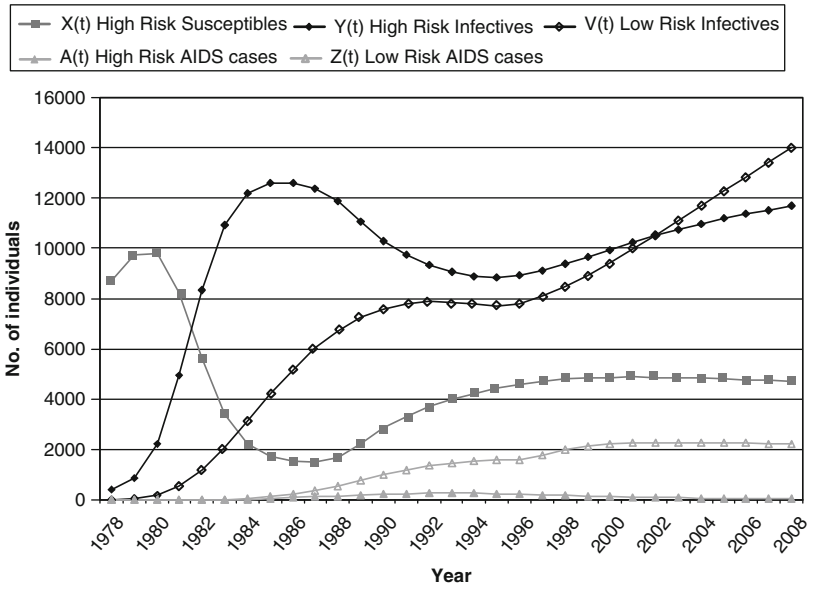


Figure 10.8 Extrapolated prevalence in sub-populations up to 2008

relatively constant at just under 5000 susceptibles present each year, for the later years of the epidemic, and both high-risk and low-risk AIDS prevalences ($A(t)$ and $Z(t)$, respectively) are maintaining low counts over the years leading up to 2008, the number of people living with HIV is increasing exponentially year on year. Coincident with the introduction of combination therapies in 1995, prevalence, in both high-risk and low-risk HIV infective categories ($Y(t)$ and $V(t)$, respectively), is rising to levels that may be deemed unsustainable in the near future, increasing to amounts never reached before. This is not surprising since treatments such as HAART have reportedly postponed the onset of AIDS, and/or death, to such an extent that it is now believed that virions in the blood can be virtually cleared, given that treatments are initiated early enough with the correct knowledge and expertise applied. Also, as seen in Figure 10.7, incidence levels are constant, so more and more people are gathering in the infective stage of the epidemic. It can also be noted from Figure 10.8 that the prevalence of low-risk Infectives actually overtakes that of the high-risk infectives in 2002, due to the constant transition of individuals from high-risk to low-risk behaviour. As a consequence of the greater prevalence numbers of HIV, future estimates of HIV incidence might increase since there are more people living with HIV in the population to infect other people. Furthermore, treatments are improving the health status of people living with HIV and, consequently, individuals may be more sexually active than before, increasing the risks of further transmission of the virus into the susceptible population. These behaviours may result in an increase in HIV incidence in the future, which would then return an even greater rise in HIV prevalence in years to come. Concerns exist as to whether increased levels of HIV prevalence are sustainable in the long term. Many researchers believe that methods of intervention, such as prevention and awareness campaigns, need to be stepped up to meet the levels of success of new treatments in order to ensure the continuation of the beneficial treatment results seen so far. The cost of treatment for each and every individual living with HIV in the UK may become unaffordable if high prevalence levels persist, and consequently, treatment uptake may diminish if prevalence levels continue to rise, resulting in an increase in AIDS incidence and deaths from AIDS to unprecedented levels.

Figure 10.9 illustrates one possible scenario if prevention and awareness campaigns were introduced successfully in the UK in 2005. To model this, the infectivity parameter has been reduced to a value of zero ($\beta c = 0$) from 2007 onwards, with 2 year smoothing from 2005–2006. This equivalently means that no more infections of HIV occur after 2006

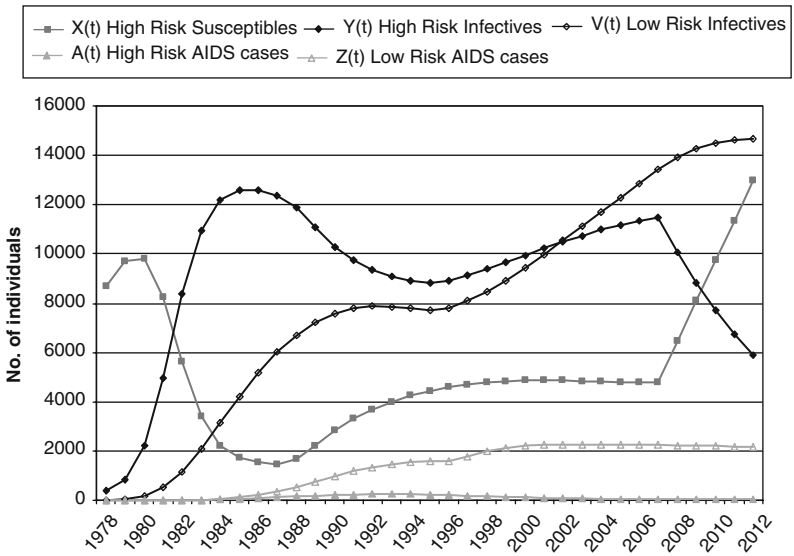


Figure 10.9 Extrapolated prevalence in sub-populations, given a change in behaviour, up to 2012

due to hypothetical prevention tactics being introduced in 2005, that is, the entire homo/bisexual population in the UK no longer partakes in high-risk activities from 2007 onwards as a result of government awareness campaigns and media publicity targeting HIV/AIDS, commencing in 2005. The extrapolations in Figure 10.9 for the sub-populations, and the relevant prevalence estimates, are extended to the year 2012 to demonstrate the longer-term effect on the low-risk HIV infectives ($V(t)$) when introducing effective preventative measures. The high-risk infective sub-population immediately drops as there are no new infectives entering this sub-population from 2007 onwards. At the same time as this movement in $Y(t)$ occurs, the susceptible population, $X(t)$, increases in size rapidly since there is no longer an exit from this sub-population into the infective category. The low-risk Infective category, $V(t)$, takes longer to react since it is not directly linked to the number of new infections. As time progresses, and the number of prevalent high-risk infectives drops, the prevalence of low-risk infectives begins to slowly fall, this is starting to happen by 2012 as can be seen in Figure 10.9. If extrapolated further, it can be seen that $V(t)$ drops steadily year on year. Thus, according to the extrapolations created by this model, assuming that highly effective prevention methods are enforced in the year 2005

to such an extent that new infections are brought to a standstill, and all other conditions remain the same as for previous years, the HIV/AIDS epidemic can be restrained to controllable and treatable levels within a few years, with the best case scenario being the defeat of the epidemic altogether.

References

Bailey NTJ (1991). The use of operational modeling of HIV/AIDS in a systems approach to public health decision making. *Math Biosci* **107**: 413–430.

Bailey NTJ (1992). Parameter estimation in the operational modelling of HIV/AIDS. In: Mardia KV (ed). *Art of Statistical Science*, Chapter 19. John Wiley & Sons Ltd, New York, pp 277–298.

Bailey NTJ (1993). An improved hybrid HIV/AIDS model geared to specific public health data and decision making. *Math Biosci* **117**: 221–237.

Bailey NTJ (1994a). HIV/AIDS: core group dynamics and public health action. In: Kaplan EH and Brandeau ML (eds). *Modelling the AIDS Epidemic: Planning, Policy and Prediction*. Raven Press Ltd, New York, pp 29–52.

Bailey NTJ (1994b). Estimating HIV incidence and AIDS projections: prediction and validation in the public health modelling of HIV/AIDS. *Stat Med* **13**: 1933–1943.

Bailey NTJ (1997). A revised assessment of the HIV/AIDS incubation period, assuming a very short early period of high infectivity and using only San Francisco public health data on prevalence and incidence. *Stat Med* **16**: 2447–2458.

Beck EJ *et al* for the National Prospective Monitoring System Steering Group (1999). Decreased morbidity and use of hospital services in English HIV-infected individuals with increased uptake of anti-retroviral therapy 1996–1997. *AIDS* **13**: 2157–2164.

CASCADE Collaboration (2003). Determinants of survival following HIV-1 seroconversion after the introduction of HAART. *Lancet* **362**: 1267–1274.

CDSC (2000). *HIV and AIDS in the UK. An Epidemiological Review: 2000*. PHLS, London.

CDSC (2002). *HIV and AIDS in the UK in 2001. An Update: November 2002*. PHLS, London.

Dangerfield BC and Roberts CA (1989). A role for system dynamics in modelling the spread of AIDS. *Trans Inst Msmt and Crtl* **11**: 187–195.

Dangerfield B and Roberts C (1994). Fitting a model of the spread of AIDS to data from five European countries. In: Dangerfield B and Roberts C (eds). *OR Work in HIV/AIDS*, 2nd edn. ORS Publication Unit, UK, pp 7–14.

Dangerfield B and Roberts C (1996). Relating a transmission model of AIDS spread to data: some international comparisons. In: Isham V and Medley G (eds). *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Publications of the Newton Institute, Cambridge University Press, Cambridge, UK, pp 473–476.

Dangerfield BC, Fang Y and Roberts CA (2001). Model-based scenarios for the epidemiology of HIV/AIDS: the consequences of highly active antiretroviral therapy. *Systems Dynamics Rev* **17**(2): 119–150.

Day report (1996). The incidence and prevalence of AIDS and prevalence of other severe HIV disease in England and Wales for 1995 to 1999: projections using

data to the end of 1994. Report of a Working Group. Communicable Disease Report 6 (Review Number 1, 5 January).

De Angelis D, Gilks WR and Day NE (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic. *Appl Stat* **47**: 1–33.

DH (2000). Prevalence of HIV and hepatitis infections in the United Kingdom 2000. Annual report of the Unlinked Anonymous Prevalence Monitoring Programme. Department of Health, December 2001.

DH (2002). Prevalence of HIV and hepatitis infections in the United Kingdom 2001. Annual report of the Unlinked Anonymous Prevalence Monitoring Programme. Department of Health; December 2002.

DH1 (2002). Table 6 Deaths and death rates per 1,000 population: age-group and sex, 2002. Age, sex and marital status *Series*. *DH1* no. 35.

Downs A and Houweling H (1997). *Proceedings from a joint meeting of the Working Group Modelling Support and Working Group Epidemiological Impact*. EC Concerted Action Multinational Scenario Analysis Concerning Epidemiological, Social and Economic Impacts of HIV/AIDS on Society held at the National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands (3–4 July).

EuroHIV (2003). *HIV/AIDS Surveillance in Europe*. End-year Report 2003 No. 70. European Centre for the Epidemiological Monitoring of AIDS. WHO and UNAIDS Collaborating Centre on AIDS.

Evans BG *et al* (1993). Sexually transmitted diseases and HIV-1 infection among homosexual men in England and Wales. *Br Med J* **306**: 426–428.

Griffiths JD, Williams JE and Wheeler K (1990). Modelling the course of the AIDS epidemic. In: Dangerfield B and Roberts C (eds). *OR Work in HIV/AIDS*, 1st edn. ORS Publication Unit, UK.

Griffiths J and Williams J (1994). Charting the AIDS Epidemic in the UK. In: Dangerfield B and Roberts C (eds). *OR Work in HIV/AIDS*, 2nd edn. ORS Publication Unit, UK, pp 3–6.

Griffiths JD and Williams JE (1995). *Contribution to meeting of Modelling Working Group*. EC Concerted Action Multinational Scenario Analysis, Paris (8–9 June).

Griffiths J, Lowrie D and Williams J (2000). An age-structured model for the AIDS epidemic. *Eur J Opl Res* **124**: 1–14.

Hethcote HW and Van Ark JW (1980). Modelling HIV transmission and AIDS in the United States. Lecture Notes in Biomathematics, Springer-Verlag, Berlin, Heidelberg.

Hethcote HW, Van Ark JW and Karon JM (1991). A simulation model of AIDS in San Francisco: I Model formulation and parameter estimation. II Simulations, therapy and sensitivity analysis. *Math Biosci* **106**: 203–247.

HPA (2005). AIDS/HIV quarterly surveillance tables. Cumulative UK data to end December 2004, No: 65: 04/4, January 2005.

Johnson AM and Gill ON (1989). Evidence for recent changes in sexual behaviour in homosexual men in England and Wales. *Philosophical Transac Royal Soc London Ser B—Biol Sci* **325**: 151–161.

Kendall MG and Stuart A (1973). *The Advanced Theory of Statistics*, Vol. 2, 3rd edn (Inference and Relationship). Griffin: London.

Longini Jr IM, Clark WS, Satten GA, Byers RH and Karon JM (1996). Staged Markov Models Based on CD4+ T-lymphocytes for the natural history of HIV infection. In: Isham V and Medley G (eds). *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Publications of the Newton Institute, Cambridge University Press, Cambridge, UK, pp 439–459.

Lowrie DL (2000). *An age dependant model for the AIDS epidemic*. PhD thesis, Cardiff University, UK.

McHenry A *et al*. (2002). Numbers of adults with diagnosed HIV infection 1996–2005—adjusted totals and extrapolations for England, Wales and Northern Ireland. *Commun Dis Public Health 2002* **5**(2): 97–100.

NIH (2002). Treatment of HIV infection. Office of Communications and Public Liaison, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fact Sheet, January.

Pasqualucci C, Rava L, Rossi C and Schinaia G (1998). Estimating the size of the HIV/AIDS epidemic: complementary use of the empirical bayesian back-calculation and the mover-stayer model for gathering the largest amount of information. *Simulation* **71**: 213–227.

PHLS (2002). Sexual health in Britain: recent changes in high-risk sexual behaviours and the epidemiology of sexually transmitted infections including HIV. HIV/STI Division, PHLS Communicable Disease Surveillance Centre, July 2002.

Roberts CA and Dangerfield BC (1990a). Modelling the epidemiological consequences of HIV infection and AIDS: a contribution from operational research. *J Opl Res Soc* **41**: 273–289.

Roberts C and Dangerfield B (1990b). A system dynamics modelling framework for understanding the epidemiology of HIV/AIDS. In: Dangerfield B and Roberts C (eds). *OR Work in HIV/AIDS*, 1st edn. ORS Publication Unit, UK.

Rossi C and Schinaia G (1998). The mover-stayer model for the HIV/AIDS epidemic in action. *Interfaces* **28**: 127–143.

Sabin CA, Lee CA and Phillips AN (1994). The use of back-calculation to estimate the prevalence of severe immunodeficiency induced by the human immunodeficiency virus in England and Wales. *J Royal Stat Soc Ser A—Stat Soc* **157**: 41–56.

Smith DE, Walker BD, Cooper DA, Rosenburg ES and Kaldor JM (2004). Is antiretroviral treatment of primary HIV infection clinically justified on the basis of current evidence? *AIDS* **18**: 709–718.

Snary EL (2000). *Modelling treatment effects in the HIV/AIDS epidemic*. PhD thesis, University of Wales, Cardiff, UK.

Van Sighem AI *et al* on behalf of the ATHENA cohort study group (2003). Mortality and Progression to AIDS after starting highly active antiretroviral therapy. *AIDS* **17**: 2227–2236.

Part III
OR for Inventory Management in Emergency Services

11

Impact of the Influenza Season on a Hospital from a Pharmaceutical Inventory Management Perspective

A. R. Vila-Parrish, J. S. Ivy and B. He
North Carolina State University, Raleigh, U.S.A.

The outbreak of an infectious disease may put significant pressure on a healthcare system, especially when there is a surge of patients. In this paper, we develop two simulation models: (1) disease outbreak model and (2) a medication inventory model. These two models are used to identify inventory policies for managing medication during disease outbreaks. Specifically, we use historical influenza data as an input to the inventory simulation model, which incorporates the impact of disease spread, patients' health conditions, and medication shelf life. We formulate a dynamic program and use a reduced version of this model to provide inputs to our inventory simulation model. We compare three different simulation-based policies and perform sensitivity analysis on several parameters such as the gross attack rate and inventory holding cost parameters. Our results provide insight regarding the management of perishable medication inventory at hospitals during an outbreak of an infectious disease.

11.1 Introduction

The U.S. spent 2.5 trillion dollars on health care in 2010, which was 17.3% of the 2009 GDP (Fritze, 2010). The total economic cost due to influenza is estimated to be $87.1 billion each year. Inpatient costs

Reprinted from *Health Systems*, 4: 12–28, 'Impact of the Influenza Season on a Hospital from a Pharmaceutical Inventory Management Perspective', by A. R. Vila-Parrish, J. S. Ivy and B. He. With kind permission from Operational Research Society Ltd. All rights reserved.

make up $78.2 billion of this expenditure, which is 90% of the total influenza-based cost (Molinari *et al*, 2007). The seasonality of influenza can significantly impact the inventory policy of a pharmacy. During these outbreaks, the medication demand at a hospital pharmacy is much greater than usual, leaving hospitals to manage the surge in demand while reducing costs. Furthermore, there have been few studies conducted that explore pharmacy inventory system in hospitals, particularly during a medical surge event such as an outbreak of flu. Optimal inventory policies can significantly affect the system cost, service level, and health care outcomes. This paper develops a model for evaluating hospital pharmacy inventory policies that quantifies the impact of surges in disease, the health conditions of inpatients, and the perishability of the medication.

In our model, patients are generated by the Susceptible/Infected/Recovered (SIR) model (refer to Hethcote (2000) for a recent summary), and thus, arrive according to a time-varying probability function. Patients are categorized by age and risk group, and the corresponding medication need is defined by different probability functions. We focus on the subset of the most severely ill patients who are hospitalized as a result of acquiring the flu. During hospitalization, a patient's health status may change at the end of each day according to an age-group-specific Markov transition matrix. In addition, patients are allocated medication daily. The pharmacy is responsible for ordering the appropriate quantity of raw material, converting raw material into dispensable form for use the next day, and removing expired medication. The decisions made by the pharmacy significantly affect the system cost.

This paper is organized as follows: the next section is a brief review of the related literature. The subsequent section presents a detailed description of the SIR model. The section after that discusses the hospitalization model, then describes the calculation of the parameters of the model, and finally proposes four inventory policies. The penultimate section applies a simulation method to compare the performance of our proposed policies. Details of the simulation model built in Arena are provided in the Appendix.

11.2 Literature review

Perishable inventory models have been developed for many applications, including food, electronic products, medical supply, etc. (refer to Nahmias, 1982; Raafat, 1991; Goyal & Giri, 2001; for a review of the research on this topic dating back to the 1960s).

Of the literature in perishable inventory, many models assumed uniform demand in order to simplify their discussion (Heng *et al*, 1991; Raafat *et al*, 1991; Goh *et al*, 1993); however, since this is usually not what happens in the real world, some papers have relaxed this assumption and modeled time-varying demand. Most of these models assumed that the demand was either a linear (Xu & Wang, 1992; Chung & Ting, 1993) or an exponential function of time (Hollier & Mak, 1983; Goyal & Giri, 2001). Dye *et al* (2006) developed a model with demand as a linear or exponential function of time, and proposed a search method to find the optimal inventory policy. Xu & Wang (1990) described a situation where the inventory was deteriorating, demand was time-varying and deterministic, which they solved using a Dynamic Programming (DP).

Various papers (Nandakumar & Morton, 1993; Perry, 1997; Liu & Lian, 1999) developed inventory models for products with fixed shelf lives. Nandakumar & Morton (1993) discussed a problem where the inventory item has a known fixed shelf life and demand is independent and identically distributed. They bounded the optimal order quantity using a near-myopic heuristic approach, which gave results very close to the optimal solutions. Haijema *et al* (2007) introduced a blood bank inventory problem with two kinds of blood, young and old, in inventory. Our paper differs from the above mainly in that there are two items with different shelf lives in inventory where the model is multi-echelon, that is, one item is used to produce the other.

The research on the two-echelon inventory system also sheds light on our paper. Kanchanasuntorn & Techanitisawad (2006) studied this kind of system with perishable goods, and a fixed interval between review periods. The two-echelon system has the same feature as our inventory model because there is an underlying production relationship between the two levels; however, in our model, demand can be satisfied during the review period by both inventory levels.

An inventory system with one-way substitution is investigated by Dutta & Chakraborty (2010). One-way substitution allows one product to be used as a substitute for another product when the other product is out of stock, but not vice versa. Our model resembles this one-way substitution in the following manner: when the finished goods are out of stock, raw material, if available, can be converted into the finished goods for use instantaneously with an associated premium. In their paper, Dutta and Chakraborty solved an inventory problem with two items and one-way substitution. A numerical search approach was developed to find the optimal ordering policy. However, they assumed

that there was no cost for substitution, and both items could be ordered, in contrast to our model in which only the raw material may be ordered.

Vila-Parrish *et al* (2008) developed a model with three types of patients. Each type had its own arrival rate and medication demand probability density function. They proposed an adaptive policy that used information regarding the in-system patient mix to make decisions. In contrast, in our model demand is determined by the type and number of patients and their specific needs, all of which are randomly assigned to each patient. In addition, the distribution of the number of incoming patients is time varying and the stochastic nature of the distribution is defined by incorporating the epidemiological features of the infectious disease into the model. Vila-Parrish *et al* (2012) studied a similar problem but only allowed for a single order for raw material in the entire planning horizon.

We formulate a two-part simulation model to represent the impact of the flu on inpatient pharmacy inventory demand. We also develop two approaches for solving this problem, simulation and dynamic programming. A two-part simulation model is developed to simulate the transmission of the disease and patients' medication needs during hospitalization. We present the two parts of the simulation model in sections 'The SIR infectious disease model' and 'The hospitalization model', respectively.

11.3 The SIR infectious disease model

The simple SIR model is used to simulate the progression of the flu (Hethcote, 2000). The SIR model uses a set of differential equations to calculate the fraction of susceptible ($s(t)$), infective ($i(t)$), and recovered ($r(t)$) from a certain population.

$$\begin{aligned}\frac{\mathrm{ds}(t)}{\mathrm{dt}} &= -\beta i(t)s(t)\\ \frac{\mathrm{di}(t)}{\mathrm{dt}} &= \beta i(t)s(t) - \gamma i(t).\\ r(t) &= 1 - s(t) - i(t)\end{aligned}$$

The SIR model assumes that after entering the infective class, the probability that the patient is still infective t time units later is $e^{-\gamma t}$. Therefore, the infectious period, which is the average length of time during which a patient can transmit the flu to other susceptible people, is equal to $1/\gamma$. The average number of new infections caused by an infected individual during the infectious period, β, is $\beta = \sigma\gamma$, where σ denotes the contact number that is defined as the average total effective contacts of an

infective individual during his/her infectious period. We use a discrete-time simulation model to approximate the solution of the above set of differential equations because this allows us to link the infectious disease model with the hospitalization model that is discussed in the section 'The hospitalization model'. Given the initial values of $s(0)$ and $i(0)$, the simulation model calculates $i(t)$, $s(t)$, and $r(t)$ every hour (simulation time) to seek a balance between accuracy and efficiency. A shorter time length can slow down the simulation while not improving the accuracy very much. A detailed description of the simulation model in Arena can be found in the Appendix. Every λ h, the simulation model records the number of susceptible individuals in the population by multiplying $s(t)$ with the total population (*Pop*), and compares it with the number of susceptible individuals λ h ago, thus giving the expected number of new cases (*NewCases*) in the last λ h. In our model, we assume $\lambda = 12$. Table 11.1 summarizes of the definitions discussed so far.

For the SIR model, σ and $s(0)$ are estimated from the H3N2 influenza epidemic that broke out in Hong Kong in 1968 (Hethcote, 2000). The average infectious period for a typical flu is assumed to be 7 days (Seasonal Influenza, 2010). So $\gamma = 1/7$ day^{-1} = 5.9524×10^{-3} h^{-1}. We assume that $i(0) = 0.002$. The total number of registered hospitals in the U.S. is 5,810 according to the 2002 data (Dudden *et al,* 2006). The 2002 U.S. population was estimated to be 288,368,698 (Table NA-EST2002-01 2002), resulting in the average population that a single U.S. hospital serves is 49,633. We will use 50,000 in general. While this estimate can vary by year and source, we have found that the 50,000 patient population per hospital does not vary greatly. Table 11.2 summarizes the above input parameters.

Table 11.1 Expressions for the total number of susceptible at times t and $t+\lambda$, and number of new cases during time length of λ

Expression	**Definition**
$s(t)Pop$	The total number of susceptible at time t
$s(t+\lambda)Pop$	The total number of susceptible at time $t+\lambda$
$NewCases = [s(t)-s(t+\lambda)]Pop$	The total number of new cases generated during time t and $t+\lambda$

Table 11.2 Input parameters for the SIR model

Variable	σ	γ	s(0)	i(0)	Pop
Value	1.44	5.9524e–3 h^{-1}	0.911	0.002	50,000

The SIR model describes the transmission of the flu among a segment of the population and calculates the expected number of new cases occurring over a specified period of time (e.g., 12 h). The expected number of new cases calculated by this model serves as the connection between the SIR model and the hospitalization model, which is described in the next section.

11.4 The hospitalization model

The hospitalization model has two parts: (i) the patient condition model and (ii) the pharmacy model. The subsection 'Patient condition model' discusses part (i) including the model parameterization and performance. The subsection 'The pharmacy model' describes part (ii), which models the decision-making process at the pharmacy. Because the patient's condition may change every day, affecting medication demand, the pharmacy must make decisions daily to ensure that enough medication has been prepared while keeping the inventory cost low. In this subsection, we develop and compare four inventory policies to reduce the inventory cost while satisfying dynamic patient demand.

11.4.1 Patient condition model

This model simulates a patient's condition while being hospitalized, for example, how a patient's health states change and how long he/she stays in the hospital. We assume that the interval between new cases that are generated in the hospitalization model will have an exponential distribution with a mean interarrival time, which we assume to be 12 h divided by the total number of new cases generated during the last 12 h by the SIR model. The mean is updated every 12 (simulation) hours. Not all of these new cases will be hospitalized. The hospitalization rate depends on age and the admitting health conditions. We divide the population into three age-groups, 0~19 years old (y/o), 20~64 y/o, and 65+ y/o according to the study by Meltzer *et al* (1999). There are two admitting conditions when a patient is hospitalized, z_1 (non-high risk, i.e., a patient without serious complications like asthma, heart disease, etc.) and z_2 (high-risk, i.e., a patient with serious complications). During their stay in the hospital, patients will move among four states (z_1, z_2, H = Healthy, D = Dead) according to their age-groups and conditions. Some of the serious complications such as asthma and heart disease will not be cured after the flu is treated, therefore, some patients may be discharged even if they are in the high-risk group. Other serious complications such as high fever and lung inflammation can develop after the

patient is admitted, thus a patient can be admitted as a non-high-risk patient and transferred to high risk later. We assume that this process is Markovian, that is, the patients' future states only depend on their current states. The transition matrix, **P**, is defined as follows.

$$\mathbf{P} = \begin{array}{c} \\ H \\ D \\ z_1 \\ z_2 \end{array} \begin{array}{c} \begin{array}{cccc} H & D & z_1 & z_2 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ x_1 & x_2 & x_3 & x_4 \\ x_5 & x_6 & x_7 & x_8 \end{pmatrix} \end{array} \tag{1}$$

Figure 11.1 shows the state transition diagram for a patient after s/he is admitted to the hospital. The variables defined in Table 11.3 represent a patient's age-group, risk group, and final health outcome. For each age-group, the Markov transition matrix is different, as well as the ratio of non-high/high-risk admitting conditions. We calculate the admitting condition of the patients in each age-group, and then determine the Markov transition matrix for each age-group in the next two sections.

11.4.1.1 Age and risk groups distributions of admitted patients

There are limited resources regarding disease spread and its impact on hospital resources. We use data from Meltzer *et al* (1999), where a large amount of the data is provided either based on statistics or simply based

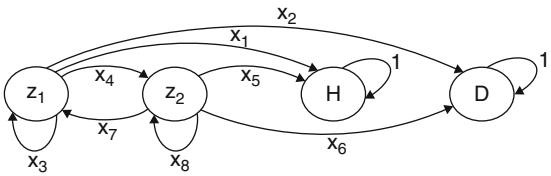


Figure 11.1 State transition diagram that describes a patient's condition over time, given that the patient is admitted to the hospital in state z_1 or z_2, and leaves the hospital from state H or D

Table 11.3 Variables used to define a patient's age-group, risk group, and final health outcome

U	The age-group. u = 1, 2, 3 for 0–19 y/o, 20–64 y/o, and 65+ y/o, respectively
V	The risk group. v = 1, 2 for non-high risk and high risk, respectively
W	The health outcome. w=1 for hospitalized and recovered, and 2 for hospitalized but deceased

on assumption. Although this is not perfect, it is the most thorough data source for estimating the input parameters.

The assumed percentage at 'high risk' (Meltzer *et al,* 1999) is the probability that a person in age-group u of the general population is at high risk and is denoted by a_u. The probability that a person in certain age-group from the general population is at high risk is: a_1 = 6.4% (0–19 y/o), a_2 = 14.4% (20–64 y/o), a_3 = 40.0% (65+ y/o) (Meltzer *et al,* 1999).

The gross attack rate (GAR) is defined as the number of clinical cases of influenza per unit population. The medical definition of a clinical case of influenza is 'fever greater than or equal to 100 (37.8 C) plus either a cough or a sore throat without a known diagnosis' (Seasonal Flu, 2009). Those who are infected but only have very mild symptoms are not considered to be clinical cases (Meltzer *et al,* 1999) and are unlikely to go to the hospital. But in the SIR model, these non-clinical cases are considered and will actually be generated. Meltzer *et al* (1999) assumed that only a subset of the general population will become clinical cases and thus assumed 15–35% of the general population is clinically important. Due to the lack of precise data, we assume that 15–35% of the infected population (i.e., the new cases generated by the SIR model) is clinically important and therefore should be considered in our system. Also, because it is difficult to give a good estimation of GAR, we assume that different risk groups within a certain age-group have the same GAR. The assumed GAR for each age-group is shown in Table 11.4.

We estimate the total number of clinical cases for the three age-groups by multiplying the population in each age-group with its corresponding Expected Gross Attack Rate (EGAR). For a new clinical case, the probability of being in age-group u is determined by the number of clinical cases in age-group u divided by the total number of clinical cases. The estimated population and the number of clinical cases for each age-group, and the probabilities associated with each age-group are shown in Table 11.5 (these probabilities will be carried directly to the Arena simulation model).

Table 11.4 Assumed GAR for different age-groups (uniformly distributed) (Meltzer *et al*, 1999)

Age-group	Lower (%)	Upper (%)	Expected GAR (EGAR) (%)
0–19	35	42	38.5
20–64	12	23	17.5
65+	12	19	17.5

Table 11.5 U.S. population and total number of clinical cases by age-groups

Age-group	Population (2007) (Table NC-EST2007-01 2007)	EGAR (from Table 11.4) (%)	Clinical cases (Population* EGAR)	Probability of being in the age-group (%)
0–19	82,361,752	38.5	31,709,274.52	45.74
20–64	181,371,447	17.5	31,740,003.23	45.79
65+	37,887,958	15.5	5,872,633.49	8.47

Table 11.6 Rates of health outcomes by age and risk group (Meltzer *et al*, 1999)

	Lower	Most likely	Upper	Distribution	Expected number per 1,000 clinical cases (denoted by e_{uvw})
Hospitalizations (per 1,000 clinical cases)					
Non-high risk					
0–19 y/o (rate)	0.57	N/A	6.9	Uniform	3.735
20–64 y/o (rate)	1.50	N/A	12.0	Uniform	6.750
65 + y/o (rate)	12.50	N/A	15.8	Uniform	14.150
High risk					
0–19 y/o (rate)	6.0	N/A	21.4	Uniform	13.70
20–64 y/o (rate)	6.9	N/A	22.3	Uniform	14.60
65 + y/o (rate)	33.3	N/A	68.4	Uniform	50.85
Deaths (per 1,000 clinical cases)					
Non-high risk					
0–19 y/o (rate)	0.041	0.07	0.30	Triangular	0.137
20–64 y/o (rate)	0.210	0.31	0.41	Triangular	0.310
65 + y/o (rate)	2.300	3.51	4.52	Triangular	3.443
High risk					
0–19 y/o (rate)	0.4	0.6	21.9	Triangular	7.633
20–64 y/o (rate)	0.8	N/A	24.9	Uniform	12.850
65 + y/o (rate)	23.0	N/A	29.6	Uniform	26.300

Given the probability that an incoming patient will be in a certain age and risk group, we find the simulation parameters that describe the health outcome for each age and risk group. Again, we assume that for a new infection, the probability of being in age-group u is the same as for a new clinical case. Table 11.6 presents the probability distributions for the rates of health outcomes for different age and risk groups that are estimated by Meltzer *et al* (1999). The data in Table 11.6 are interpreted as follows (taking the 0–19 y/o, non-high-risk group as an example): For every 1,000 new cases that are in the 0–19 y/o age-group and are not at

high risk, the number of patients who are hospitalized and recover is uniformly distributed with a lower bound of 0.57 and an upper bound of 6.9, so the expected number of hospitalizations per 1,000 cases is 3.735.

We assume that the death outcome only happens after the patient is hospitalized. We calculate the total expected number of hospitalizations (denoted by te_{uv}) per 1,000 cases as the sum of the expected number of patients hospitalized and finally recovered (e_{uv1}) and the expected number of patients who die (e_{uv2}) in the corresponding age and risk groups, for example, for the 0–19 y/o, non-high-risk group, in every 1,000 new cases, the number of hospitalizations, including both recovered and dead cases is 3.872 (= 3.735+0.137). Table 11.7 summarizes the proportions of patients hospitalized and recovered (e_{uv1}) and the proportion of patients who die (e_{uv2}) in each age and risk group, and the proportion of hospitalization in each age and risk group.

The percentage of non-high-risk patients hospitalized in age-group u is denoted by y_u, and equals to proportion of non-high-risk hospitalization divided by the sum of proportions of non-high and high-risk hospitalization in age-group u.

$$y_u = \frac{te_{u1}(1-\alpha_u)}{te_{u1}(1-\alpha_u)+te_{u2}\alpha_u} \tag{2}$$

The portion of inpatients admitted with non-high (high) risk condition and finally recovered in age-group u, $pu(q_u)$, is the proportion of non-high (high) risk recovered patients divided by the proportion of hospitalization per 1,000 cases in age-group u. The death rate per 1,000 hospitalized patients in age-group u, d_u, is 1 minus the sum of percentage of non-high-risk patients hospitalized and recovered ($y_u p_u$) and percentage of high-risk patients hospitalized and recovered ($(1–y_u)q_u$) in age-group u.

Table 11.7 Proportions of hospitalized but finally recovered patients by age and risk groups

	Recovered per 1,000 cases (e_{uv1})		**Death per 1,000 cases (e_{uv2})**		**Hospitalization per 1,000 cases (e_{uv1}+e_{uv2})**	
Age-group	**Non-high**	**High**	**Non-high**	**High**	**Non-high**	**High**
0–19	3.735	13.70	0.137	7.633	3.872	21.333
20–64	6.750	14.60	0.310	8.567	7.060	27.450
65+	14.15	50.85	3.443	17.533	17.593	77.150

Table 11.8 Percentage of non-high-risk inpatients (y_u), inpatients admitted with non-high/high-risk condition and finally recovered (p_u/q_u), and death rate (d_u)

Age-group	y_u	p_u	q_u	d_u
0–19	0.726	0.965	0.642	0.124
20–64	0.605	0.956	0.688	0.150
65+	0.255	0.804	0.773	0.219

$$d_u = 1 - (y_u p_u + (1 - y_u)\, q_u) \tag{3}$$

Table 11.8 presents a summary of these percentages for each age and risk group. Please refer to Figure 11.6 of the Appendixes A and B to see the details about how we implement these parameters in the simulation model in Arena.

At this point, we have established all of the parameters that connect our SIR model with the hospitalization model. Next, we solve to find the age-group-dependent Markov transition matrix, which regulates the patients' movement among different states. The final outcome of the patients must be in accordance with certain parameters such as the recovery and death percentages given different admitting conditions, age-groups, and the length of stay.

11.4.1.2 Patient condition dynamics within the hospital

We assume a patient's condition changes according to a Markov chain, which is age-dependent. There are four states in the Markov chain, Healthy (H), Dead (D), non-high risk (z_1), and high risk (z_2). Healthy and Dead are absorbing states as they represent departure from the hospital. We use properties of absorbing-state Markov chains to derive the transition probabilities (refer to Eq. (1)). As discussed in Appendix A, the percentage of non-high-risk inpatients and the percentage recovered are used to derive age-dependent transition probability matrices based on a combination of the time to absorption estimates and the fact that the rows of the transition matrices must sum to one. Table 11.9 shows the results obtained from MATLAB after applying the criteria for deriving the transition probabilities. These criteria are described in detail in Appendix A. For the age-group of 0–19, only nine candidates satisfy the criteria, among which we select the one with the smallest absolute error (AE) as our final solution, which is bolded. We choose the solution with the smallest AE for the 20–64 age-group. For the 65+ age-group, we choose the solution with the smallest ratio of x_4/x_3 because the probability for a non-high-risk patient transferring to high-risk condition is significantly

Table 11.9 Results of transition probability matrix for the three age-groups

Age-group	x_1	x_3	x_4	x_5	x_6	x_7	x_8	*AE*
0~19	0.25	0.70	0.05	0.05	0.20	0.35	0.40	0.075138
	0.25	0.70	0.05	0.05	0.20	0.40	0.35	0.073371
	0.25	0.70	0.05	0.10	0.15	0.20	0.55	0.053600
	0.25	0.70	0.05	0.10	0.15	0.25	0.50	0.099345
	0.25	0.70	0.05	0.15	0.10	0.05	0.70	0.073371
	0.30	**0.65**	**0.05**	**0.05**	**0.10**	**0.15**	**0.70**	**0.044779**
	0.55	0.35	0.10	0.05	0.05	0.05	0.85	0.083935
	0.65	0.20	0.15	0.05	0.05	0.05	0.85	0.071289
	0.75	0.05	0.20	0.05	0.05	0.05	0.85	0.097826
20~64	**0.15**	**0.80**	**0.05**	**0.10**	**0.20**	**0.45**	**0.25**	**0.081647**
	0.15	0.80	0.05	0.10	0.25	0.65	0	0.092101
65+	0.20	0.05	0.75	0.05	0.05	0.60	0.30	0.064558
	0.20	0.10	0.70	0.05	0.05	0.55	0.35	0.068550
	0.20	0.20	0.60	0.05	0.05	0.50	0.40	0.087611
	0.20	**0.25**	**0.55**	**0.05**	**0.05**	**0.45**	**0.45**	**0.076364**

higher than staying in the non-high-risk condition for other solutions. The chosen solutions are also displayed in bold. Figures 11.10 and 11.11 provide a description of this part of the model in Arena.

After obtaining the Markov transition matrix for each age-group, we compare these transition probabilities and the parameters in Table 11.2 with other reported data including Thompson *et al*'s (2003, 2004) and Meltzer *et al*'s (1999) in terms of the total hospitalization rate, total death rate, and length of stay in hospital. We discuss this topic in the section 'Analysis of the performance of the disease simulation model'.

11.4.1.3 Analysis of the performance of the disease simulation model

In the U.S., it is estimated that there are 36,000 deaths and 200,000 hospitalizations each year due to influenza (Thompson *et al,* 2003; Thompson *et al,* 2004). The U.S. population is approximately 300 million as of July 2008. Scaling this to a population size of 50,000 implies that there are (36,000/300,000,000) × 50,000≈6 deaths and (20,000/300,000,000) × 50,000≈33 hospitalizations. We run the simulation model for 30 replications with simulation length of 200 days in each replication and GAR of 15% and compare key simulation results with the data from Thompson's and Meltzer's papers in Table 11.10. Compared with other sources, our model results are statistically equivalent.

In order to determine an appropriate GAR for our model, we performed a test against different GAR ranges from 15 to 35%, with a 5% increase at each step. Results are shown in Table 11.11.

Table 11.10 Comparison between simulation results and other literature

	Simulation results	Half width	Minimum average	Maximum average	Other literature
Total deaths	5.63	0.81	2	10	6[a]
Total hospitalization	33.33	2.19	20	47	33[a]
0~19 length of stay (days)	4.18	0.46	1.49	6.80	4[b]
20~64 length of stay (days)	5.40	0.59	2.38	8.42	5.8[b]
65+ length of stay (days)	6.92	0.75	1.34	12.16	7.1[b]

[a]Thompson *et al*, 2003; Thompson *et al*, 2004.
[b]Meltzer *et al*, 1999.

Table 11.11 Sensitivity analysis of GAR

GAR (%)	Total deaths	Total hospitalizations	Age-group 1 length of stay	Age-group 2 length of stay	Age-group 3 length of stay
15	5.63	33.33	4.18	5.40	6.92
20	6.27	44.80	3.68	5.54	6.31
25	9.43	52.40	3.89	5.30	6.59
30	11.10	61.27	3.51	5.22	6.11
35	13.37	72.97	3.62	5.15	6.61

Figure 11.2 is a plot of the total number of deaths and the total number of hospitalizations *vs* GAR and shows that both the total number of deaths and hospitalizations appear to have a linear relationship with GAR. This is because a higher GAR means there are more clinical cases, hence more hospitalizations and deaths. The length of stay is not affected by GAR because it is determined by the characteristics of the Markov transition matrix, which is independent of GAR. In the following section, we develop the pharmacy model, which makes production and ordering decisions in each period.

11.4.2 The pharmacy model

The pharmacy in a hospital is basically an inventory system. There are two types of pharmaceutical items in inventory. One is the raw material, which has a long (we assume infinite) shelf life; the other type of item is the finished good (FG), which has a short shelf life (we assume it is 2 days). We assume the conversion rate is one to one, meaning one

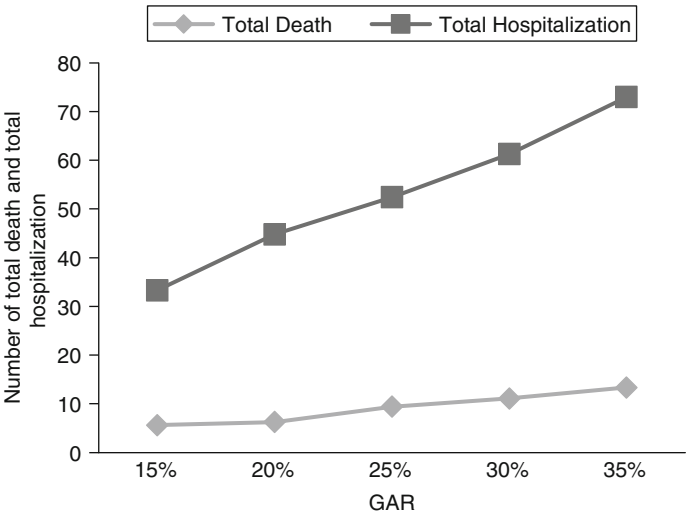


Figure 11.2 Effect of GAR on total hospitalization and total death

unit of raw material can be made into one unit of finished goods. The pharmacy orders raw material with a 1-day lead time for the order. The finished goods can be made from the raw material or purchased from outside sources for emergency use when both the raw material and the finished goods are out of stock. After finished goods expire, they are discarded with a cost. Figure 11.3 summarizes how this system functions.

At the beginning of each period (day), the hospital pharmacy receives the raw material order placed one period before, produces finished goods, and places a new order for raw material. Then the medication is distributed to the patients who are hospitalized during the period. The issuing policy for the finished goods is first in first out. The medication demand follows a discretized exponential distribution with different means associated with different patient conditions and age-groups. If there is enough raw material but not enough finished goods, then the pharmacy can make new finished goods from the raw material on hand, resulting in a finished goods stockout cost. If there is not enough FG or raw material, patients' needs are satisfied by expediting. This stockout of both finished goods and raw material has an impact on total cost. Specifically, if there is neither enough finished goods nor enough raw material, then the pharmacy must order finished goods from an outside source immediately (which it receives instantly) to satisfy demand and generate a raw material stockout cost. At the end of each period, the patients' conditions change according to the transition probability

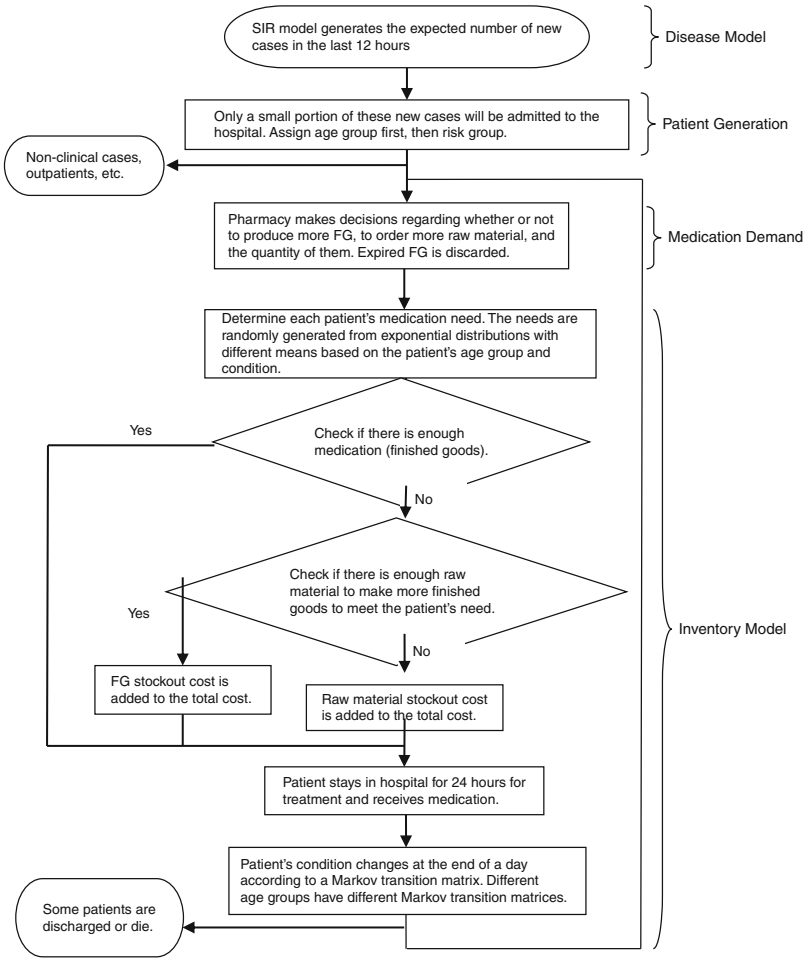


Figure 11.3 Flow chart of the hospital pharmacy inventory system model

matrix. Figures 11.7, 11.8, and 11.9 in Appendix B illustrate how we simulate this process in Arena. In the next section, we develop a mathematical model of the hospitalization and inventory system.

11.4.2.1 Model notation and assumptions

In order to formulate the inventory system models and further develop the objective function, the following notation and assumptions are used throughout this paper.

Assumptions:

(i) d_{uv} is a discrete random variable. Its probability mass function is a discretized exponential distribution with mean μ_{uv}. The probability mass function is:

$$P(\mathrm{d}_{uv} = \delta) = \int_{\delta}^{\delta+1} \mu_{uv} e^{-\mu_{uv} x} \mathrm{dx}.$$

(ii) The two variables $a(\tau)$ and $RO(\tau)$ are the decisions that the pharmacy will make at the beginning of every period except period 0.

The timeline associated with the events corresponding to the variables defined in Table 11.12 is shown in Figure 11.4. The events occur in the following order:

(a) The number of patients and the conditions (states) of the patients in the hospital, $n_{uv}(\tau)$, are updated. The amount of raw material, $m(\tau)$, is updated.
(b) The pharmacy disposes expired finished goods.
(c) An outstanding order of raw material placed one period ago arrives.
(d) The pharmacy decides the amount of finished goods to make ($a(\tau)$) and the amount of raw material to order ($RO(\tau)$).

Table 11.12 Notation

d_{uv}	the medication needs for age-group u and risk group v
$n_{uv}(\tau)$	the number of patients in the hospital in age-group u and risk group v at the beginning of period τ after transition of patients' states
$N_{uv}(\tau)$	the number of patients in age-group u and risk group v admitted into the hospital during period τ
$m(\tau)$	the amount of raw material on hand at the beginning of period τ, before the pharmacy receives the outstanding order and before the production of finished goods for period τ
$FG(\tau, 1)$, $FG(\tau, 2)$	the number of finished goods at the beginning of period τ with remaining shelf life of 1 and 2 days, respectively
$a(\tau)$	the amount of finished goods to produce at the beginning of period τ
$RO(\tau)$	the amount of raw material to order at the beginning of period τ, which will be received at the end of period τ
$D_\tau(n_{uv}(\tau), N_{uv}(\tau))$	demand realized during period τ, which will be simplified as $D_\tau(\cdot)$

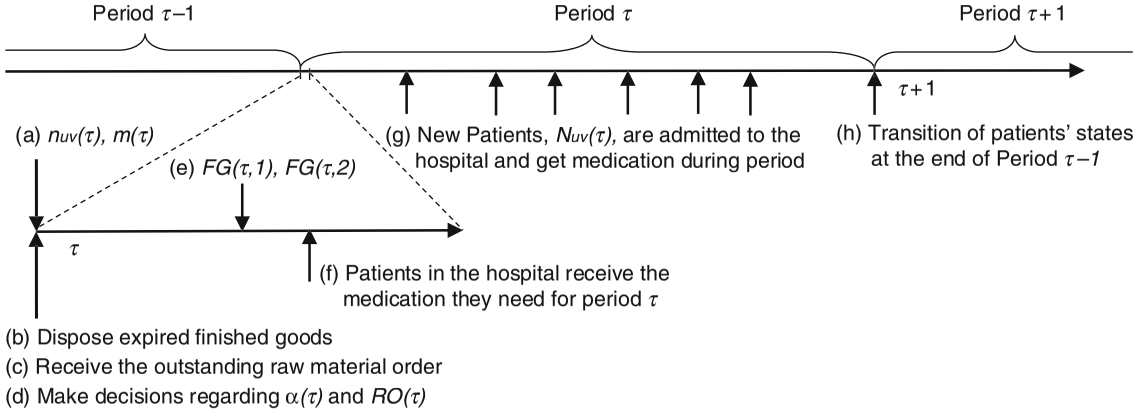


Figure 11.4 Timeline of events with corresponding variable definitions

(e) The amount of finished goods, $FG(\tau, 1)$ and $FG(\tau, 2)$, is updated after production.
(f) The pharmacy distributes the medication to patients in hospital.
(g) New patients, $N_{uv}(\tau)$, are admitted to the hospital and get medication throughout the period.
(h) Patients' states change according to the Markov chains at the end of the period.

In the next section, we will formulate the total cost function of the system.

11.4.2.2 Total cost function

The objective is to minimize the total cost of the system. The inventory status at the beginning of period τ is defined by the raw and finished goods inventory on hand at the beginning of period τ: $m(\tau)$, $FG(\tau, 1)$, and $FG(\tau, 2)$. We define the relationship between the inventory status at the beginning of period τ and τ+1 as follows. The amount of finished goods on hand at the beginning of period τ+1, with a 2-day shelf life is:

$$FG(\tau+1,2) = a(\tau+1). \tag{4}$$

If we define the following: $v^+ = \max(v, 0)$, then $m(\tau+1)$, the amount of raw material on hand at the end of period τ, before receiving the outstanding raw material order and before finished goods production, is:

$$\begin{aligned} m(\tau+1) = \{&m(\tau) + RO(\tau-1) - a(\tau) \\ &-[D_\tau(\cdot) - (FG(\tau,1) + FG(\tau,2))]^+\}^+. \end{aligned} \tag{5}$$

The amount of finished goods on hand with 1-day shelf life at the beginning of period τ is:

$$FG(\tau+1,1) = [FG(\tau,2) - (D_\tau(\cdot) - FG(\tau,1))^+]^+. \tag{6}$$

The total cost of the system consists of the following parts: (taking period τ as an example)

(i) Raw material holding cost with cost-coefficient RH (per unit per day)

$$\begin{aligned} RH\{&m(\tau) + RO(\tau-1) - a(\tau) \\ &-[D_\tau(\cdot) - (FG(\tau,1) + FG(\tau,2))]^+\}^+. \end{aligned} \tag{7}$$

(ii) Expired FG cost with cost-coefficient EF (per unit)

$$EF[FG(\tau,1)-(D_\tau(\cdot)]^+. \tag{8}$$

(iii) Raw material stockout cost with cost-coefficient RS (per unit)

$$\begin{aligned} &RS[D_\tau(\cdot)-(m(\tau)+RO(\tau-1)-a(\tau)\\ &\quad+FG(\tau,1)+FG(\tau,2))]^+. \end{aligned} \tag{9}$$

(iv) FG stockout with cost-coefficient FS (per unit)

$$FS[D_\tau(\cdot)-(FG(\tau,1)+FG(\tau,2))]^+. \tag{10}$$

(v) Purchasing cost for raw material, A (per unit)

$$ARO(\tau). \tag{11}$$

(vi) Production cost, B (per unit)

$$Ba(\tau). \tag{12}$$

The total cost incurred in period τ, C_τ, is the sum of (7)–(12):

$$\begin{aligned} &C_\tau(m(\tau),FG(\tau,1),FG(\tau,2),D_\tau(\cdot),a(\tau),RO(\tau))\\ &\quad=RH\{m(\tau)+RO(\tau-1)-a(\tau)\\ &\quad-[D_\tau(\cdot)-FG(\tau,1)+FG(\tau,2))]^+\}^+\\ &\quad+EF[FG(\tau,1)-D_\tau(\cdot)]^+ +[RS(D_\tau(\cdot)-m(\tau).\\ &\quad+RO(\tau-1)-a(\tau)+FG(\tau,1)+FG(\tau,2))]^+\\ &\quad+FS[D_\tau(\cdot)-(FG(\tau,1)+FG(\tau,2))]^+ +ARO(\tau)+Ba(\tau) \end{aligned} \tag{13}$$

For simulation-based policies, the objective function is:

$$\min\left(\sum_\tau C_\tau(\cdot)\right). \tag{14}$$

For the dynamic programming-based procedure, TC_τ, the total cost from period 0 to period τ is defined by a recursive function, which is the minimum expected cost with respect to demand $D_\tau(\cdot)$:

$$\begin{aligned} &TC_\tau(m(\tau),FG(\tau,1),FG(\tau,2),D_\tau(\cdot))\\ &\quad=\min_{0\le a(\tau)\le m(\tau)+RO(\tau-1)RO(\tau)\ge 0}\{E_{D_{\tau t}(\cdot)}[C_\tau(\cdot)]\\ &\quad+E_{a(\tau-1),RO(t-1)}\\ &\quad\left[TC_{\tau-1}\begin{pmatrix}m(\tau-1),FG(\tau-1,1,))\\ FG(\tau-1,2),D_{\tau-1}(\cdot)\end{pmatrix}\right]\Bigg\},\\ &\quad\tau=1,2,\ldots,T. \end{aligned} \tag{15}$$

The computational effort required to apply backwards recursion for the DP is significant. In order to determine the minimum cost and resulting near-optimal policy, we evaluate all possible combinations of inventory status and decision variable values. Each combination has three elements that reflect the inventory status, $r(\tau)$, $FG(\tau, 1)$, and $FG(\tau, 2)$, and two decision variables, $a(\tau)$ and $RO(\tau)$. The relationship between $FG(\tau, 2)$ and $a(\tau)$, that is, $FG(\tau, 2) = a(\tau)$, eliminates one of the five elements. For example, when we ran the simulation model for 200 days and 30 replications. The largest daily medication demand is 53 units. If we only evaluate 0–49 for each of the four elements, there will be $50^4 = 6.25$ million possible combinations for each stage (period), except for period 0, where we assume that the inventory is predetermined and there are no decisions to make in this period. Assuming a horizon of 199 periods, there will be approximately 1.25 billion steps in total. Implementing the DP procedure in C++, it will take roughly 100 days to find the optimal solution. To address this issue, we consider a reduced problem size and use the solution of this smaller problem to seed our simulation-based inventory models. In the following section, we propose three inventory policies and compare their performance.

11.4.2.3 Proposed inventory policies and their performance

We consider various ordering and production policies and present and compare the results of each in the discussion that follows. For an inventory policy to be considered efficient, it has to make the appropriate decisions about the amount of medication to produce in period τ, $a(\tau)$, and the amount of raw material to order in period τ, $RO(\tau)$. The cost structure of the problem, that is, the inventory carrying cost and the stockout cost, play an important role along with the expected medication demand. Since the probability distribution for medication demand is extremely complicated due to the uncertainty regarding the spread of the disease and patient behavior while hospitalized, a heuristic method is a good option to determine $a(\tau)$ and $RO(\tau)$.

Policy I The Markov property implies that medication needs for the period τ are related to the current period's hospitalized patient mix, that is, $n_{uv}(\tau)$. Policy I is based on this property. At the beginning of each period, there are two decisions to make: (i) the amount of medication to produce in period τ, $a(\tau)$ and (ii) the amount of raw material to order in period τ, $RO(\tau)$, which has a lead time of 1 day.

The production quantity $a(\tau)$ is determined by two factors. The first is the estimate of the medication needs for the patients who are already in

the system and will continue to stay in the hospital for the next period. This quantity is determined by multiplying the number of patients in different categories, $n_{uv}(\tau)$ (i.e., different age and risk groups) by their specific average needs, μ_{uv}, for all u and v. (Recall that the patients' states are assumed not to change until after time period τ.) The second factor is the estimate of the medication needs for incoming patients, which is calculated by multiplying the number of patients who were admitted in the previous period, denoted by $N_{uv}(\tau–1)$, by their specific needs. The estimated medication need in period τ is:

$$\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))]. \tag{16}$$

The amount of medication that needs to be prepared is:

$$\left\{\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))]-FG(t,1)\right\}^{+}. \tag{17}$$

The estimated medication quantity to prepare cannot exceed the total raw material on hand at the time when the pharmacy is making the finished goods for period τ, which is:

$$\begin{aligned}&\{m(\tau-1)-[D_{\tau-1}(\cdot)-(FG(\tau-1,1)+FG(\tau-1,2))]^{+}\}^{+}.\\&\quad+RO(\tau-1)\end{aligned} \tag{18}$$

So for Policy I, we can write the amount of medication to prepare as:

$$\begin{aligned}a(\tau)=\min\Bigg\{&\Bigg\{\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))]\\&-FG(\tau,1)\}^{+},\{m(\tau-1)-[D_{\tau-1}(\cdot)-(FG(\tau-1,1)^{+}\\&+FG(\tau-1,2))]^{+}\}^{+}+RO(\tau-1)\}.\end{aligned} \tag{19}$$

For $RO(\tau)$, we will place an order that is equal to the expected need:

$$RO(\tau)=\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))]. \tag{20}$$

Policy I is generated only using the expected demand as the decisions. Recall the Newsvendor Problem, where the optimum number of papers to order is not always the expected demand. Policy II, while still based on the expected demand, allows some flexibility.

Policy II Policy II utilizes OptQuest from Arena to optimize the selection of coefficients, ρ and θ, which serve as multipliers for the decision

variables $a(\tau)$ and $RO(\tau)$ from Policy I. This policy allows for a global scaling of the order quantities that will be applied to all periods. We define the decision variables $a(\tau)$ and $RO(\tau)$ for Policy II as follows:

$$\begin{aligned} a(\tau) = \min\Bigg\{\Bigg[\rho\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))] \\ \{-FG(\tau,1)\}^{+}+0.5\Bigg\}\{m(\tau-1)-[D_{\tau-1}(\cdot) \\ -(FG(\tau-1,1)+FG(\tau-1,2))]^{+}\}^{+} \\ +RO(\tau-1)\}, \end{aligned} \tag{21}$$

$$RO(\tau) = \left\lfloor \theta\sum_{u,v}[\mu_{uv}(n_{uv}(\tau)+N_{uv}(\tau-1))]+0.5 \right\rfloor, \tag{22}$$

where $\lfloor x+0.5 \rfloor$ is to round a number to the nearest integer. The values of ρ and θ are found by OptQuest such that they minimize the total cost of the system $\Sigma_{\tau} C_{\tau}(\cdot)$.

Policies I and II utilize the most recent information of the system, that is, the patient mix prior to making decisions and the forecast of the incoming patients in the next period. Policy III uses all medication demand records from historical data.

Policy III For Policy III, we only use the average demand extracted from historical data. The historical data are generated by simulation and the expected demand for each day (period) is the average of the demand of the same day (period) from all past years (replications). We assume $X_{\tau,yr}$ is the realized demand in period τ in year yr, where yr= 1,2,... ,YR. Then $\overline{X}_{\tau}$, the average demand in period τ is:

$$\overline{X}\tau = \left\lfloor \frac{\sum_{yr} X_{\tau,yr}}{YR} + 0.5 \right\rfloor. \tag{23}$$

The available amount of raw material left on hand prior to making decisions is the same as in Policies I and II (refer to Eq. (18)). Hence, the decisions are:

$$\begin{aligned} a(\tau) = \min\{\overline{X}_{\tau}, \{m(\tau-1)-D_{\tau-1}(\cdot)-(FG(\tau-1,1) \\ +FG(\tau-1,2))]^{+}\}^{+}+RO(\tau-1)\}, \end{aligned} \tag{24}$$

$$RO(\tau) = \overline{X}_{\tau}. \tag{25}$$

Policy IV is different from Policies I through III in the sense that it is not heuristic, but is based on dynamic programming.

Policy IV For Policy IV, the objective function is Eq. (15). The near-optimal decision variables $a(\tau)$ and $RO(\tau)$ are obtained by solving the backwards recursive function of (15). In the next section, we evaluate Policies I through III in Arena, and compare them with the results from Policy IV (DP).

11.5 Numerical Results

11.5.1 Cost Parameters and Demand Parameters

Before running the simulation model and the DP procedure, the cost parameters and the patients' demand rate are estimated. The values of these parameters are determined as follows. We assume that the raw material holding cost and ordering cost are negligible. The finished goods incur a cost when they expire. This cost may be expensive because it may require special processing to dispose of expired medication. If a finished goods stockout occurs, the pharmacy will make an emergency production to fulfill demand, which costs more than the normal production. If the raw material is out, then the pharmacy has to seek the medication from an outside source. This will result in an extremely high cost. The base cost coefficients are shown in Table 11.13. We define the patient demand rate μ_{uv} as shown in Table 11.14. Unless otherwise stated, we will use this base scenario in the rest of the paper.

11.5.2 Results

With all the parameters defined, we first run the simulation model and record the realized demand generated by Arena for 30 replications, which equals to *YR* in Eq. (23), with 200 days (periods) in each replication and no patients in the system at the beginning of the simulation. We omit a warm-up period in order to capture the effects of the start of

Table 11.13 Cost coefficients

Cost parameters	Cost	Cost parameters	Cost
Raw material holding cost with cost-coefficient RH (per unit per day)	1	FG stockout cost with cost-coefficient FS (per unit)	10
Expired FG cost with cost-coefficient EF (per unit)	8	Ordering cost for raw material, A (per unit)	1
Raw material stockout cost with cost-coefficient RS (per unit)	20	Production cost, B (per unit)	1

Table 11.14 Patients' demand rate μ_{ij}

Age-group u	Risk group v	
	1	2
1	2 units	4 units
2	1 unit	3 units
3	3 units	5 units

Table 11.15 Cost structure of Policies I and II in base scenario for 200 periods

	Corresponding equation	Policy I		Policy II	
		Cost ($)	% of total cost	Cost ($)	% of total cost
Total	(18)	7,277.60	100.00	6,070.47	100.00
Raw holding	(11)	1,297.53	17.83	1,150.60	18.95
Raw s/o	(13)	2,363.33	32.47	1,473.33	24.27
FG s/o	(14)	2,560.33	35.18	1,255.67	20.68
FG expiration	(12)	136.00	1.87	904.53	14.90
Ordering	(15)	531.87	7.31	671.27	11.06
Production	(16)	388.53	5.34	615.07	10.13

a pandemic, which can be compared to a system that starts empty. The run length of 200 days is based on the estimated October–May timing of the flu season in the U.S. as reported by the Center for Disease Control. We run 30 replications because it is sufficient to reduce the 95% confident intervals of the total cost to 10% of its average value, which can be seen in Tables 11.16 and 11.17. We use the recorded demand to calculate the demand distribution that will be used in the DP procedure. Let $M_{\tau j}$ be the total number of times demand is equal to j in period τ of all replications. The total number of replications is the same as *YR,* which is defined in Policy III, considering each replication represents one year. So the probability that the demand equals to j in period τ,$\mathbf{P}(D_\tau(\cdot) = j)$, is defined by:

$$\mathbf{P}(\mathrm{D}_\tau(\cdot) = j) = \frac{M_{\tau j}}{YR}. \tag{26}$$

Second, we perform the DP procedure with the demand distribution obtained from Eq. (26). Finally, we use the optimal solution from the

Table 11.16 Performance of proposed policies under different GARs

GAR (%)	m(0)	DP solution (Policy IV)	Simulation		
			Policy I	Policy II	Policy III
15	2	\$6,134.65	\$7,277.60±617.23	\$6,070.47±439.81 ρ=2.25 θ = 1.25	\$16,013.53±1,470.98
25	5	\$8,947.29	\$10,887.27±1,048.66	\$8,565.60±558.66 ρ=2.90 θ=1.45	\$22,409.67±2,231.69
35	5	\$11,065.4	\$13,400.87±792.69	\$10,217.00±585.11 ρ=3.80 θ=1.35	\$29,619.70±3,126.77

reduced DP as the starting value for the three simulation models that we compare.

The 'optimal' cost by solving the reduced DP is \$6,133.61 with initial raw inventory $m(0) = 2$, initial finished goods inventory $FG(0,1) = 0$, $FG(0, 2) = 0$. We realize that the range of decision variables (from 0 to 7) may not contain the optimal decisions, thus it is likely to find a lower cost if we could evaluate more possible combinations of decision variables.

We set the initial raw inventory $m(0)$ to be 2, which is found by DP, for all three simulation-based policies and present the results in Table 11.17. The expected cost of Policy I with the same starting inventory is \$7,277.60 with half width of \$617.23. For Policy II, OptQuest found $\rho = 2.25$, $\theta = 1.25$ to be optimal. (Throughout the paper, the range evaluated for both ρ and θ in OptQuest is set to be from 0.05 to 4 with a step size of 0.05.) Using $\rho = 2.25$, $\theta = 1.25$ to run the simulation for 30 replications, the cost is \$6,070.47±439.81. Policy III gives an average cost of \$16,013.53 with half width of \$1,470.98. The initial results obtained under the base scenario show that Policies I and III have significantly higher costs than Policy II.

Policy I only takes the expected medication demand into consideration due to the unknown distribution of medication demand. Policy II takes advantage of OptQuest for Arena. It takes into consideration the fact that the optimal decisions are affected by not only the expected demand, but also the demand distribution and cost coefficients. This is very similar to the Newsvendor Problem. In the various scenarios we tested, Policy II outperformed the other two policies. Policy III treats the historical data as the primary information, ignoring the fact that the medication demand is largely decided by the patients in the system; therefore, its performance is the worst.

In terms of the cost split, Table 11.15 compares the elements of the total cost for Policy I and Policy II. It is worth noting that for Policy I,

Table 11.17 Comparison of total cost of the proposed policies with respect to input costs, raw holding, expired finished good, raw and finished good stockout, ordering and production costs

	m(0)	Simulation		
		Policy I	**Policy II**	**Policy III**
Base scenario	2	\$7,277.60±617.23	\$6,070.47±439.81 ρ=*2.25* θ=1.25	\$16,013.53±1,470.98
Raw holding cost = 5	0	\$12,322.47±1,221.69	\$8,026.83±571.57 ρ=3.75 θ=1.15	\$44,815.17±6,020.66
Expired FG cost = 4	2	\$7,209.60±615.80	\$5,396.70±373.68 ρ=2.65 θ=1.35	\$14,538.20±1,459.30
Raw s/o cost = 10	0	\$6,079.40±488.37	\$5,160.37±381.56 ρ=2.55 θ=1.15	\$14,076.60±1,217.04
FG s/o cost = 5	1	\$5,998.37±505.42	\$5,477.20±403.77 ρ=1.90 θ=1.15	\$14,398.40±1,397.18
Ordering cost = 2	2	\$7,809.47±653.65	\$6,727.67±497.14 ρ=2.35 θ=1.20	\$16,017.50±1,471.34
Production cost=2	2	\$7,666.13±644.08	\$6,678.80±495.17 ρ=2.35 θ=1.20	\$16,438.40±1,469.71

the raw material holding cost is much lower than the raw material stockout cost, which means that Policy I ordered fewer raw materials than desired. Policy II adjusted for this problem so that the raw material holding cost was comparable to the raw material stockout cost. A similar adjustment is found in the FG stockout cost and FG expiration cost. By balancing within each cost pair, that is, the raw material holding/stockout cost and the FG stockout/expiration cost, Policy II reduced the total cost by 16.59% compared with Policy I.

11.5.3 Sensitivity Analysis

Because GAR and the cost coefficients are obtained mostly by estimation and are uncertain, we would like to evaluate their affect on the inventory policies and optimal solutions when they vary.

11.5.3.1 Effect of attack rate on optimal solutions

As stated in the subsection 'Age and risk groups distributions of admitted patients', GAR is difficult to estimate, yet it has a strong influence on the total number of patients generated in the simulation mode; hence, GAR can affect medication demand significantly. We perform sensitivity analysis on GAR using three rates: 15%, 25%, and 35%, and evaluate the impact on policy performance.

First, we determine the optimal initial state by running the DP procedure. Then we use this initial condition to run the simulation for the three policies. We restrict the initial states to be 0 for finished goods ($FG(0, 1) = 0$, $FG(0, 2) = 0$), the production quantity ($a(0) = 0$), and the order quantity ($RO(0) = 0$), except for the raw inventory, $m(0)$, which will be evaluated from 0 to 7.

When we use 25% as GAR as opposed to 15%, which is used in the base scenario, the DP finds the optimal solution to be \$8,947.29 with initial raw inventory $r(0) = 5$, initial finished goods inventory $FG(0, 1) = 0$, $FG(0, 2) = 0$. With the same starting inventory, the simulation gives an expected total cost of \$10,887.27 with half width of \$1,048.66 for Policy I; a cost of \$8,565.50 with half width of \$558.66 for Policy II; a cost of \$22,409.67 with half width of \$2,231.69 for Policy III. We also compared the results from DP and simulation under a 35% GAR. See Table 11.16 for a detailed comparison.

11.5.3.2 Effect of cost parameters on optimal solutions

In Table 11.17, we compare the effect of different cost parameters on the performance of our proposed policies. Policy II performs the best among the three. Varying the cost coefficients changes not only the

minimal cost, but also the optimal initial inventory status. When the raw material holding cost coefficient is raised to 5/unit/day from 1/unit/day, it drives the optimal initial inventory status down to 0, because holding less initial raw material is favorable under a high holding cost. Similarly, decreasing the raw material stockout cost coefficient lowers the initial raw material inventory. The values of ρ and θ of Policy II correspond to the changes in the cost coefficients very well. As defined in subsection 'Proposed inventory policies and their performance', ρ is the coefficient related to the production decision, θ is related to the order decision. Increasing the raw material holding cost coefficient results in ordering less raw material (smaller θ) and converting more raw material on hand to finished goods (bigger ρ) if possible. Similar conclusions can be drawn for the FG expired cost coefficient, the raw material stockout cost coefficient, and the FG stockout coefficient. The total cost is not sensitive to the ordering cost and production cost, which is supported by the slight change of ρ and θ compared with base scenario.

11.6 Discussion

In this paper, we study a hospital pharmacy inventory system where medication demand is driven by an influenza outbreak. Infected patients of different age and risk groups are admitted to the hospital and transition between various health/disease states according a Markov chain. We develop a DP model and use the solution to a reduced version of the DP as an input for the three simulation-based policies. Of the three proposed polices, Policy II outperforms the other two polices. The sensitivity analysis shows that under a high GAR, which leads to high medication demand, Policy II is very favorable due to its easy implementation and shorter computation time.

There are certainly some limitations of our model and methodology. First, all the input parameters for the SIR model and the Markov transition probability matrix are based on other parameters that are uncertain. When facing the seasonal influenza, it is difficult to get accurate statistics because a lot of flu cases are not reported due to mild symptoms. Many parameters are based on experience and estimation. Though some key performance indicators of our model correspond to numbers reported in other literature, for example, the total number of hospitalizations, total number of deaths, and length of hospitalization, the probability distributions associated with those indicator variables remain unknown. Despite this limitation, our model provides a

framework for gleaning insight regarding how a hospital system might respond in the face of an influenza outbreak.

Second, the complexity of the DP prevents us from finding the optimal policy; therefore, we solve a smaller problem instance and obtain a sub-optimal policy and use it to inform our simulation models. Third, in the real world, the shortage of certain medications may affect the treatment of patients. This could result in changes to the Markov transition matrix for some patients who are not treated in time. It is worth noting that the pharmacy at a hospital may have production or storage limits. At the peak of the influenza season or under an unexpected attack of flu, an external source for medication supply may face a shortage as well, thus affecting the availability of medication at a hospital. All these factors can be incorporated into the model given the appropriate data to characterize these events.

References

Chung KJ and Ting PS (1993) A heuristic for replenishment of deteriorating items with a linear trend in demand. *Journal of the Operational Research Society* **44**(**12**), 1235–1241.

Dudden RF, Corcoran K, Kaplan J, Magouirk J, Rand DC and Smith BT (2006) The medical library association benchmarking network: results. *Journal of the Medical Library Association* **94**(**2**), 118–129.

Dutta P and Chakraborty D (2010) Incorporating one-way substitution policy into the newsboy problem with imprecise customer demand. *European Journal of Operational Research*, 99–110.

Dye C, Chang H and Teng J (2006) A deteriorating inventory model with time-varying demand and shortage-dependent partial back logging. *European Journal of Operational Research* **4**(**72**), 417–429.

Fritze J (2010) Medical expenses have 'very steep rate of growth'. *USA TODAY* [WWW document] http://www.usatoday30.usatoday.com/news/health/2010-02-04-health-care-costs_N.html.

Goh H, Greenberg BS and Matsuo H (1993) Two-stage perishable inventory models. *Management Science* **39**(5), 633–649.

Goyal SK and Giri BC (2001) Recent trends in modeling of deteriorating inventory. *European Journal of Operational Research* **134**(**1**), 1–16.

Grinstead C and Snell J (1997) Introduction to Probability. *American Mathematical Society.*

Haijema R, van der Wal J and van Dijk NM (2007) Blood platelet production: optimization by dynamic programming and simulation. *Computers & Operations Research* **34**(**3**), 760–779.

Heng KJ, Labban J and Linn RJ (1991) An order-level lot size inventory model for deteriorating items with finite replenishment rate. *Computers and Industrial Engineering* **20**(**2**), 187–197.

Hethcote HW (2000) The mathematics of infectious diseases. *SIAM Review* **42**(**4**), 559–653.

Hollier RH and Mak KL (1983) Inventory replenishment policies for deteriorating items in a declining market. *International Journal of Production Research* **21**(**6**), 813–826.

Kanchanasuntorn K and Techanitisawad A (2006) An approximate periodic model for fixed-life perishable products in a two-echelon inventory distribution system. *International Journal of Production Economics* **100**(**1**), 101–115.

Liu L and Lian Z (1999) (s,S) continuous review models for inventory with fixed lifetimes. *Operations Research* **74**(**1**), 150–158.

Meltzer MI, Cox NJ and Fukuda K (1999) The economic impact of pandemic influenza in the United States: priorities for intervention. *Emerging Infectious Diseases* **5**(**5**), 659–671.

Molinari NM *et al* (2007) The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine* **25**(**27**), 5086–5096.

Nahmias S (1982) Perishable inventory theory: a review. *Operations Research* **30**(**4**), 680–708.

Nandakumar P and Morton TE (1993) Near myopic heuristic for the fixed life perishability problem. *Management Science* **39**(**12**), 1490–1498.

Perry D (1997) A double band control policy of a Brownian perishable inventory system. *Probability in Engineering and Informational Sciences* **11**(**3**), 361–373.

Raafat F (1991) Survey of literature on continuously deteriorating inventory models. *Journals and Operational Research Society* **42**(**1**), 27–37.

Raafat F, Wolfe PM and Eldin HK (1991) An inventory model for deteriorating items. *Computers and Industrial Engineering* **20**(**1**), 89–94.

Seasonal Flu. (2009) Nebraska Department of Health and Human Service. [WWW document] http://www.hhs.state.ne.us/han/update042909-2.pdf.

Seasonal Influenza. (2010) Center for Disease Control and Prevention. [WWW document] http://www.cdc.gov/flu/about/qa/disease.htm.

Table NA-EST2002-01 – National Population Estimates: April 1, 2000 to July 1, 2002. (2002) Population Division, U.S. Census Bureau [WWW document] http://www.census.gov/popest/archives/2000s/vintage_2002/NA-EST2002-01.html.

Table NC-EST2007-01 – Annual Estimates of the Population by Sex and Five-Year Age Groups for the United States: April 1, 2000 to July 1, 2007. (2007) Population Division, U.S. Census Bureau [WWW document] http://www.census.gov/popest/national/asrh/NC-EST2007-sa.html.

Thompson W *et al* (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA: The Journal of the American Medical Association* **289**(**2**), 179–186.

Thompson W *et al* (2004) Influenza-associated hospitalizations in the United States. *JAMA: The Journal of the American Medical Association* **292**(**11**), 1333–1340.

Vila-Parrish A, Ivy JS and King RE (2008) A simulation-based approach for inventory modeling of perishable pharmaceuticals. *Proceedings of the 2008 Winter Simulation Conference,* 1532–1538.

Vila-Parrish A, Ivy J, King R and Abel S (2012) Patient-based pharmaceutical inventory management – a two-stage inventory and production model for perishable products with Markovian demand. *Health Systems* **1**(**1**), 69–83.

Xu H and Wang H (1990) An economic ordering policy model for deteriorating items with time proportional demand. *European Journal of Operational Research* **46**(**1**), 21–27.

Xu H and Wang H (1992) Optimal inventory policy for perishable items with time proportional demand. *IIE Transactions* **24**(**5**), 105–110.

Appendix A

We assume that a patient's condition (i.e., state) updates once a day, we define g_{uv} as the average length of stay for patients who are admitted into the hospital in age-group u and risk group v. For each specific age-group, L_u is the weighted average length of stay of patients in age-group u, where y_u is the percentage of non-high-risk patients as defined in Table 11.8.

$$L_u = g_{u1}y_u + g_{u2}(1 - y_u) \text{ for } u = 1,2,3,. \tag{A.1}$$

The estimation of length of stay (in days) is L_1 = 4 (0–17 years old (y/o)), L_2 = 5.8 (18–64 y/o), L_3 = 7.1 (65+ y/o) (Meltzer *et al*, 1999). Note that for the length of stay, Meltzer *et al* (1999) grouped 0–17 y/o together instead of 0–19 y/o. We assume that the 2-year difference in age-groups will not significantly change the average length of stay.

According to Theorem 11.4 and Theorem 11.5 (Grinstead & Laurie Snell, 1997), which state that starting from state s_v, the expected number of steps that the chain is absorbed into a certain state is the vth entry of $(I–Q)^{-1}c$, where c is a column vector with all entries being 1. Thus, for each age-group u,

$$g_{u1} = \frac{x_4 - x_8 + 1}{(1 - x_3)(1 - x_8) - x_4x_7}, \tag{A.2}$$

$$g_{u2} = \frac{x_7 - x_3 + 1}{(1 - x_3)(1 - x_8) - x_4x_7}, \tag{A.3}$$

where x_n's are the entries of corresponding age-group-specific transition probability matrix for age-group u. For simplification, we will use y, p, q, and L to represent y_u, q_u, p_u, and L_u. Substituting (A.2), (A.3) into (A.1), we have

$$\begin{aligned} &y\frac{x_4 - x_8 + 1}{(1 - x_3)(1 - x_8) - x_4x_7} \\ &+(1 - y)\frac{x_7 - x_3 + 1}{(1 - x_3)(1 - x_8) - x_4x_7} = L. \end{aligned} \tag{A.4}$$

Theorem 11.6 in Grinstead & Laurie Snell (1999) states the probability that patients admitted with different conditions end up in H or D is $(I–Q)^{-1}R$, where R is from the rearranged transition matrix (1)

$$R = \begin{matrix} \\ z_1 \\ z_2 \end{matrix} \begin{bmatrix} \text{H} & \text{D} \\ \dfrac{x_1(1 - x_8) + x_4x_5}{(1 - x_3)(1 - x_8) - x_4x_7} & \dfrac{x_4x_6}{(1 - x_3)(1 - x_8) - x_4x_7} \\ \dfrac{x_1x_7 + x_5(1 - x_3)}{(1 - x_3)(1 - x_8) - x_4x_7} & \dfrac{x_6(1 - x_3)}{(1 - x_3)(1 - x_8) - x_4x_7} \end{bmatrix}. \tag{A.5}$$

We have calculated the portion of patients admitted with non-high-risk conditions and finally recovered (p_u) and with high-risk condition and finally

recovered (q_u) in Table 11.8. Thus the portion of patients who are admitted with a non-high risk condition and die during their stay is $1–p_u$, and the portion of patients who are admitted with high-risk condition and die is $1–q_u$.

$$R = \begin{matrix} & \text{H} & \text{D} \\ z_1 & p_u & 1-p_u \\ z_2 & q_u & 1-q_u \end{matrix} \tag{A.6}$$

Equating matrix (A.5) with matrix (A.6), we have:

$$\begin{bmatrix} \dfrac{x_1(1-x_8)+x_4x_5}{(1-x_3)(1-x_8)-x_4x_7} & \dfrac{x_4x_6}{(1-x_3)(1-x_8)-x_4x_7} \\ \dfrac{x_1x_7+x_5(1-x_3)}{(1-x_3)(1-x_8)-x_4x_7} & \dfrac{x_6(1-x_3)}{(1-x_3)(1-x_8)-x_4x_7} \end{bmatrix} = \begin{bmatrix} p_u & 1-p_u \\ q_u & 1-q_u \end{bmatrix} \tag{A.7}$$

It is also true that the summation of each row of the transition matrix has to be 1.

$$x_1 + x_2 + x_3 + x_4 = 1. \tag{A8}$$
$$x_5 + x_6 + x_7 + x_8 = 1. \tag{A9}$$

Eqs. (A.4), (A.7)–(A.9) are a set of nonlinear equations and the unknown variables of x_n's are subject to $x_n \in [0, 1]$. It is reasonable to assume that x_2, the probability that a non-high-risk inpatient will die the next day, is zero. This set of nonlinear equations is very difficult to solve, resulting in an intractable problem, thus we use a brute force way of enumerating a large set of vectors in the solution space. The main criterion for selecting a good candidate as our approximate solution is to select those with small AE (e.g., <0.1). Another possible condition that must hold is that the probability cannot be 0 or 1 for x_i's except x_2 because no matter which state (z_1 or z_2) the patient is in, the transition probability to other states (z_1, z_2, H, D) is unlikely to be either 0 or 1. The candidate solutions obtained from MATLAB after applying the criteria are shown in Table 11.9.

Appendix B

The simulation model consists of two parts: (i) the SIR model and (ii) the hospitalization model. We implement the model in Arena using discrete-event simulation.

The SIR Model

This model describes the progression of an infectious disease among a certain population using a set of differential equations to determine the fraction of susceptible ($s(t)$), infective ($i(t)$), and recovered ($r(t)$) in the population at time t (refer to the section 'The SIR infectious disease model'). The solutions of the equations are approximated by discrete-event simulation in the model. We calculate $s(t)$, $i(t)$, and $r(t)$ and their derivatives every 1 h (simulation time) and assume those

values are constant during each 1 h (simulation time). The compartment in Figure 11.5 is the realization of this model in Arena. The output of this model is the expected number of new cases in a certain time period (we assume 12 h), which connects the SIR model with the hospitalization model.

The hospitalization model

After we calculate the expected number of new cases in 12 h, we can generate patients assuming they arrive according to a Poisson process with interarrival time equal to 12h divided by the expected number of new cases in 12 h. Next, these patients are assigned with an age-group with probabilities shown in Table 11.5 and a risk group with probabilities α_u's described at the beginning of the subsection 'Age and risk groups distributions of admitted patients'. Then under the assumption that the percentage of clinical cases in different age and risk groups are the same, we eliminate those non-clinical cases by chance. Finally, only a small portion of the clinical cases are admitted into the hospital according to the probabilities, which are age and risk group dependent, listed in Table 11.7. So far, we have generated the patients who will be admitted into the hospital. Figure 11.6 shows the generation of inpatients.

At the beginning of each period, except period 0, the pharmacy will make decisions of producing finished goods and purchasing raw materials. At the beginning of period 0, the pharmacy does not make any decisions and the patients admitted into the hospital during period 0 consume the initial inventory, which is

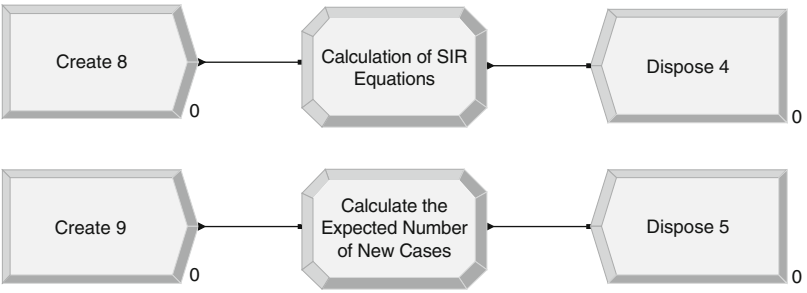


Figure 11.5 The SIR model in Arena

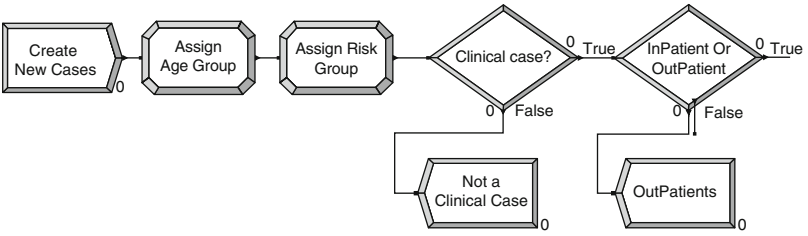


Figure 11.6 The process of generating inpatients

predetermined (see Column 2 of Table 11.17), and expedited if needed. Figure 11.7 shows the simulation module of the pharmacy. The patients admitted in period 0 do not need to wait for this module to execute and will go directly to the module described next.

During each period, the patients are allocated the medication they need by going through the module in Figure 11.8 and are held until the beginning of the next period. The release of the patients is triggered by the module shown in Figure 11.9. It sends a signal to the hold module in Figure 11.8 (The Patients Hospitalized module) at the end of each period. After the patients are released, they will go through the status update module where their health statuses change.

At the end of each period, patients' statuses are updated by going through the status update module shown in Figure 11.10. It separates the patients into different age-groups and the patients in each age-group will go into sub-models 1, 2, or 3. After their health statuses are updated, some of them will leave the system in the sub-models if their status is dead (D) or healthy (H). Then the patients who stay in the hospital wait for a short period of time (we set it to be 0.002 simulation hour) in order for the pharmacy to make its decisions and to finish production and ordering. After the wait, the simulation goes to the beginning of the next period and the patients will go back to the medication allocation process described in Figure 11.8. This completes the simulation in one period. The detail of sub-model 1 is shown in Figure 11.11 and the other two are similar to it.

For Policies I through III, changes are made to variables in the pharmacy decision-making module, which is shown in Figure 11.7. In particular for Policy III, one more module is added to read the historical data associated with medication demand between 'Decision Making and Cost Calculation' and 'Calculation Cost'. For the sensitivity analysis on GAR, changes are made in the 'Clinical Cases?' module in the patient generation module shown in Figure 11.6.

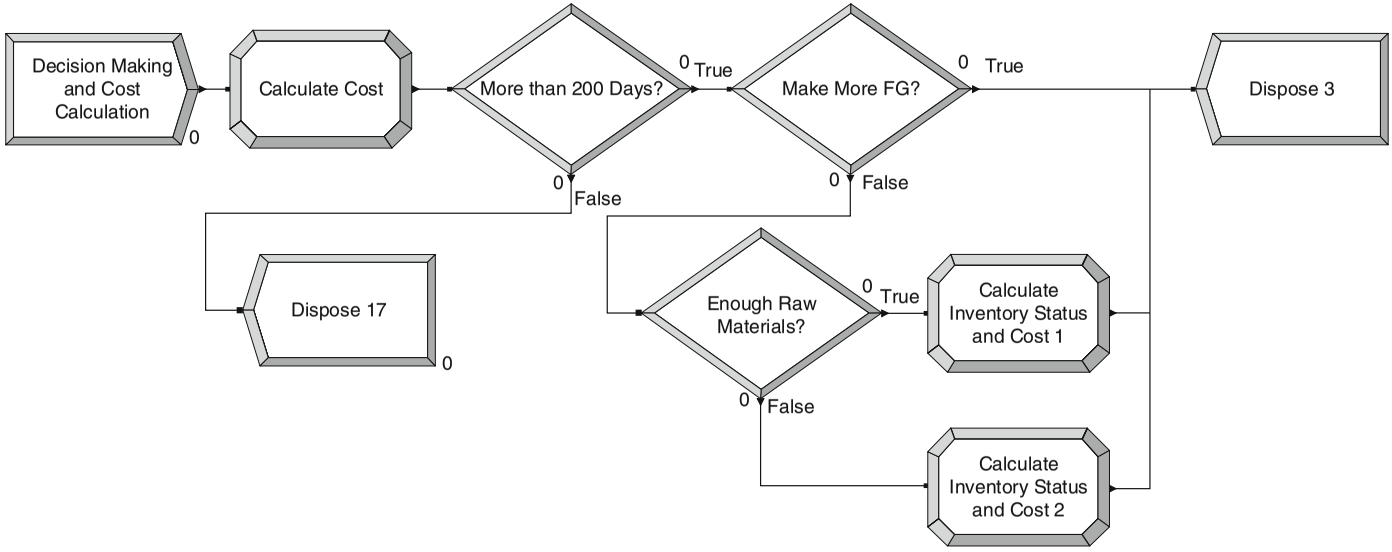


Figure 11.7 The pharmacy decision-making process

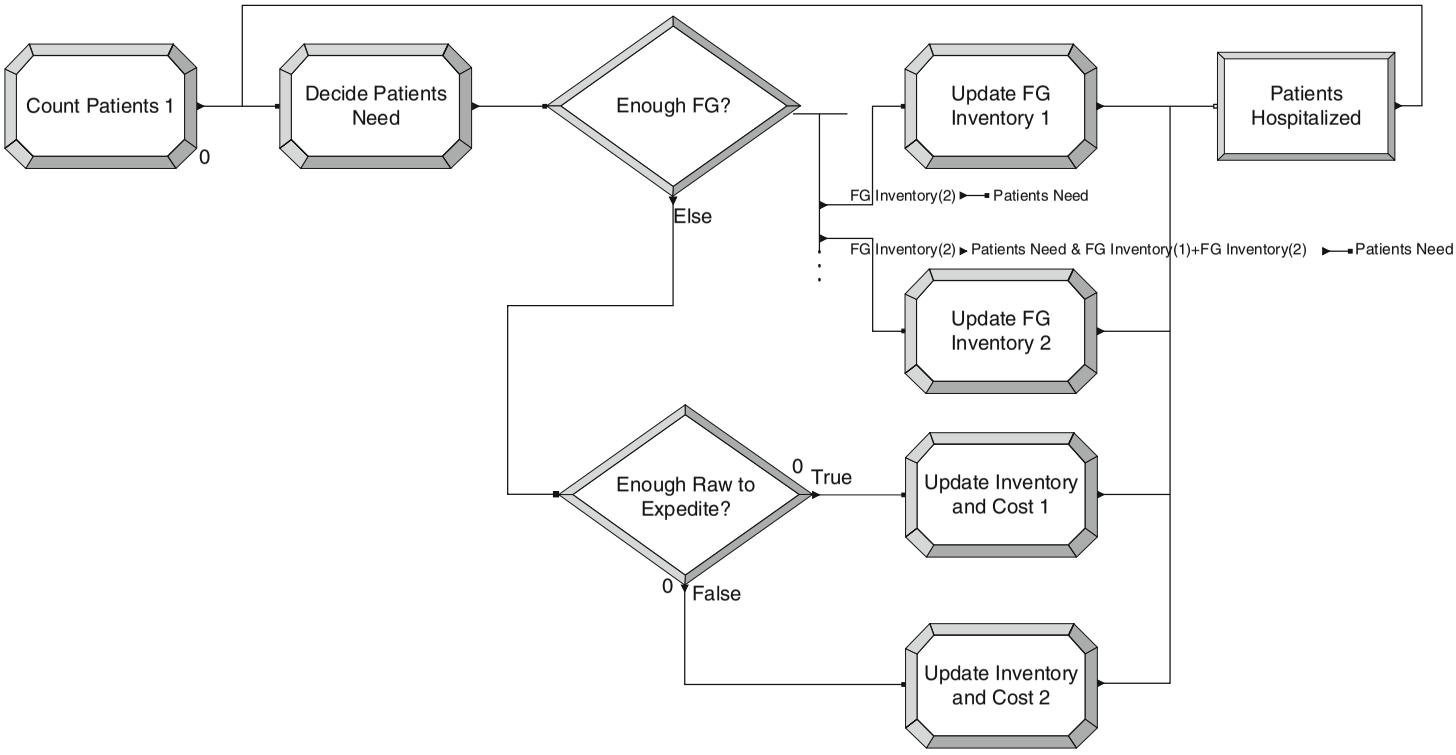


Figure 11.8 The medication allocation process

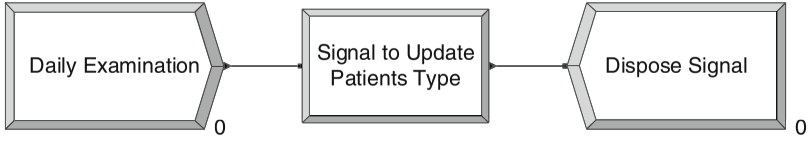


Figure 11.9 The signal module to trigger the health status update of the patients at the end of each period

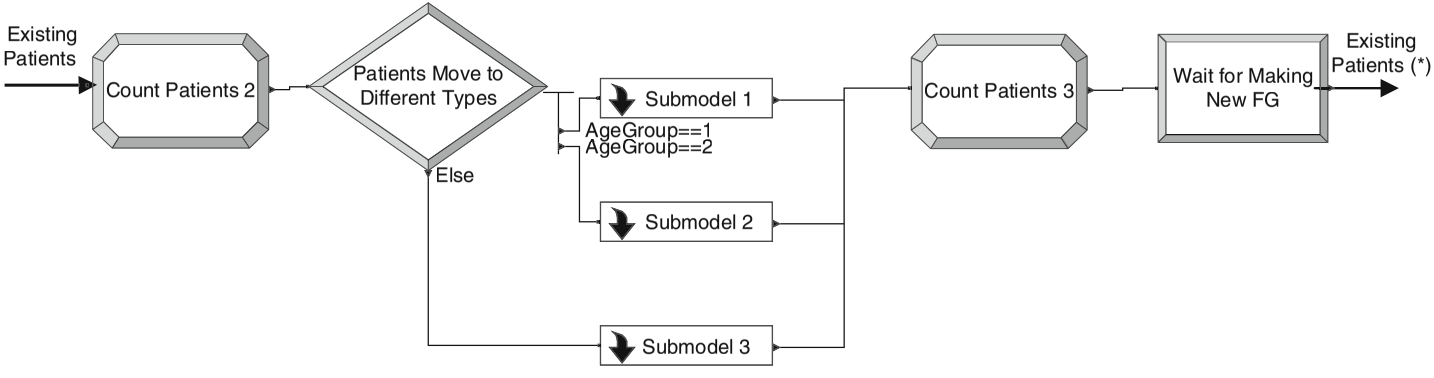


Figure 11.10 The status update module

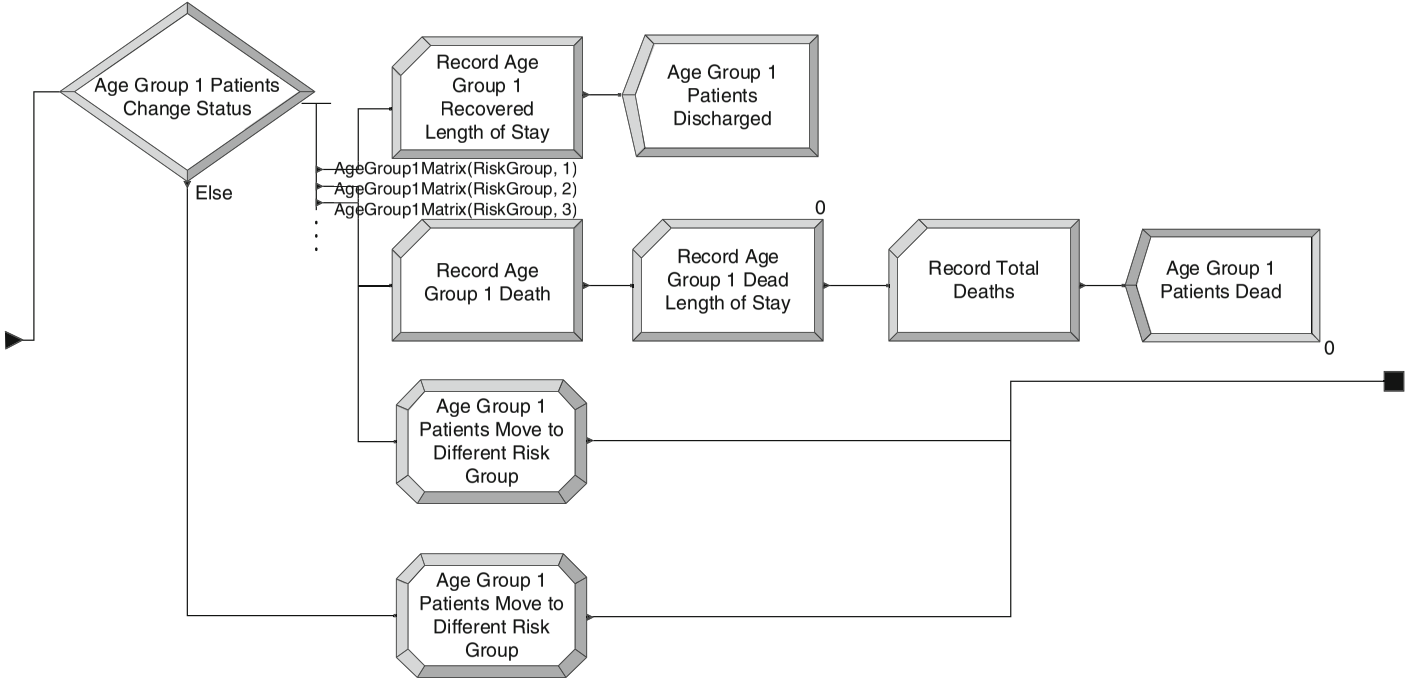


Figure 11.11 Details of sub-model 1, which updates the health status of patients in age-group 1

12

Perishable Inventory Management System with a Minimum Volume Constraint

Z. Shen, M. Dessouky and F. Ordonez
University of Southern California, Los Angeles, USA

The federal government maintains large quantities of medical supplies in stock as part of its Strategic National Stockpile (SNS) to protect the American public in case of a public health emergency. Managing these large perishable inventories effectively can help reduce the cost of the SNS and improves national security. In this paper, we propose a modified Economic Manufacturing Quantity (EMQ) model for perishable inventory with a minimum volume constraint, which is applicable to managing the inventory of medicines for the Strategic National Stockpile. We demonstrate that minimizing the cost of maintaining such a system can be formulated as a non-convex non-smooth unconstrained optimization problem. The property of this model is analysed and an efficient exact algorithm is presented to solve this problem. In the numerical experiment part, we perform sensitivity analysis on several government-controlled system parameters to illustrate how the government can obtain lower costs or a larger stockpile at the same cost by allowing more freedom in the management of the stocks.

12.1 Introduction

In most inventory systems, it is assumed that stock items can be stored indefinitely to meet future demands. However, the effects of perishability cannot be ignored for certain types of inventories, which may become partially or entirely unsuitable for consumption as time passes. Typical examples are fresh produce, blood cells, chemicals, photographic films,

Reprinted from *Journal of the Operational Research Society*, 62(12): 2063–2082, 2011, 'Perishable Inventory Management System with a Minimum Volume Constraint', by Z. Shen, M. Dessouky and F. Ordonez. With kind permission from Operational Research Society Ltd. All rights reserved.

drugs and other pharmaceuticals. In this work, we investigate a perishable inventory management system with a constant market demand rate in a production environment with a minimum inventory volume ($I_{\min}$) requirement that must be kept at all times.

This work is motivated by the supply chain management for a large-scale emergency response. As part of the national emergency preparedness plan, the US federal government maintains a Strategic National Stockpile (SNS). For example, in a potential anthrax attack, the stockpile contains enough medicine to treat 10 million people. This stockpile represents enough Cipro, a common antibiotic with a 9 year lifespan that works against anthrax and other infections, to meet regular market demand for several years. The federal government pays pharmaceutical companies to produce and store these large inventories, keeping them ready for use at a moment's notice in case of an emergency. About 80% of the SNS is in this form, which is called Vendor Managed Inventories (VMIs).

Currently, the SNS policy allows manufacturers to sell the pills at a predefined date prior to expiration rather than let the drugs spoil; however, considering the fact that the size of the stockpile is huge compared with the regular market demand while the drugs are so close to their expiration date, the potential salvage value is low. From the manufacturer's perspective, if it can apply a more sophisticated inventory holding policy, which allows the constant usage of the stockpile to meet the regular market demand, and at the same time refill with new production to maintain the minimum stockpile requirement; then, the firm can save on the total cost in maintaining the stockpile inventory, hence, making it possible to further reduce the price charged to the government. From the government's perspective, if it allows firms to sell the pills earlier, there is an opportunity to capture a significant amount of salvage value for the unsold stockpile. The challenge of this problem lies in efficiently maintaining a minimum level of perishable inventory.

In this work, our goal is to propose an inventory model from the manufacturer's perspective with the following three key parameters in mind that are given as the government-controlled system parameters: (1) how often the stockpile is refreshed and released to the open market, (2) what is a suitable cost effective minimum inventory requirement, and (3) how much should the government pay to the manufacturer for each pill stored in the stock pile. These three parameters play key roles in determining the cost and revenue of the manufacturer. Through sensitivity analysis, we demonstrate how decision makers from the government's perspective can use the proposed model to set policy (determine

the value of the above three parameters), and illustrate the possibility of reducing the cost to the government for the same level of VMI by leveraging the regular market demand.

In the SNS package, there are different medical supplies used for different emergency scenarios; for some single type of medical supplies (like antibiotics), there might be multiple suppliers. In this paper, we limit our study to a single manufacturer environment; the models for multiple manufacturers competing for contracts or how to model the strategy for the manufacturer–government interaction and negotiation are out of the scope of this study. In this work, we are concerned with a single fixed–life perishable inventory with large minimum stock requirement. Cipro is used as one example in our numerical experiments, but our model is general and could be applied to other inventory systems with similar characteristics. In such an inventory system, we need to satisfy two types of demand: the regular market demand and the minimum inventory requirement for emergency preparedness; and we aim to minimize the operational cost from the manufacturer's perspective.

Such a large perishable inventory can be maintained by setting a fixed constant production rate to replenish the stock while using this inventory to satisfy the regular demand. This type of model, in the spirit of a just in time production shop, assumes that it can be optimal to produce at less than the maximum production rate. For the majority of this paper however, we focus on extending the traditional Economic Manufacturing Quantity (EMQ) lot sizing model. In the EMQ model, it is assumed to be optimal to operate the machine at the maximum allowable production rate when making the product since there is value to the idle time. The company can use the idle time for either machine maintenance, the production of other products, or to reduce the setup costs. For a single item perishable inventory system, if there is no minimum stock requirement or the constant amount to be kept in the system is relatively small, we can readily extend the traditional EMQ model to address the perishability property of the stock by properly upper-bounding the EMQ cycle, to guarantee that the entire inventory can be completely refreshed before it expires. However, when the minimum inventory is comparable with the total regular market consumption during the shelf-life, trivial extensions to the perishable inventory policy are no longer adequate. It is therefore imperative to develop a new inventory policy specially geared to a system with a high minimum volume requirement on top of the regular market demand, such as the perishable VMI system for the SNS, to minimize the operational cost of maintaining such a system.

This paper is organized as follows: we first review the relevant literature in Section 2. In Section 3, we propose the mathematical models. We first discuss the assumptions and policies we adopt. The first model (Section 3.1) assumes a continuous constant production rate. We then study a simple variant of the EMQ model, with perishable items and zero or a 'small' minimum inventory requirement in Section 3.2. Next, we propose a modified EMQ model for the more complex scenario, eg, on perishable items with a 'high' minimum inventory constraint in Section 3.3. One of the key differences between the scenarios in Sections 3.2 and 3.3 is whether the entire inventory can be consumed by the regular market before it expires. We also show that in order to guarantee the repetitive cycle of the stockpile in the modified EMQ model in Section 3.3, it is necessary to constrain the maximum inventory cycle, a key parameter, which is derived from another government-controlled parameter in the proposed system. In Section 4, we present the detailed calculation on the total cost and boundary conditions.

For this we decompose the cost of such a system into four components: inventory holding costs, fixed setup costs, manufacturing costs and salvage costs. We can express the four parts of the total cost as a non-convex and non-smooth function of Q. Section 5 covers the exact solution approach and its complexity analysis. We conduct two different numerical experiments on an anthrax attack example in Section 6. The first experiment demonstrates the advantage of our proposed model over a standard model, which runs two separate systems to meet the regular market demand and the minimum inventory requirement respectively, and a constant production model. In the second experiment, we perform sensitivity analysis on those government controlled system parameters to provide some insights for both parties (the firm and the government) in how to negotiate contract terms. Finally, we present some concluding remarks in Section 7.

12.2 Literature review

In the existing perishable inventory management literature, policies in four different aspects have attracted attention from the research community.

- *Ordering Policy* focuses on *when* and *how much* to order; a well-known review is from Nahmias (1982).
- *Issuing Policy* concerns the sequence in which items are removed from a stockpile of finitely many units of varying ages; the most

general approach is first-in-first-out (FIFO) and it is proved to be optimal for perishable goods with random supply and demand and fixed life-time under several possible objective functions by Pierskalla and Roach (1972).

- *Disposal Policy* is applied when the strategic disposal of part of the inventory is desirable, such as slow moving stock; the topic on when and how much to dispose under stochastic demand and perishing has been studied by Rosenfield (1989, 1992).
- *Pricing Policy* is closely coupled with the ordering policy in the multi-period newsvendor problem. The price is a decision variable and the forecasted demand is price-sensitive. The pricing and ordering quantity decision can be made either sequentially or simultaneously (Gallego and van Ryzin, 1994; Abad, 1996; Burnetas and Smith, 2000; Chun, 2003).

For the perishable inventory system, the *ordering policy* is the most researched policy of the four above. Research has been done assuming: fixed or continuously deteriorating lifetime; periodic or continuous review; different distributions of the demand process; lost sale or back-order, etc. Based on different sets of assumptions, various modelling method and solution approaches have been applied.

1. *Zero first-order derivative point (stationary point) over total cost function:* It is usually straightforward to write the governing equation on the inventory level over time and then obtain the inventory carrying cost, along with the fixed ordering cost, the purchasing cost, the shortage cost, and the salvage cost. This can be modelled as an unconstrained nonlinear system. If the total relevant cost function is continuous and second-order differentiable over the decision variable of the current system (in most cases, the ordering quantity or reorder level), then we can obtain the first-order stationary point as the optimum (Ravichandran, 1995; Giri and Chaudhuri, 1998; Liu and Lian, 1999).
2. *Heuristics/Approximations:* For stochastic demand circumstances, exact optimal policies are not only difficult to compute, but also demanding to implement due to the requirement to keep track of the age distribution of the stock. Nahmias (1982) provided a good summary on the early works of approximated optimal policies and heuristics to obtain the optimal policy parameters. Nandakumar and Morton (1993) developed heuristics from 'near myopic' bounds and demonstrated the accuracy with less than 0.1% average error

over a wide range of problems. Goh *et al* (1993) applied different approximation methods to study a two-stage perishable inventory model.

3. *Markovian model:* The queuing model with impatient customers has been used to analogize perishable inventory systems. The queue corresponds to the inventory stockpile, service process to the demand, arrival of customers to the replenishment of inventory, and the time a customer will stay in queue before leaving due to impatience corresponds to the shelf-life. Early works (Chazan and Gal, 1977; Graves, 1982) usually wrote the descriptive transition probabilities then obtained the stationary distribution to evaluate the performance measures such as the expected out-dating amount. Weiss (1980) and Liu and Lian (1999) also use the performance measures to construct a cost function from which the optimal policy parameter can be computed.
4. *Dynamic programming (DP):* Since the dynamic economic lot size problem was first proposed by Wagner and Whitin (1958), which reviews the inventory periodically and the demand is deterministic in every review period, dynamic programming techniques have been widely adopted in solving many variations and extensions in the non-perishable inventory context. Hsu (2000) provided a brief review on this topic. There is a well-known zero-inventory property under which no inventory is carried into a production period that is necessary for an optimal solution. Hsu (2000) demonstrated that this zero-inventory property may not hold for any optimal solution with perishable items and proposed a new DP recursion based on an interval division property that solved the problem in polynomial time. Further extension to perishable systems that allow backorder and coexisted stochastic and deterministic demand can be found in Hsu and Lowe (2001) and Sobel and Zhang (2001).
5. *Fuzzy theory:* Recently, Katagiri and Ishii (2002) introduced fuzzy set theory in a perishable inventory control model with a fuzzy shortage cost and a fuzzy salvage cost; hence, the expected profit function is represented with a fuzzy set. The effect of the fuzziness on the obtained ordering quantity is investigated.

There is abundant literature in perishable inventory management to model and solve different types of real-life problems. To the best of our knowledge, there is no prior work with a minimum volume constraint on the inventory size throughout the planning horizon. This extension is trivial when the required minimum volume is not significant compared

with the amount consumed by a regular market demand rate within the shelf life as it can be timely and completely refreshed by the regular demand. However, when the minimum volume is huge—comparable with the total regular market demand over the shelf life—then a strategic inventory ordering and disposing policy is needed to guarantee the freshness and readiness of the required minimum inventory as well as the low cost of maintaining such an inventory system. This is the case of the medical stockpiles required for SNS in the large-scale emergency context. In this work, we address this modelling gap by formulating a perishable inventory management model with minimum inventory constraint and providing an exact solution approach to this model.

12.3 Model

We assume a single fixed-life perishable item is produced, consumed and stored for an infinite continuous time horizon. We denote T_s as the shelf-life of the product measured in terms of some given time unit—a period. There is a known regular market demand with a constant rate of D items per period. The production can start at any time at a constant maximum allowable rate of P that is greater than D; and there is a constant cost A associated with each production setup. The holding cost h, unit purchase cost w and unit salvage cost w are all time invariant. This inventory system must maintain at every point in time at least $I_{\min}$ of non-spoiled inventory and we assume that the inventory is consumed using the natural FIFO policy.

In this problem, since the items are perishable, to maintain an $I_{\min}$ amount of fresh inventory, the items must be produced regularly. In particular the entire $I_{\min}$ must be produced within its shelf life, that is

$$I_{\min} \leqslant P \cdot T_S. \tag{1}$$

Without this *stability condition* it is not possible to build an inventory of $I_{\min}$ items that have not expired.

We first consider a simple continuous production model that decides a constant production rate, and then extend the results to a classical EMQ model in the following subsections.

12.3.1 Continuous production model

In continous production models it is assumed that it can be optimal to produce less than the maximum allowable production rate of P. With no setup costs, the constant production rate can be set to any value between

$\max\{D, I_{min}/T_s\}$ and P items per period. Note that the production rate must be greater than D to satisfy the regular demand and greater than $I_{min/}T_s$ so that the perishable inventory can be built in T_s periods, like in the stability condition (1). Note also that (1) implies that $I_{min}/T_s \leqslant P$.

If the production rate is set to the lower bound and $D \geqslant I_{min}/T_s$ then the oldest D units in I_{min} are used to meet the regular market demand and are replaced by the newly produced items, keeping the inventory constant. If, however, $D < I_{min}/T_s$ then the oldest items in I_{min} were produced exactly T_s–1 periods ago and are about to expire. We should use D of these oldest items to meet the regular demand in the period and dispose of the remaining I_{min}/T_s–D oldest items, again keeping the inventory constant. In general, a continous production model that sets the production rate at a value $v \in [\max\{D, I_{min}/T_s\}, P]$ and uses the oldest items in inventory to satisfy the regular demand discarding any excess, would maintain a constant inventory which is greater than or equal to I_{min} made up of $\lceil I_{min}/v \rceil$ groups of v items produced 0, 1, 2, ..., $\lceil I_{min}/v \rceil - 1$ periods ago, respectively. Each period, v items are produced and D of the last v items meet the regular demand, with the remaining v–D items (of age $\lceil I_{min}/v \rceil$) being discarded.

Note that this model may not set the machine production rate at its maximum speed of P and neglects the consideration of setup costs. For the remainder of the paper, we assume it is optimal to operate the machine at the maximum allowable production rate when making the product since there is value to the idle time. The company can use the idle time for either machine maintenance or the production of other products. Hence, the next two models studied are based on the classical EMQ inventory model.

12.3.2 EMQ model with perishability

For non-perishable items and no minimum inventory, the classical EMQ model provides an optimal inventory management policy for this problem. This cyclic solution orders a fixed amount, known as the production batch size, Q every $T = Q/D$ units of time. The EMQ cycle T begins with a production phase that lasts $T_1 = (Q/P) = (D/P)T$ and is followed by an idle phase lasting $T_2 = ((P–D)/P)T$. The optimal batch size of the EMQ model EMQ^* is identified by minimizing a convex inventory cost function.

The proposition below shows that when the minimum inventory requirement is small ($I_{min} \leqslant DT_s$) the EMQ model can be directly applied to the perishable stockpile with a minimum inventory requirement by adding a constraint to ensure the age of the inventory does not reach expiration. Before we prove the general case, consider a perishable

inventory system with $I_{\min} = 0$. The idea is that, if the inventory is consumed following a FIFO policy, the items consumed by time T_1 are produced up to time $T'' = (D/P)T_1$. We obtain an upper bound on the age of the inventory assuming that the next item produced is kept in inventory for the duration of the cycle, which gives an age of $T–T'' = ((P^2–D^2)/P^2)T$. Ensuring that this upper bound is less than the shelf-life T_s guarantees that the EMQ policy is valid (Figure 12.1).

Proposition 1 *An EMQ model for a perishable inventory with minimum inventory requirement of* $I_{\min} \leq DT_s$ *has an optimal production batch size given by*

$$Q^* = \min\left\{EMQ^*, T_s\frac{DP}{P-D} - I_{\min}\frac{P+D}{P-D}, \left(T_s - \frac{I_{\min}}{P}\right)\frac{DP^2}{P^2-D^2}\right\}.$$

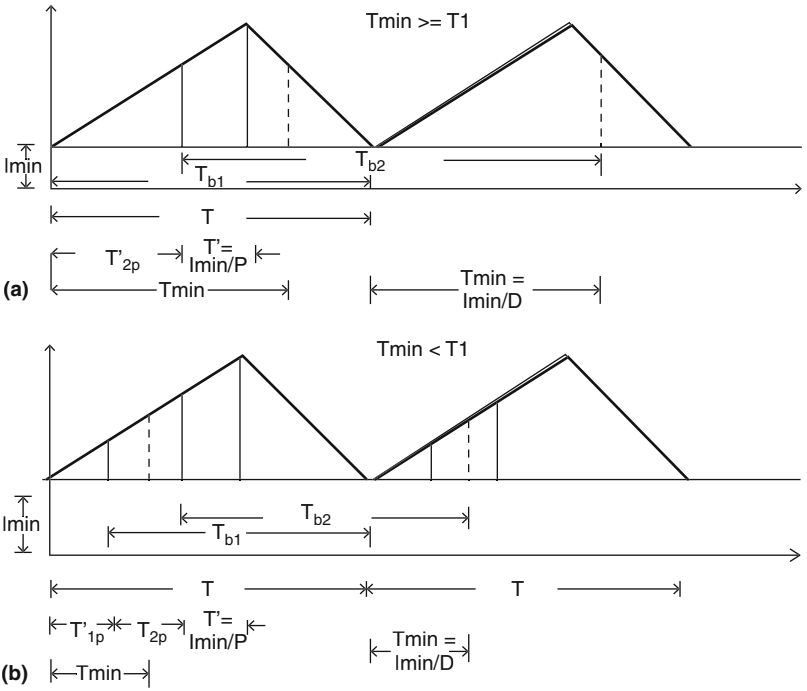


Figure 12.1 Perishable inventory system with small $I_{\min}$

The proof of Proposition 1 is in Appendix A. Proposition 1 identifies the situation where the classical EMQ model can be used to manage a perishable inventory system with minimum inventory requirement. However, when such a condition ($I_{\min} \leqslant T_s D$) cannot be met, a more sophisticated model is required, which is the focus for the rest of this paper.

12.3.3 Modified EMQ model

In this subsection, we introduce the modified EMQ model that incorporates both regular market demand and emergency demand ($I_{\min}$) for a perishable stockpile. We consider an EMQ like model because of the simplicity and wide use of these types of models.

Similar to the traditional EMQ model, we propose a periodic inventory management strategy where the amount produced each period is set and this will determine the inventory cycle length. Figure 12.2 gives an illustration of the inventory plot for the modified EMQ model we propose. We define the *inventory cycle* (T_{inv}) as the minimum length of time that an inventory pattern repeats. We assume that excess supplies will be disposed (salvaged) once at the end of each T_{inv}, so that exactly $I_{\min}$ items are present at the beginning of the next inventory cycle. Note that, the salvaged products have different ages at the time of disposal.

We propose the following ordering policy, illustrated in Figure 12.2: for any given Q, we initially run a regular EMQ cycle (which we call the 'underlying regular EMQ cycle' with cycle length $T = Q/D$) and make

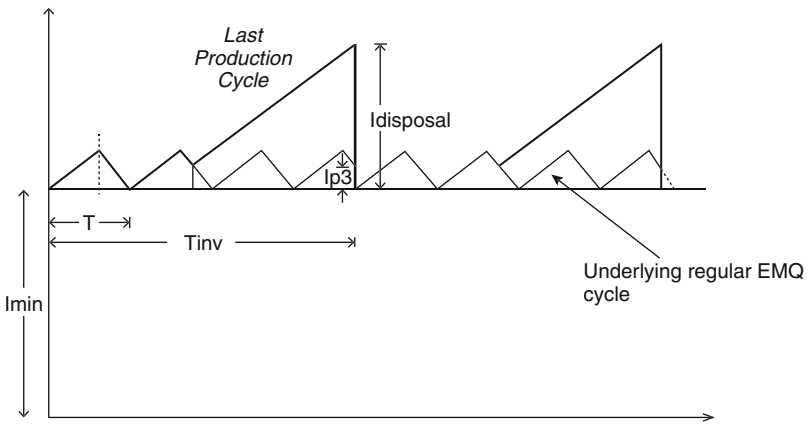


Figure 12.2 Illustration for the modified EMQ model

some adjustment near the end of the *inventory cycle* to satisfy the minimum inventory requirement. This adjustment may require a special or *last production cycle*. Since the inventory pattern repeats every T_{inv}, similar to the stability condition (1), we must ensure that an $I_{\min}$ inventory can be produced every T_{inv} periods. In addition, as seen in the previous subsection, the modified EMQ model is necessary when the $I_{\min}$ cannot be consumed by the regular market demand in T_{inv}. Therefore the modified EMQ model is necessary when

$$DT_{inv} < I_{\min} \leqslant PT_{inv}. \tag{2}$$

Given the *production batch size Q*, we show in the next section that we can determine all relevant quantities to describe this model as a function of *Q*. In particular these quantities include, the length of the underlying regular EMQ cycle, $T = Q/D$, when to initiate the *last production cycle,* and the length of this *last production cycle*. The different relations between these quantities also show that there are five different situations for this model.

Figure 12.3 presents the five possible situations of the inventory model, of which Figure 12.2 is one of them. Below we explain these cases and their classification, which is based on three criteria:

1. if *a last production cycle* is needed to replace the extra disposing part of the inventory;
2. where the *inventory cycle* ends relative to a regular underlying EMQ cycle (in the production period—the uphill region in the inventory graph, or in the non-production period—the downhill region);
3. where the *last production cycle* starts (if it starts at the same downhill region of a regular underlying EMQ cycle as it ends or not).

The disposing policy of the model requires that we dispose all the inventory above the $I_{\min}$ level at the end of each *inventory cycle* to restart a new cycle with exactly $I_{\min}$. The amount to be disposed is either the part of the minimum inventory ($I_{\min}$), which cannot be consumed by the regular market demand within an inventory cycle (which is $I_{disposal} = I_{\min} - DT_{inv}$) or the part of the inventory above the minimum inventory volume requirement generated by a regular EMQ cycle at the end of an *inventory cycle* (defined as I_{p3}), whichever is bigger. Cases 1 and 2 in Figure 12.3 depict the situation that I_{p3} is greater than $I_{disposal}$ and we dispose I_{p3} without a *last production cycle.* Cases 3, 4 and 5 fit the situation where I_{p3} is less than $I_{disposal}$ and we initiate a *last production cycle* to produce the amount $I_{disposal} - I_{p3}$ and

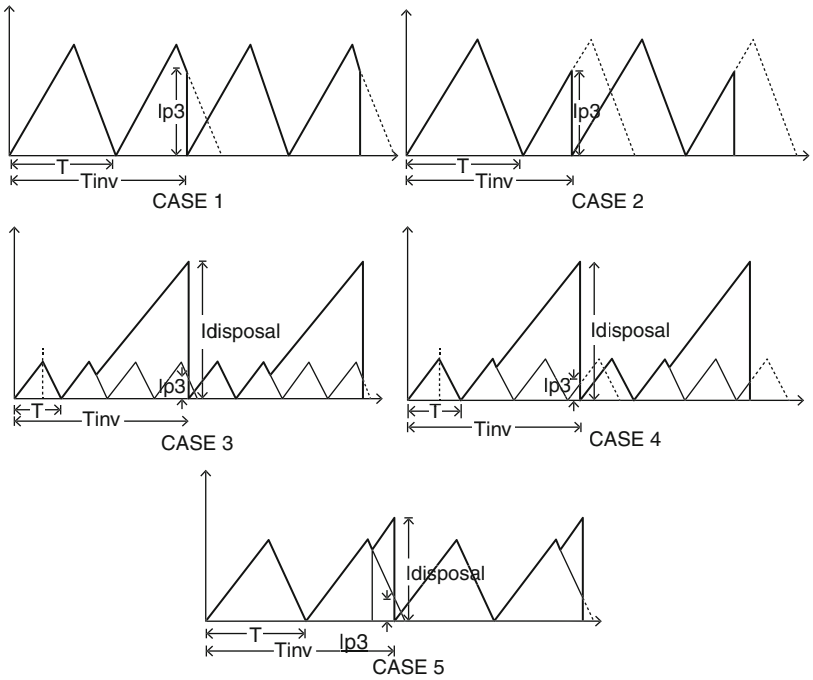


Figure 12.3 All 5 cases of possible scenarios

dispose exactly $I_{disposal}$ at the end of each T_{inv}. We further classify the cases based on where an *inventory cycle* ends relative to the underlying EMQ cycle. In cases 2 and 4, the *inventory cycle* ends at an uphill region (the production period); otherwise, in cases 1, 3 and 5, the cycle ends in a downhill region (the non-production period). Furthermore, the 3 cases which have a *last production cycle,* are further classified depending on where the last production cycle starts. If it starts at the same non-production period as it ends, it is case 5; otherwise, if it starts at some earlier regular EMQ cycle, we have cases 3 and 4. With the above three criteria, we can distinguish these five cases uniquely. These enumerate all possible scenarios.

The conditions in (2) guarantee that the minimum inventory has not expired by producing it in every inventory cycle T_{inv}. Since $I_{\min}$ has to be maintained and there is a FIFO issuing policy, the $I_{\min}$ amount produced in one inventory cycle is consumed starting in the next cycle. This observation leads to the following lemma that imposes a constraint on T_{inv} to guarantee the freshness of all items in inventory.

Lemma 1: *(Maximum Inventory Cycle Length) Setting the maximum inventory cycle length to at most half shelf-life* $T_{inv} \leq (1/2)T_s$*; guarantees that*

- *the complete inventory is always within its shelf-life;*
- $I_{\min}$ *is always younger than half the shelf-life and the salvaged items are aged between the half shelf-life and full shelf-life;*
- *the age distribution of the stockpile repeats itself every* T_{inv}.

Proof We consider the case when $I_{\min} = PT_{inv}$, which means that we must maintain a constant production at a rate P to produce $I_{\min}$ every T_{inv}, (see Figure 12.4). This modified EMQ model is the case of one single last production cycle for the entire T_{inv}. In the case of a smaller $I_{\min}$, the analysis below is valid and provides an upper bound on the age of items in inventory. Since we keep a constant production of rate of *P,* the minimum inventory $I_{\min}$ at the beginning of an inventory cycle is made up of $\lceil I_{\min}/P \rceil = T_{inv}$ groups of P items each, produced 0,1, ..., $T_{inv} - 1$ periods ago.

Each period of this inventory cyle, we produce a new set of P items and consume D of the oldest items available. To simplify the accounting, we will upperbound the age of the items kept in inventory assuming that every period we consume D items produced T_{inv}–1 periods ago, and keep the remaining P–D items in inventory until the end of the inventory cycle. Therefore, in the last period of this inventory cycle,

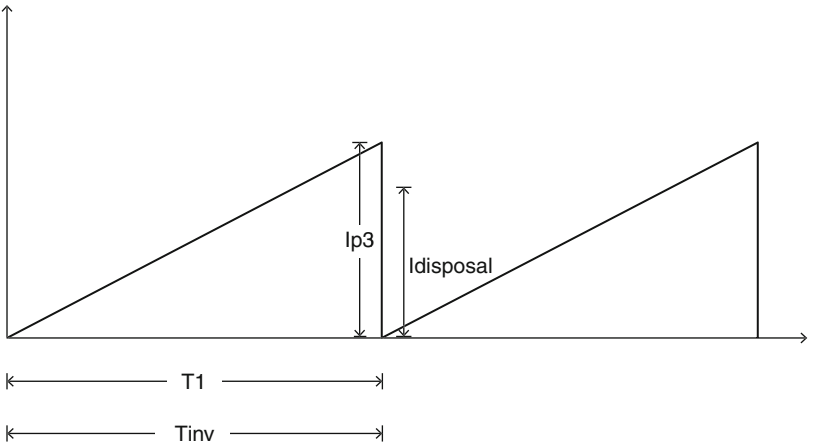


Figure 12.4 Graph illustration for an extreme boundary case

and before discarding the excess inventory, in addition to the $I_{\min}$ inventory described above, we have $T_{inv}(P–D)$ items in groups of $P–D$ items, each produced T_{inv}, $T_{inv}+1$, ..., $2T_{inv}–1$ periods ago.

The upper bound on cycle length $2T_{inv} \leqslant T_s$ ensures that the entire inventory is within the shelf life T_s. After the oldest $T_{inv}(P–D)$ items are discarded we recover the inventory level and age distribution that existed at the start of the inventory cycle, showing that we can repeat the process every T_{inv}. The actual inventory under the FIFO issuing policy (and with a smaller $I_{\min}$) would be younger than the situation considered here, therefore $T_{inv} \leqslant (1/2)T_s$ also ensures that all the inventory is within shelf life and the process is repeatable.

12.4 Total cost evaluation and boundary conditions

In this Section, we first introduce the notation, then discuss the calculation steps required by all five cases. In all five cases, the total cost depends continuously only with respect to Q.

12.4.1 Notation

We first introduce the notation used in the calculation. We will continue using the parameters we defined previously, and the variables below are notation specially used in our proposed model.

$I_{\max}$	the maximum inventory level in a regular EMQ cycle
$I_{disposal}$	the minimum amount to be disposed every T_{inv}
N	number of complete regular EMQ cycles in a T_{inv}
T_{p3}	the remainder of T_{inv} divided by T
I_{P3}	the inventory level of a regular underlying EMQ at the end of a T_{inv}
T_{disposal}	production time for the extra inventory to be disposed, which is max $(I_{disposal}–I_{p3}, 0)$
δ	the non-production time in T_{p3}
T_{p1}	from the start of the *last production cycle* to the end of the current underlying EMQ cycle
I_{p1}	the inventory level at the beginning of the *last production cycle*
N_1	number of complete regular EMQ cycles within the *last production cycle*
M	number of regular EMQ orders in a T_{inv}

12.4.2 Cost decomposition

Now we are ready to calculate the inventory level for the different cases to prepare the total cost computation. We use the same T, T_1 and T_2

formula as defined in Section 3.2. Below are the basic quantities which share the same formula across all five cases.

$$I_{\max} = Q\left(1 - \frac{D}{P}\right) \tag{3}$$

$$I_{disposal} = I_{\min} - T_{inv} \cdot D \tag{4}$$

$$N = \left\lfloor \frac{T_{inv}}{T} \right\rfloor \tag{5}$$

$$T_{p3} = T_{inv} \% T = T_{inv} - N \cdot T \tag{6}$$

$$I_{p3} = \begin{cases} D \cdot (T - T_{p3}) & \text{for cases 1,3,5; where } T_{inv} \text{ ends} \\ & \text{during non – production period;} \\ (P - D) \cdot T_{p3} & \text{for cases 2,4; where } T_{inv} \\ & \text{ends during production period.} \end{cases} \tag{7}$$

For cases 3, 4 or 5, we have the formula for the non-production time in the last incomplete EMQ cycle (T_{p3})

$$\delta = \max(T_{p3} - T_1, 0) \tag{8}$$

As we assume that at the end of each *inventory cycle,* the inventory level is set back to $I_{\min}$. At T_{inv}, the inventory level I_{p3}, is less than the required disposal amount $I_{disposal}$ for cases 3, 4 and 5. The time that is needed to produce the extra disposal amount ($I_{disposal}$–I_{p3}) is

$$T_{disposal} = \frac{I_{disposal} - I_{p3}}{P} \tag{9}$$

Note that ($I_{disposal}$–I_{p3}) is just part of the amount produced in the last production cycle, which cannot be covered by the production periods in a regular underlying EMQ cycle within an *inventory cycle*. Hence, $T_{disposal}$ only occupies part of the last production cycle with the remaining time used to produce items to meet the regular market demand. Another way to express the classification of cases 3, 4 and 5 by δ and $T_{disposal}$ is: if $\delta = 0$, it is case 4; else when $\delta > 0$, if $\delta > T_{disposal}$, it goes to case 5 and otherwise to case 3.

We can decompose the total cost TC of maintaining this perishable inventory system within a single *inventory cycle* into four parts: inventory holding cost (TC_{inv}), fixed ordering cost (TC_N), unit purchase cost ($TC_{purchase}$) and salvage cost on the disposal part ($TC_{salvage}$). That is

$$TC = TC_{inv} + TC_N + TC_{purchase} + TC_{salvage} \tag{10}$$

We first look at the computation on the purchase cost and salvage cost. The total purchase cost is the amount to produce within an *inventory cycle* times the unit price. The total salvage cost is the amount to dispose at the end of each *inventory cycle* times the unit salvage value. Hence, we have the following formulas:

For cases 1 and 2:

$$TC_{purchase} = (I_{\min} + I_{p3} - I_{disposal}) \cdot v \tag{11}$$

$$TC_{Salvage} = I_{p3} \cdot w \tag{12}$$

For cases 3, 4 and 5:

$$TC_{Purchase} = I_{\min} \cdot v \tag{13}$$

$$TC_{Salvage} = I_{disposal} \cdot w \tag{14}$$

Note that for cases 3, 4 and 5, $TC_{Purchase}$ and $TC_{Salvage}$ are fixed and independent of the *production batch size* (Q); hence, they can be removed from the total relevant cost calculation.

Next, we look at the computation on the total ordering cost and inventory holding cost, which are more complicated. We first give the general calculation formula here and then expand them with respect to Q for each case later.

For cases 1 and 2, since there is no *last production cycle,* the number of orders is the number of complete regular EMQ cycles in an *inventory cycle* plus 1. And the total inventory carrying cost can be calculated by the area under the inventory plot. For case 1 where the T_{inv}, ends in the downhill region, the area under the inventory plot would be $N + 1$ regular EMQ triangles minus the cut-off small triangle in the shadow, in Figure 12.5. For case 2 where the T_{inv}, ends in the uphill region, the

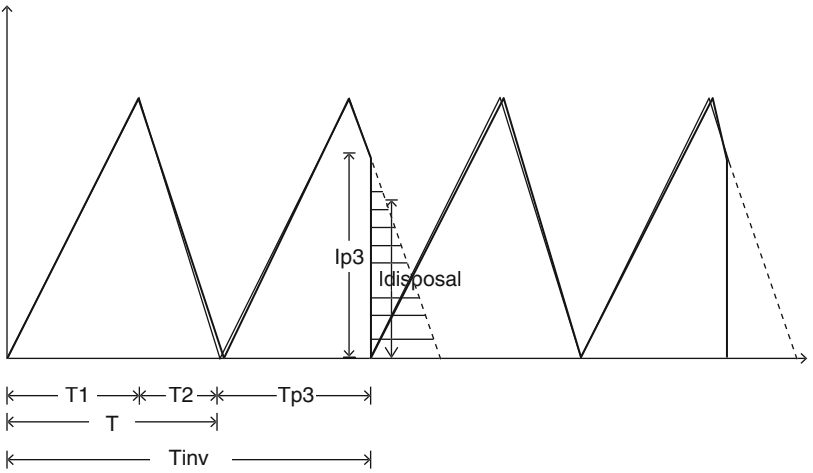


Figure 12.5 Graph illustration for case 1

area would be N regular EMQ triangles plus the small extra triangle in the shadow, in Figure 12.Bl. The formula is as follows:

$$TC_N = (N+1)\cdot A \tag{15}$$

$$TC_{Inv} = \begin{cases} \frac{1}{2}(N+1)\cdot T\cdot I_{\max} - \frac{1}{2}(T-T_{p3})\cdot I_{p3} & \text{for case 1;} \\ \frac{1}{2}N\cdot T\cdot I_{\max} + \frac{1}{2}T_{p3}\cdot I_{p3} & \text{for case 2.} \end{cases} \tag{16}$$

For cases 3 and 4, we can only use the non-production time (T_2) of regular EMQ cycles to produce the $I_{disposal}$–I_{p3} amount (production time of the EMQ cycle is already in use to satisfy the regular demand D). The number of complete EMQ cycles that would be covered by the last production cycle is

$$N_1 = \left\lfloor \frac{T_{disposal} - \delta}{T_2} \right\rfloor \tag{17}$$

If M is the number of complete regular EMQ cycles before the last production period starts, then there are $M + 1$ orders per inventory cycle.

$$M = N - N_1 \tag{18}$$

$$TC_N = (M+1)\cdot A \tag{19}$$

In cases 3 and 4 (see Figure 12.C1 and Figure 12.B2) the last production cycle must start on a non-production period (T_2), and the time of this last production cycle during the current T_2 is

$$\begin{aligned} T_{p1} &= (T_{disposal} - \delta)\%T_2 \\ &= (T_{disposal} - \delta) - N_1 \cdot T_2 \end{aligned} \tag{20}$$

Hence, we have the height of the short parallel lateral (E_1E_2) of the trapezoid ($E_1E_2E_3E_4$), which is

$$I_{p1} = T_{p1} \cdot D \tag{21}$$

The total area under the inventory plot for both cases 3 and 4 is M regular EMQ triangles minus the cut-off small shaded triangle plus the area of the trapezoid. The long parallel lateral is exactly the amount to be disposed (which is the length of E_3E_4 that equals to $I_{disposal}$). Hence, the total inventory carrying cost can be expressed as

$$\begin{aligned} TC_{Inv} = {} & \frac{1}{2} M \cdot T \cdot I_{\max} - \frac{1}{2} T_{p1} \cdot I_{p1} \\ & + \frac{1}{2}(I_{p1} + I_{disposal}) \\ & \cdot (T_{inv} - (M \cdot T - T_{p1})) \end{aligned} \tag{22}$$

For case 5 (see Figure 12.B3), the last production cycle starts at the same downhill slope as the T_{inv} ends, and the area under the inventory plot would be similar to the calculation of cases 3 and 4 except for the width of the cutoff small triangle in the shadow and the width of the trapezoid:

$$N_1 = 0 \tag{23}$$

$$I_{p1} = (T - T_{p3} + T_{disposal}) \cdot D \tag{24}$$

$$TC_N = (N+2) \cdot A = (M+1) \cdot A \tag{25}$$

$$TC_{Inv} = \frac{1}{2}(N+1)\cdot T\cdot I_{\max} - \frac{1}{2}(T - T_{p3} + T_{disposal})\cdot I_{p1} + \frac{1}{2}(I_{p1} + I_{disposal})\cdot T_{disposal} \quad (26)$$

12.4.3 Total cost and boundary condition of case 1

Next, based on the results of the previous subsection, we present the detailed calculation on the total cost as a function with respect to Q and the boundary conditions for case 1. This is a relatively simple case. For a more complicated example, please see Appendix C that shows the calculation on case 3. The other cases can be found in Shen (2008).

$$\begin{aligned} TRC(Q) &= I_{p3}\cdot(v+w) + \frac{1}{2}h[(N+1)\cdot T\cdot I_{\max} \\ &\quad -(T-T_{p3})\cdot D(T-T_{p3})] \\ &= \left(\frac{Q}{D} - T_{inv} + N\frac{Q}{D}\right)\cdot D\cdot(v+w) \\ &\quad + \frac{1}{2}h\left[(N+1)\frac{Q^2}{D}\left(1-\frac{D}{P}\right)\right. \\ &\quad \left. - D\cdot\left(\frac{Q}{D} - T_{inv} + N\frac{Q}{D}\right)^2\right] \end{aligned}$$

We let $TRC(Q) = a\cdot Q^2 + b\cdot Q + c$, then

$$\begin{aligned} a &= \frac{1}{2}h\cdot\left[(N+1)\cdot\left(\frac{1}{D}-\frac{1}{P}\right) - D\cdot\frac{(N+1)^2}{D^2}\right] \\ &= -\frac{h}{2}\frac{(N+1)}{P}\cdot\left(1+N\cdot\frac{P}{D}\right) \\ b &= (N+1)\cdot(v+w) + h\cdot T_{inv}\cdot(N+1) \\ &= (N+1)\cdot[(v+w) + h\cdot T_{inv}] \end{aligned}$$

The $TRC(Q)$ is a quadratic function of Q and the stationary point of the quadratic curve is (where $a < 0$):

$$Q^*_{case\,1} = -\frac{b}{2a} = \frac{P[(v+w) + h\cdot T_{inv}]}{h\left(1+N\frac{P}{D}\right)}$$

The boundary points for case 1 on one side is that T_{inv} ends at the downhill region when I_{p3} is exactly the amount of $I_{disposal}$, which is $I_{\min}$–$T_{inv}D$. That is:

$$I_{\min} - T_{inv}D = D\left[(N+1)\frac{Q}{D} - T_{inv}\right];$$

$$\text{hence: } Q_{lb} = \frac{I_{\min}}{N+1}.$$

The other boundary point lies when T_{inv} ends at the point where $T_{p3} = T_1$. That is

$$T_{inv} = N\frac{Q}{D} + \frac{Q}{P};$$

$$\text{hence: } Q_{ub} = \frac{T_{inv}}{\frac{N}{D} + \frac{1}{P}}.$$

12.5 Solution approach

12.5.1 Local and global optimality

With the total cost calculation formula in Section 4 for the different cases, it is straight forward to plot the total relevant cost with respect to Q (see Figure 12.6). The x-axis is the Q value and the y-axis is the total relevant cost. The below smooth plot represents the total relevant cost with respect to Q for a regular EMQ model without the $I_{\min}$ constraint. It is a quadratic function over Q and the optimal Q value can be readily obtained at the stationary point (where the first order derivative is equal to zero). The top irregular plot denotes the total cost with respect to Q for our proposed model. We can see from the plot that the total cost is non-continuous and non-differentiable at those boundary points between two different cases. In this section, we propose an exact solution method to obtain the optimal Q that reaches the minimum total cost with pseudo-polynomial complexity. We first discuss the local optimality property for each segment in the total cost plot. Then we prove the property which guarantees the global optimality. Finally, we present the complete algorithm and demonstrate its complexity.

12.5.1.1 Local optimality

Theorem I: (Local Optimality) *Within each case with fixed N and M, the total cost is a quadratic function of Q and the local minimum is either at a boundary point or at the zero first order derivative point.*

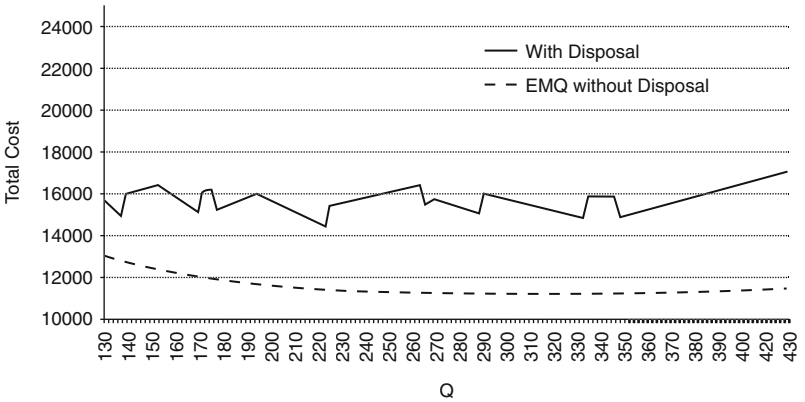


Figure 12.6 Plot of non-continuous, non-differentiable total cost function with minimum inventory

Proof The quadratic property of the total cost with fixed N and M for each case (corresponding to a segment in the plot) is evident from the calculation in Section 4. The local minimum lies at either the boundary point with smaller value or the zero first order derivative point (stationary point).

12.5.1.2 Global optimality

After we obtain the local optimality of each segment, we can compare the local minimum values for the different segments and select the lowest one as the global optimum. However, since the number of segments goes to infinity as Q decreases (or N increases), we need to show that only a limited number of segments are required to calculate the local optimal values as potential global optimality candidates.

As in the regular EMQ model, the optimal Q value occurs at a point such that the total ordering cost is at the same level as the total inventory carrying cost. These two components are balanced. In our proposed model, we can use the analogy of the optimality condition for a regular EMQ model to explore the global optimum near the region where the ordering cost and the inventory carrying cost are most balanced. It is straightforward to demonstrate that only when the N value is low, cases 1, 2 and 5 may occur, since low N leads to large Q, which also implies large regular underlying EMQ cycles that makes I_{p3} in a scale that is comparable with $I_{disposal}$. When N gets larger, cases 3 and 4 segments alternate between each other. If we prove that the total cost will monotonically increase as N increases after a threshold value $\bar{N}$, then it

implies that only a limited number of local optimum need to be computed as global minimum candidates. Next we will prove this property.

Theorem II: (Global Optimality) *When N is greater than a threshold value* $\bar{N}$*, the total cost will monotonically increase as N increases.*

Proof We first prove that the total cost of a fixed *N* value for both cases 3 and 4 is bounded; then we prove that the bound monotonically increases as *N* increases after a certain threshold value.

First, we use the graph to prove the bounds on the total cost with fixed *N*. Figure 12.7 (a) illustrates the general situation of the inventory plot with four regular EMQ orderings before the *last production cycle*. If we consider the area under the complete triangle for the *last production cycle* as fixed, then the rest area for the inventory carrying cost would be the area in shadow in Figure 12.7 (a). The shadow area is bounded by the shadow area of the three small triangles in Figure 12.7 (b) from below; and is bounded by the shadow area of the four large triangles in Figure 12.7 (c) from above. This is evident since the *last production cycle* triangle is the same for all three situations (Figure 12.7 (a), (b) and (c)) and the time $T_{inv} - T'_{disposal}$ is evenly divided by 3 and 4, respectively, in

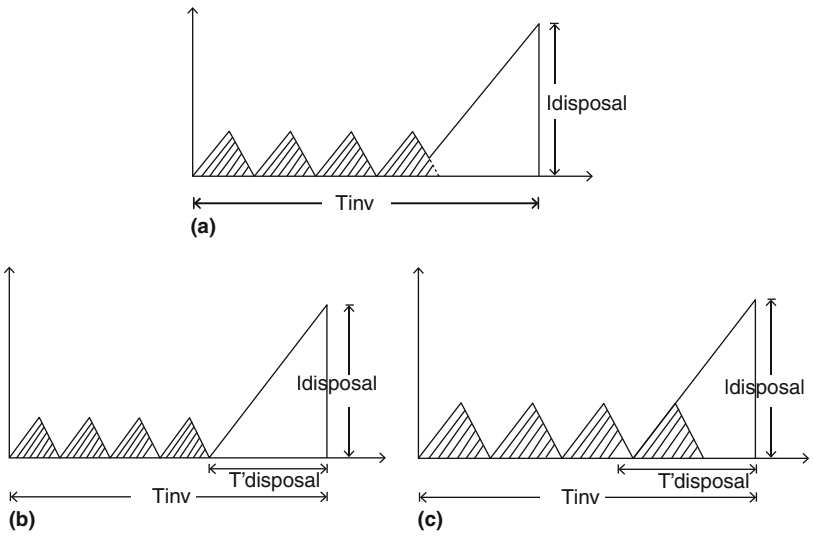


Figure 12.7 Graph illustration for the global optimum proof

Figures 12.7 (c) and (b). (Please note that $T'_{disposal}$ here represents the length of the *last production cycle,* which is different from the $T_{disposal}$ defined and used in Sections 3.3 and 4.) Thus, the area of one triangle in Figure 12.7 (a) is smaller than the one triangle area in Figure 12.7 (c), and is larger than the one triangle area in Figure 12.7 (b). Since the shadow area in Figure 12.7 (a) is between the area for 3 and 4 triangles, the upper and lower bounds on it is proved.

Next, we prove the monotonically increasing property of the bounds after a certain threshold. For both cases 3 and 4, the purchase cost and salvage cost are independent to *Q;* only ordering cost and inventory carrying cost are sensitive to *Q*. We let $T_{triangle} = T_{inv} - T'_{disposal}$ denote the time that is occupied by the regular EMQ cycle triangles. The total area for exactly N triangles within $T_{triangle}$ is $Area(N) = (1/2)DT(1-(D/P))TN$ where $T = T_{triangle}/N$, therefore $Area(N) = D/2(1-(D/P))T^2_{triangle}/N$. As N increases to $N + 1$, the total inventory carrying cost decreases the amount of $Area(N) - Area(N + 1)$ and the total ordering cost increases by A (where A is the fixed cost associated with each production setup). Hence, it is proved that the total cost at the boundary points (where $T_{triangle}$ can be exactly divided into an integer number of EMQ cycles) is monotonically increasing beyond $\bar{N}$ where $\bar{N}$ is the smallest integer value satisfied by $Area(N) - Area(N + 1) \leqslant A$.

12.5.2 Exact solution algorithm and complexity analysis

Guaranteed by *Theorem I* and *Theorem II,* we have the complete algorithm to reach the global optimum for the modified EMQ model we proposed in Section 3.3 with pseudo-polynomial complexity.

Theorem III: (Complexity) *The exact solution algorithm at most needs to explore* $5\bar{N}$ *local minimum points to reach the global minimum, where* $\bar{N} = \left\lfloor \sqrt{D/2(1-D/P)T^2_{triangle}/A} \right\rfloor$.

Proof From the proof in *Theorem II,* we have $Area(N) = (D/2)(1-(D/P))(T^2_{triangle})/(N)$, and the monotonically increasing threshold is given by $\bar{N}$ where $\bar{N}$ is the smallest integer value satisfied by $Area(N) - Area(N + 1)$. Substituting the expression for $Area(N)$ we have that $A \geqslant (D/2)(1-D/P)T^2_{triangle}((1/N)-(1/N+1))$. This implies $N+1 \geq \sqrt{D/2(1-D/P)T^2_{triangle}/A}$. Hence, the smallest integer N that satisfies the above inequality is $N = \left\lfloor \sqrt{D/2(1-D/P)T^2_{triangle}/A} \right\rfloor$. For each N, there are at most five different cases, thus, the maximum number of local minimum points we need to explore before reaching the global optimality is $5\bar{N}$.

12.6 Computational experiments

In this section, we conduct numerical experiments for a potential anthrax attack on our proposed model. It serves two purposes. First, we compare our proposed model that combines two types of demand (an emergency demand—the $I_{\min}$ minimum inventory requirement, and a regular market demand) into a single system, with a standard model that runs the two parts separately; second, we are interested in investigating how the government controlled parameters affect the firm's profit, and thus it provides the government a perspective to negotiate parameters with the firms producing the drugs.

This section is organized as follows. We first discuss the parameter estimation in the experiments, then compare three different models (the proposed model, a standard model and a constant production model) and conduct a sensitivity analysis for our proposed model.

12.6.1 Parameter estimation

In a potential anthrax attack, the federal government is prepared to treat 10 million exposed persons. This represents a stockpile of 1.2 billion Cipro pills (the treatment regimen is two pills a day for 60 days), as $I_{\min}$ used in our experiments. According to the *Cipro Pharmacy.com* website (http://www.ciprofloxacinpharmacy.com/active.html): 'Cipro has a shelf life of approximately 36 months. However, material has been and is currently being tested through the DOD/FDA Shelf Life Extension Program (SLEP) and has received extensions up to 7 1/2 years from original expiration date and some lots have received up to 9 years from original expiration date. Materiel shows no signs of deteriorating based on yearly test'. Based on the above statement, we use 9 years (108 months) as the shelf-life of the drug, which indicates that the government would pay for the production and storage of $I_{\min}$ every 9 years. We define a parameter Y to represent the flexibility that the government gives to firms, which is defined as the number of months before the expiration that the government allows $I_{\min}$ to be sold. With the constraint that the maximum cycle length (T_{inv} in our model) must be at most half a shelf-life (we use shelf-life minus the flexibility here to represent the actual allowable on-shelf period). In this section, we use the upper limit that $T_{inv} = (1/2)(T_s - Y) = (108 - Y)/2$, which dictates the relationship between T_{inv} and Y. We use 36 months for Y in the base case, which represents that the government allows firms to resell the required $I_{\min}$ 3 years prior to its expiration. Then $T_{inv} = 108–36 \times 2 = 36$ months = 3 years. Please note that the government still pays the firm for producing and keeping

the inventory every 9 years (108 months) regardless the flexibility (Y) it allows.

We also define p_{gov} as the price that the government pays to the firms per pill for production and storage; p_{market} as the price the firms can sell to the regular US domestic market per pill. We estimate p_{gov}, p_{market}, v and w according to Socolar and Sager (2002). Based on the same source, we assume that the government uses a tier pricing strategy to pay firms, 95 cents listed in Table 12.1 represents the price per pill for the first million pills, the second million pills, is 10 cent cheaper per pill and from the third million onwards, there is another 10 cent per pill discount. Hence, as long as the first million pills' price is given, the total price that the federal government pays to firms is fixed every year. We use 30 cents, the price of generic at Ranbaxy in India (Socolar and Sager, 2002) as the salvage value after 36 months to the secondary overseas market, and 7.5 cents if after 72 months (this will be used in the standard model's calculation only).

The parameters related to the firms' manufacturing and inventory keeping are approximated according to Singh (2001). 'In the US alone, Bayer sold $1.04 billion worth of Cipro in 1999'. This is equivalent to about 220(≈ 1040/4.5) million pills in year 1999; with a 4% per year growth rate, the demand reaches 300 million in year 2007/2008 and we use this as the demand rate D in our model. 'At best, Bayer offered to produce 200 million pills within 60 days'. This indicates a maximum production capacity as 1200 million per year, we use half of it—600 million per year as the production rate allocated for our system. The inventory holding cost (h) mainly comes from the capital cost, with an expected 10% annual investment return rate, $h \approx 0.2 \times 10\% = 0.02$ mil$/mil pill/year. We assume the ordering cost A is 100 times the h value. In Table 12.1, we summarize the estimated parameter values as discussed above, and we use them as the base case in our experiments.

As we combine the two parts of demand into our proposed model: the regular market demand with constant rate D and the emergency demand $I_{\min}$ required by the federal government, which are two distinct resources that generate revenues. According to our assumption, both revenues are independent of the *production batch size* (Q); hence, the total revenue per year is not a function of Q. As the profit equals to revenue minus cost and the revenue is independent of Q, hence, minimizing the cost is the same as maximizing the profit. We emphasize this equivalence relationship because we used minimizing the cost in the analysis in Section 4 and we will use the total profit in the following experimental Section (6.3) to illustrate how the government controlled

Table 12.1 Estimated numerical value of the parameters used in the experiment

Parameter	Unit	Estimation
$I_{\min}$	million pills	1200 (=10 × 60 × 2)
T_{inv}	month	36
Y	month	36 (= 108–2 × T_{inv})
p_{gov}	mil $/mil pill	0.95
p_{market}	mil $/mil pill	4.67
v	mil $/mil pill	0.2
w	mil $/mil pill	–0.3(3 years)/ –0.075(6 years)
D	mil pill/year	300
P	mil pill/year	600
A	mil $/time	2 (=100 × h)
h	mil $/mil pill/year	0.02

parameters ($I_{\min}$, p_{gov} and Y) affect the total profit based on the analysis in section 4.

12.6.2 Model comparison

We use the base case parameter setting to run our proposed model. We also use the same parameters to run a constant production model and a standard EMQ model (Table 12.2). A comparable constant production model would have an inventory cycle of T_{inv} = 3 years and produce at a rate of 400 million pills per year. We assume the excess inventory is stored and salvaged at the end of the inventory cycle (at 30 cents per pill). The whole 1.2 billion inventory is completely refreshed every 3 years. The standard EMQ model splits the production capacity of 600 million pills per year into two parts: 400 million pills per year to meet the regular market demand and another 200 million pills per year of constant production to build the 1.2 billion pills of $I_{\min}$, which is completely refreshed every 6 years and we can obtain 7.5 cents per pill at the secondary market. For the constant production and standard model, we force one charge of fixed costs per 3 years on constant producing machines.

Tables 12.2 and 12.3 compare the costs and idled machine capacity involved in each of these models respectively.

The detailed calculation of the table is explained in Appendix D. As the results in the table demonstrate, our model costs 77.33 million per year to run the combined system that saves 34.78 million a year compared with the standard model. This illustrates the significant benefit (30% cost-saving in a year) by using our proposed model instead of

Table 12.2 Model comparison of decomposed cost (mil $/year)

	Proposed model	**Standard model**	**Constant production model**
Emergency Inv. Carrying Cost	24	24	24
Regular Inv. Carrying cost	2	2.45	3
Fixed Setup Cost	1.33	0.66	0.66
Salvage Cost	–30	–15	–30
Variable Cost	80	100	80
Total Cost	77.33	112.11	77.66

Table 12.3 Model comparison of idle machine capacity per year

	Proposed model	**Standard model**	**Constant production model**
Average Machine Idle Time	1/3	1/2	1
Machine Production Capacity	600	400	200
Average Machine Idle Capacity	200	200	200

running two separate systems. It is caused by the fact that the standard model keeps completely separate inventory for emergency and for regular demand (which are combined in the other two models). This shows that the separation causes us to produce at a higher rate than the other two models due to the demand and lifetime constraints. The proposed model is comparable in cost to a constant production model. However, EMQ models are more favourable in certain industries since it is not only minimizing the cost, but also spares certain amount of full production capacity to other products (details see the comparison in Table 12.3). The total production capacity that is left idle (200 mil pills/year) is the same, but the distribution is different. The proposed model has spare capacity 600 mil pills/year for one-third of a year and the constant production model leaves constant capacity 200 mil pills/year for a whole year. It usually depends on the industry and other demand requirements to determine which scenario is more favourable. If it is easy to break up production and manage multiple lines then the idle capacity is equivalent. If there is little space or management facilities to handle multiple products at the same time, then the 600 mil pills/year for one-third of a year would be superior. The 200 mil pills/year for 1 year would be superior when it is easy to handle multiple products that are demanded continuously through the year. In the process industry,

production lines tend to be setup to produce a single product type and it may not be feasible to divide the production capacity. Thus, for this industry, it may be preferable to run a line at full production capacity over a period of time than a portion of capacity continuously. Therefore the proposed EMQ-based model is more desirable.

12.6.3 Sensitivity analysis

In the previous subsection, we demonstrated the advantage of our proposed model over a standard model by using the base case parameter settings. Now we are interested in the sensitivity analysis on the government controlled parameters: $I_{\min}$, Y and p_{gov}. We use different salvage strategies to investigate how the profit changes for different $I_{\min}$, Y and p_{gov} values. All figures in this subsection are plotted as the profit from the manufacturer's perspective over the parameters controlled by the government. The government can use these plots to negotiate the parameter settings with the firms.

Figures 12.8 and 12.9 plot the relationship between the total profit and the flexibility (Y) given by the government at different $I_{\min}$ level. The x-axis represents Y which ranges from 0 to 69 months and it corresponds to T_{inv} from 54 to 19.5 months. The y-axis represents the total profit and different plot lines denote different $I_{\min}$ levels.

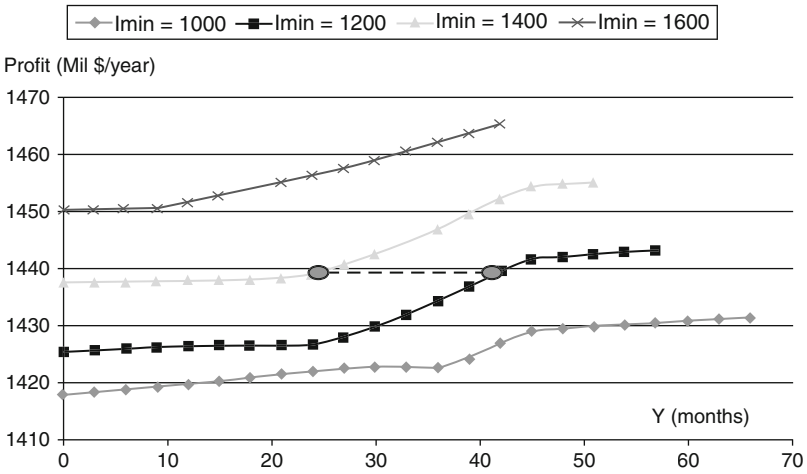


Figure 12.8 Profit v.s. Y with $Q_{\max} = 1600$ at different $I_{\min}$ level

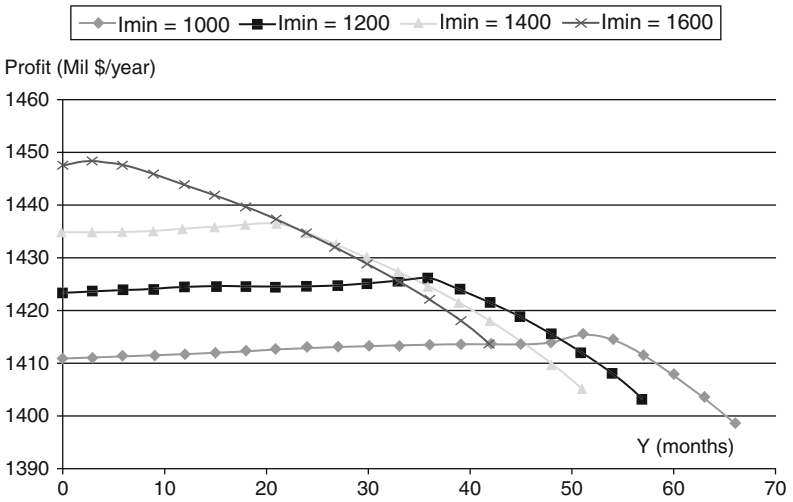


Figure 12.9 Profit v.s. *Y* which allows salvage at most 300 million pills

In Figure 12.8, we limit the *maximum production batch size* (Q_{max}) as 1600 million pills. Based on this figure, we have the following observations:

- For any fixed flexibility (Y), the higher I_{min} level, the more profit the firm can gain; this is due to the extra revenue obtained from the government is higher than the additional cost required for maintaining the extra I_{min}.
- For any fixed I_{min} level, the more flexibility allowed (larger Y, which means firms can salvage the I_{min} amount earlier), the higher profit firms can gain; this is due to the extra flexibility given to firms that they can refresh their inventory in a shorter period hence reduce the average running cost of the proposed inventory system. However, the profit increases at different slopes with different Y values. This phenomenon will be explained in a later discussion.
- The two dots, which on two different I_{min} level plot lines and at the same profit level, demonstrate the trade-off between the high flexibility for the firms and the low minimum inventory requirement for the government. If the government would pay less to the firms by requiring a smaller amount of minimum inventory, it must allow more flexibility (longer time before the expiration to salvage the pills) for the firms to obtain the same level of profitability.

In Figure 12.9, we use a different salvage strategy by assuming that we can earn salvage value for the first 300 million disposed pills only. Additional pills will be disposed at a zero salvage value. Compared with Figure 12.8, this figure shares the same plot shape when Y is low. But the plots start to have a negative slope at the points when the salvage amount reaches 300 million pills. The turning points are at different Y values for different $I_{\min}$ plots; the lower the $I_{\min}$, the larger the turning point's Y value. This is explained by the fact that a less minimum inventory requirement will reach a given salvage amount with higher flexibility (or, a larger Y); or from the mathematical formula: the salvage amount = $I_{\min} - DT_{inv} = I_{\min} - D(9-Y)/2 = I_{\min} - 4.5D + (DY/2)$. For a fixed salvage amount, a smaller $I_{\min}$ comes with a larger Y. When $I_{\min}$ is fixed, as the flexibility Y increases, the inventory cycle (T_{inv}) is reduced; this results in more salvage amount at the end of each T_{inv}. If the salvage amount exceeds the 300 million limit, only the first 300 million is valuable; however, we still need to pay the unit cost for those salvaged stocks above the 300 million. Thus, after the turning point, the more we are forced to salvage, the less profit we gain.

Figure 12.10 illustrates the relationsonship between the profit and the flexibility (Y) at different prices the government would pay firms. Again we limit the *maximum production batch size* ($Q_{\max}$) as 1600 million pills and use the base case $I_{\min}$ as 1.2 billion pills. Similar to Figure 12.8, we have the following observations:

- The higher unit price the government pays, the higher profit the firm gains.
- The two dots, which on two different p_{gov} level plot lines and at the same profit level, demonstrate the trade-off between the high flexibility for the firms and the low price for the government. If the government would pay less to the firms by reducing the unit price of a pill, it must allow more flexibility (longer time before the expiration to salvage the pills) for firms to obtain the same level of profitability.

Figure 12.11 provides a graph explanation on the different slopes at different segments on the profit v.s. flexibility (Y) plot. It uses the base case parameters and gives the inventory plots at the *optimal production batch size* Q^* for different Y values. Through the sensitivity analysis, the marginal profit gained over a fixed small ΔY in (a), (b) and (d) segments all come from the cost saving on the regular EMQ part, which is trivial, compared with the marginal profit gain over a fixed small ΔY at the (c) segment, which is from the additional gain due to the extra salvage amount.

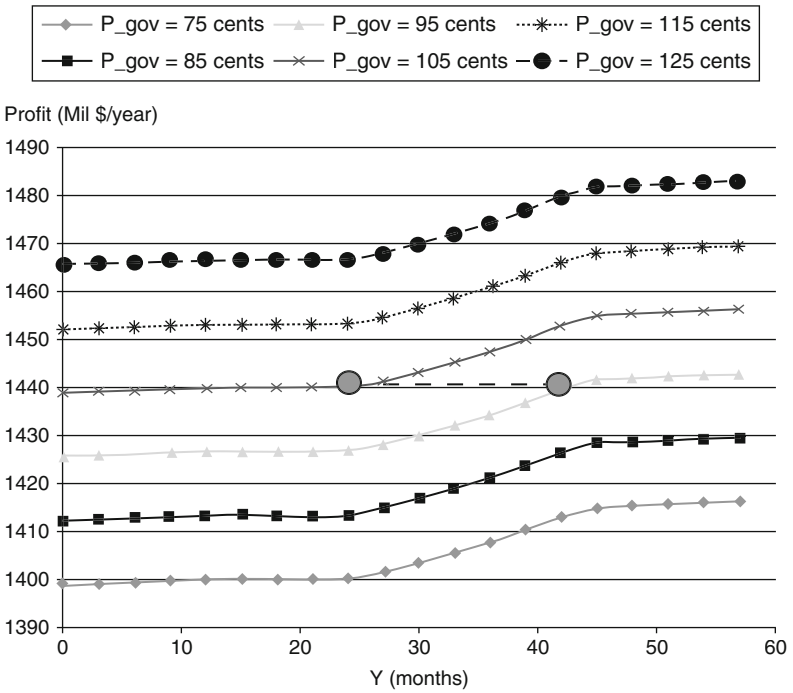


Figure 12.10 Profit v.s. Y with Q_{max} = 1600 at different unit government-paid price

In summary, in this experimental section, we estimate the model parameters for a potential anthrax attack scenario and use them to quantitatively demonstrate the advantage of our proposed model over a standard model. We also perform a sensitivity analysis on the government controlled parameters to provide the aid for the negotiation on the parameter settings between the federal government and the firms.

12.7 Conclusion and future research

For the perishable inventory management system, a trivial extension of a regular EMQ model is adequate when the required minimum inventory is not significant compared with the amount consumed by the regular market demand rate within the shelf-life. However, when we consider the VMIs in the SNS for the large-scale emergency setting,

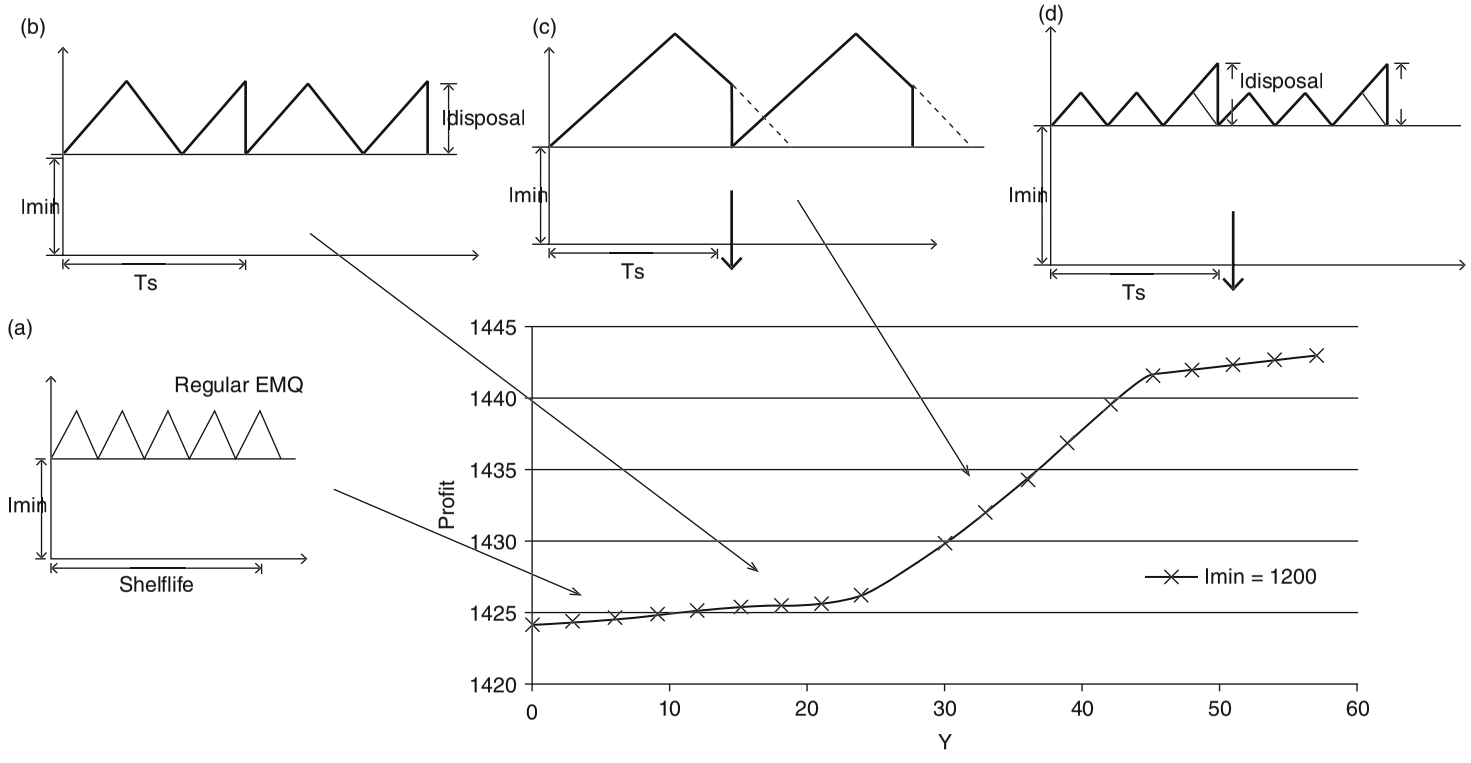


Figure 12.11 Illustration of the different slopes in the profit v.s. *Y* plot

the minimum inventory requirement is in a scale which is comparable with the total market consumption within the shelf-life. A more sophisticated inventory management strategy is required to maintain a fresh and massive stockpile. Hence, in this work, we modelled the perishable inventory management problem with a minimum inventory volume constraint as a modified economic manufacturing quantity (EMQ) model. We discussed the policies and assumptions adopted in this model from both a regular perishable inventory management context and the special constraints on the minimum stock size and the maximum inventory cycle length enforced by the large-scale emergency response context. We formulated the problem to minimize the total relevant cost with respect to the production batch size as an unconstrained non-continuous non-differentiable optimization problem. We proved the existence of the local as well as global minimum of the total cost with respect to the order quantity. Hence, an exact solution procedure is proposed and its complexity is proved to be pseudo-polynomial.

We estimated the parameters in the modified EMQ model for a potential anthrax attack scenario from various sources and used them to compare the proposed model with a standard model to show a significant cost saving of running our system as around 34 million US dollars per year, which is a savings of about 30% on the cost of the standard model. We also compare the proposed model with a constant production model. Our model has slightly smaller cost and has a more favourable spare production capacity. We performed sensitivity analysis on some government controlled parameters in the system and observed that at a given profitability level for the firm, there are trade-offs between the less amount paid by the government to firms (either by reducing the $I_{\min}$ requirement or by reducing the unit price p_{gov}) with the higher flexibility the government allows to firms (longer time before the expiration to salvage the pills).

For the inventory management problem, our current model assumed uniform unit price on the items sold to the regular market and uniform unit salvage value on the items disposed at the end of each inventory cycle regardless of their age. An interest in the future research is to combine the inventory model with a revenue model to address the potential economic impact with more sophisticated issuing, pricing and disposing strategies that incorporate the age distribution of the stockpile. Another interesting direction in future research is to model with multiple suppliers competing for contracts and study the model and strategy for the manufacturer-government interaction and negotiation.

References

Abad P (1996). Optimal pricing and lot-sizing under conditions of perishability and partial backordering. *Mngt Sci* **42**: 1093–1104.

Burnetas A and Smith C (2000). Adaptive ordering and pricing for perishable products. *Opns Res* **48:** 436–443.

Chazan D and Gal S (1977). A Markovian model for a perishable product inventory. *Mngt Sci* **23**: 512–521.

Chun YH (2003). Optimal pricing and ordering policies for perishable commodities. *Eur J Opl Res* **144:** 62–82.

Gallego G and van Ryzin G (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Mngt Sci* **40:** 999–1020.

Giri B and Chaudhuri K (1998). Deterministic models of perishable inventory with stock-dependent demand rate and nonlinear holding cost. *Eur J Opl Res* **105:** 467–474.

Goh C-H, Greenberg B and Matsuo H (1993). Two-stage perishable inventory models. *Mngt Sci* **39**: 633–649.

Graves S (1982). The application of queuing theory to continuous perishable inventory systems. *Mngt Sci* **28**: 400–406.

Hsu VN (2000). Dynamic economic lot size model with perishable inventory. *Mngt Sci* **46**: 1159–1169.

Hsu VN and Lowe T (2001). Dynamic economic lot size models with period-pair-dependent backorder and inventory costs. *Opns Res* **49**: 316–321.

Katagiri H and Ishii H (2002). Fuzzy inventory problems for perishable commodities. *Eur J Opl Res* **138:** 545–553.

Liu L and Lian Z (1999). (*s, S*) continuous review models for products with fixed lifetimes. *Opns Res* **47**: 150–158.

Nahmias S (1982). Perishable inventory theory: A review. *Opns Res* **30:** 680–708.

Nandakumar P and Morton T (1993). Near myopic heuristics for the fixed-life perishability problem. *Mngt Sci* **39**: 1490–1498.

Pierskalla W and Roach C (1972). Optimal issuing policies for perishable inventory. *Mngt Sci* **18:** 603–614.

Ravichandran N (1995). Stochastic analysis of a continuous review perishable inventory system with positive lead time and Poisson demand. *Eur J Opl Res* **84:** 444–457.

Rosenfield D (1989). Disposal of excess inventory. *Opns Res* 37: 404–409.

Rosenfield D (1992). Optimality of myopic policies in disposing excess inventory. *Opns Res* **40:** 800–803.

Shen Z (2008). *Routing and inventory models for emergency response to minimize unmet demand*. PhD thesis, University of Southern California.

Singh K (2001, Nov). War profiteering: Bayer, anthrax and international trade, http://www.corpwatch.org/article.php?id=723.

Sobel M and Zhang R (2001). Inventory policies for systems with stochastic and deterministic demand. *Opns Res* **49**: 157–162.

Socolar D and Sager A (2002, June). Cipro prices expose windfall profits. Health Services Department, Boston University School of Public Health.

Wagner HM and Whitin TM (1958). Dynamic version of the economic lot size model. *Mngt Sci* **5**: 89–96.

Weiss H (1980). Optimal ordering policies for continuous review perishable inventory models. *Opns Res* **28**: 365–374.

Appendix A

Proof of proposition 1

Proposition 1 *An EMQ model for a perishable inventory with minimum inventory requirement of* $I_{\min} \leq DT_s$ *has an optimal production batch size given by*

$$Q^* = \min\left\{EMQ^*, T_s\frac{DP}{P-D} - I_{\min}\frac{P+D}{P-D}, \left(T_s - \frac{I_{\min}}{P}\right)\frac{DP^2}{P^2-D^2}\right\}.$$

Proof When $I_{\min}$ is non-zero and in a small amount, as in Figure 12.1, with a FIFO policy, an $I_{\min}$ amount is first consumed to meet the market demand over the period $T_{\min} = I_{\min}/D$. There are two possible upper bounds of the oldest item: one is T_{b1} which is the oldest item consumed at the end of each T; another is T_{b2} which is the oldest item in $I_{\min}$ before it completely depletes at the beginning of each T cycle. In Figure 12.1 part (a), $T_{\min}$ is longer than T_1. T_1 can be decomposed into two parts: $T' = I_{\min}/P$ which produces the $I_{\min}$ amount consumed at the beginning of the next cycle and $T'_{2p} = ((T_1 - T_{\min})D)/P$. Therefore, the upper bound of the oldest item should be $T_{b2} = T' + T_2 + T_{\min} = ((P-D)/P)T + ((P+D)/PD)I_{\min} \leq T_s$. In Figure 12.1 part (b), $T_{\min}$ is shorter than T_1 which can be decomposed into three parts: T' with the same definition as in part (a), T_{2p} that produces the items consumed in T_2 period and $T'_{1p} = ((T - T_{\min})D)/P$ that produces the items consumed in T_1 after depleting $I_{\min}$ from the previous cycle. Hence, it shares the same T_{b2} bound as in part (a), we also have another upper bound of the oldest item $T_{b1} = T - T'_{1p} = ((P-D)/P)((P+D)/P)T + (I_{\min}/P) \leq T_s$. Therefore, when $I_{\min}$ is upper bounded by T_sD (when $T_{\min} \leq T_s$), the *optimal production batch size* for the perishable items with $I_{\min}$ minimum inventory requirement is $Q^* = \min\{EMQ^*, T_s(DP/(P-D)) - I_{\min}(P+D)/(P-D), (T_s - (I_{\min}/P))(DP^2/(P^2-D^2))\}$.

Appendix B

Detailed graph illustration of cases 2, 4 and 5

See Figure 12.12, 12.13 and 12.14.

Appendix C

Total cost calculation and boundary conditions for case 3

$$TRC(Q) = \frac{1}{2}h[M \cdot T \cdot I_{\max} - T_{p1} \cdot DT_{p1}) + (I_{disposal} + T_{p1} \cdot D) \cdot (T_{inv} - M \cdot T + T_{p1})]$$

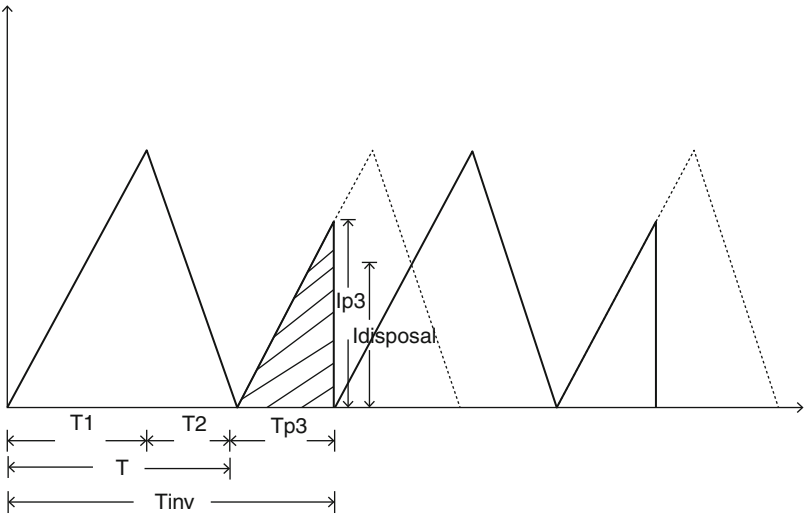


Figure 12.12 Graph illustration for case 2

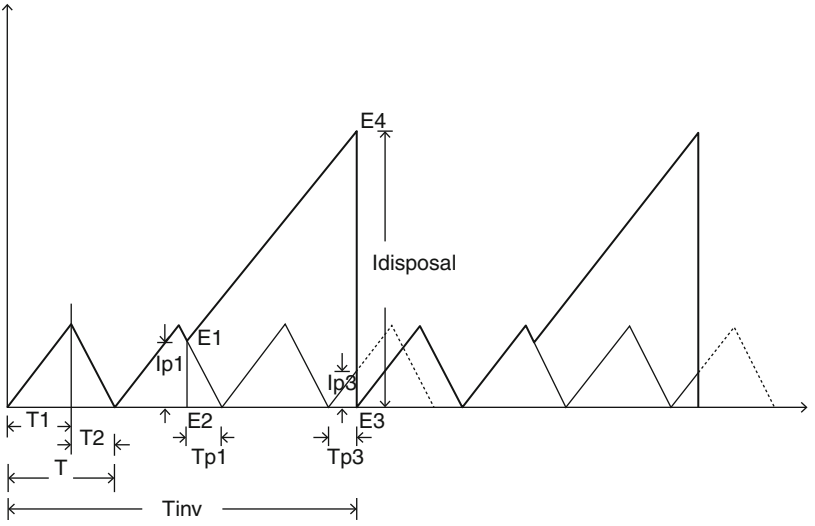


Figure 12.13 Graph illustration for case 4

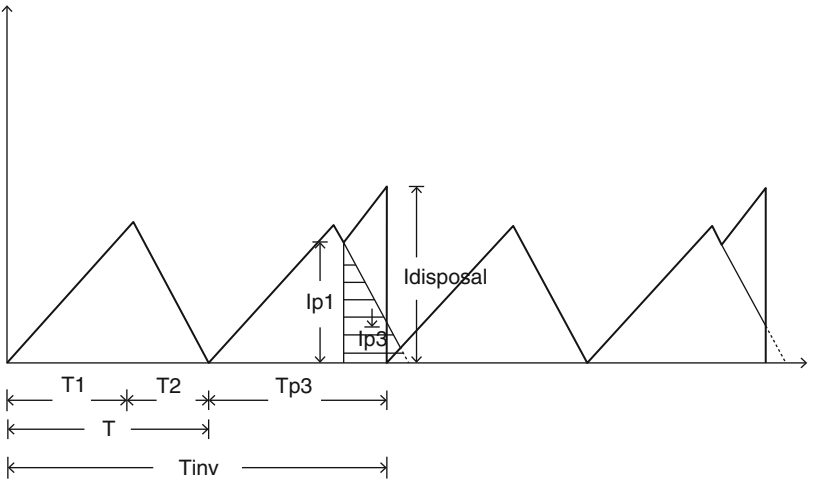


Figure 12.14 Graph illustration for case 5

We first compute T_{p1}, which can be expressed as a linear function of Q:

$$\begin{aligned} T_{p1} &= T_{disposal} - (T_{p3} - T_1) - N_1 \cdot T_2 \\ &= \frac{I_{disposal} + T_{inv} \cdot (D - P)}{P} \\ &\quad + (N + 1 - N_1) \cdot \left(\frac{1}{D} - \frac{1}{P} \right) \cdot Q \\ &= \beta_1 + \alpha_1 \cdot Q \end{aligned}$$

Where we define:

$$\alpha_1 = (M + 1) \cdot \left(\frac{1}{D} - \frac{1}{P} \right)$$

$$\beta_1 = \frac{I_{\min}}{P} - T_{inv}.$$

Hence we have:

$$\begin{aligned} TRC(Q) = \frac{1}{2} h [& M \cdot \frac{Q^2}{D} \cdot \left(1 - \frac{D}{P} \right) \\ & - (\alpha \cdot Q + \beta)^2 \cdot D \\ & + (I_{disposal} + (\alpha \cdot Q + \beta) \cdot D) \\ & \cdot \left(T_{inv} - M \cdot \frac{Q}{D} + \alpha \cdot Q + \beta \right)] \end{aligned}$$

We let $TRC(Q) = a \cdot Q^2 + b \cdot Q + c$, then (Figure 12.15)

$$\begin{aligned} a &= \frac{1}{2}h\left[M \cdot \left(\frac{1}{D} - \frac{1}{P}\right) - D \cdot \alpha^2 \right. \\ &\quad \left. + \alpha \cdot D \cdot \left(\alpha - \frac{M}{D}\right)\right] \\ &= -\frac{h}{2} \cdot M^2 \cdot \left(\frac{1}{D} - \frac{1}{P}\right) < 0 \end{aligned}$$

$$b = \frac{h}{2} \cdot \left[(2M+1) \cdot \left(T_{inv} - \frac{I_{disposal}}{P} - \frac{D}{P} \cdot T_{inv}\right) + \frac{I_{disposal}}{D}\right]$$

The $TRC(Q)$ is a quadratic function of Q and the stationary point of the quadratic curve is (where $a < 0$):

$$\begin{aligned} Q^*_{case\,3} &= -\frac{b}{2a} \\ &= \frac{(2M+1) \cdot \left(T_{inv} - \frac{I_{disposal}}{P} - \frac{D}{P} \cdot T_{inv}\right) + \frac{I_{disposal}}{D}}{2M^2 \cdot \left(\frac{1}{D} - \frac{1}{P}\right)} \end{aligned}$$

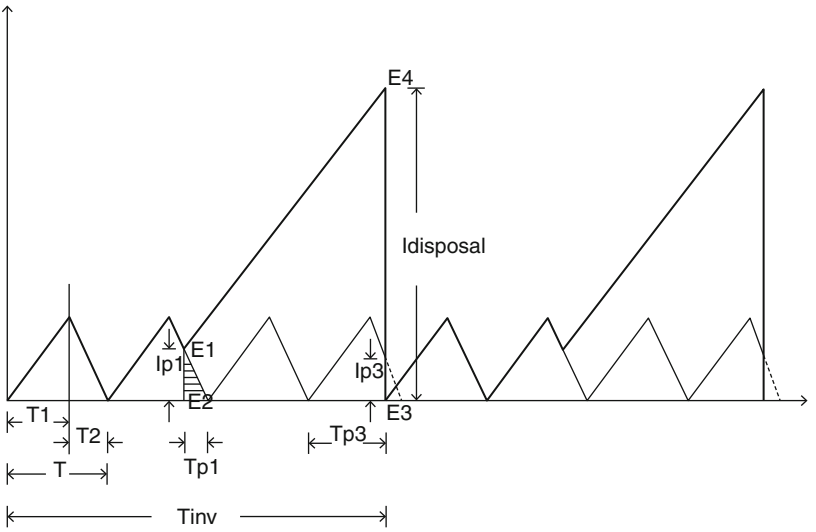


Figure 12.15 Graph illustration for case 3

Next, we show the calculation of the boundary point values for case 3. Without loss of generality, we assume that the *N* and *M* values are both fixed and as shown in Figure 12.16, the T_{inv}, period ends at somewhere in the downhill slope between *C* and *D;* the *last production cycle* starts somewhere in the downhill region between *A* and *B*. From Figure 12.16, we can infer that only if $I_{disposal}$ falls in range 1 or range 2 or range 3, it is feasible for case 3 with the parameter *M* and *N* valid (that is, T_{inv} ends between C and *D* and the *last production cycle* starts between *A* and *B*). Otherwise, if $I_{disposal}$ is too short (below range 3) or too high (above range 2) with fixed *N*, the value of parameter *M* would increase or decrease accordingly.

When $I_{disposal}$ falls into range 1, the two boundary points of case 3 with *M* and *N* are:

(1) T_{inv} ends at *D,* and the *last production cycle* starts at *SD*1;
(2) T_{inv}, ends at C, and the *last production cycle* starts at *SC*1.

Or when $I_{disposal}$ falls into range 2, the two boundary points are:

(1) T_{inv} ends at *D*, and the *last production cycle* starts at *SD2;*
(2) T_{inv} ends at *EA*2, and the *last production cycle* starts at *A*.

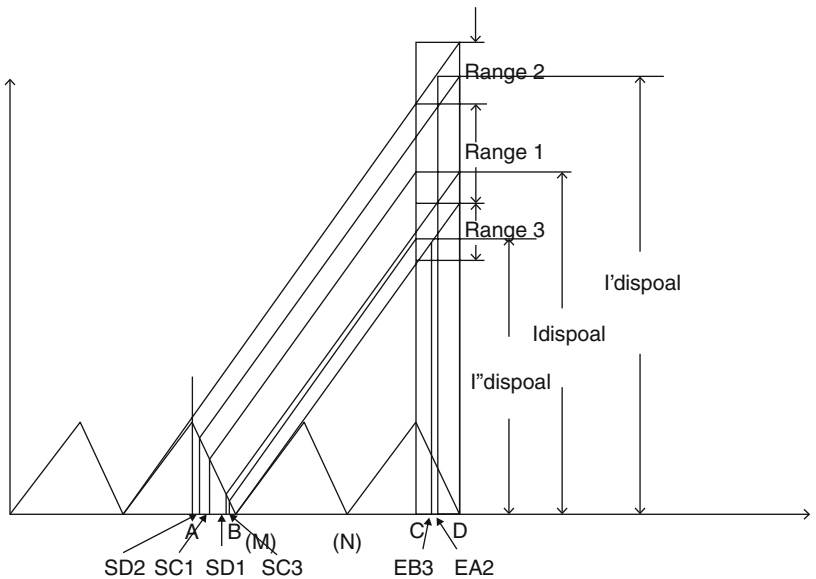


Figure 12.16 Graph illustration for the boundary value of case 3

Or when $I_{disposal}$ falls into range 3, the two boundary points are:

(1) T_{inv} ends at *EB*3, and the *last production cycle* starts at *B*;
(2) T_{inv} ends at *C*, and the *last production cycle* starts at *SC*3.

Next, we will calculate the *Q* values corresponding to these boundary points.

For fixed *N* and *M* values, when T_{inv} ends at point *D*, we have $(N + 1) \cdot T = T_{inv}$. Since $T = (Q_D/D)$, we have $(N + 1) \cdot (Q_D/D) = T_{inv}$. Hence $Q_D = (T_{inv} \cdot D)/(N + 1)$, where Q_D is the *production batch size* when T_{inv} ends at point *D*.

When T_{inv} ends at point C, we have $N \cdot T + T_1 = T_{inv}$, so $N \times (Q_C/D) + Q/P = T_{inv}$. Hence, $Q_C = T_{inv}/(N/D + 1/P)$, where Q_C is the *production batch size* when T_{inv} ends at point *C*.

If T_{inv} ends somewhere between *C* and *D* and the *last production cycle* starts at *A*, we have

$$\frac{I_{disposal} - I_{p3}}{P} = (N - M + 1) \cdot T_2 + [T_{inv} - (N \cdot T + T_1)],$$

If we use Q_A to represent the *production batch size* when the *last production cycle* starts from *A*, then:

$$\frac{I_{disposal} - \left[(N+1) \cdot \frac{Q_A}{D} - T_{inv}\right] \cdot D}{P} = (N - M + 1) \cdot Q_A \cdot \left(\frac{1}{D} - \frac{1}{P}\right) + \left[T_{inv} - \left(N \cdot \frac{Q_A}{D} + \frac{Q_A}{P}\right)\right]$$

Hence we have

$$Q_A = \frac{(P - D) \cdot T_{inv} - I_{disposal}}{P \cdot (M - 1) \cdot \left(\frac{1}{D} - \frac{1}{P}\right)}$$

If T_{inv} ends somewhere between *C* and *D* and the *last production cycle* starts at *B*, we have

$$\frac{I_{disposal} - I_{p3}}{P} = (N - M) \cdot T_2 + [T_{inv} - (N \cdot T + T_1)]$$

If we use Q_B to represent the *production batch size* when the *last production cycle* starts from *B*, with the similar calculation for Q_A, we have:

$$Q_B = \frac{(P-D)\cdot T_{inv} - I_{disposal}}{P\cdot M\cdot\left(\frac{1}{D}-\frac{1}{P}\right)}$$

Appendix D

Detailed total cost calculation for model comparison

For our proposed model, there are four parts in the total cost:

- Inventory carrying cost:
  - Emergency part: 1200 × 0.02 = $24 mil/year;
  - Regular market part: (1/2 × 150 × 2+1/2 × 300) × 0.02 = 6 million for 3 years, which is $2 million per year;
- Fixed setup cost: 2 times in each T_{inv} (3 years), which is $1.33 million per year;
- Salvage cost: 300 × (–0.3) = –90 million dollars in 3 years, which is -30 million dollars per year;
- Variable cost: $I_{\min} \times v$ = 1200 × 0.2 = 240 million dollars for 3 years, which is $80 million per year.

The total annual cost of the proposed system is: 24 + 2 + 1.33–30 + 80 = $77.33 million

For the standard model, we calculate the cost for the two systems separately:

- Regular market EMQ model:
  - Inventory carrying cost: $\sqrt{2hAD(1-D/P)} = \$2.45$ mil/year;
  - Variable cost: 0.2 × 300 = $60 mil/year;
- Emergency demand:
  - Inventory holding cost: 120 × 0.02 = $24 mil/year; Fixed Setup cost: 1 times in each T_{inv} (3 years), which is $0.66 million per year;
  - Variable cost: 1200/6 × 0.2 = $40 mil/year;
  - Salvage cost: –0.075 × 1200/6 = –$15 mil/year.

The total annual cost of the standard model is: 2.45 + 60 + 24 + 0.66 + 40 – 15 = $112.11 million.

For the constant production model, there are only two parts in the total cost:

- Inventory carrying cost:
  - Emergency part: 1200 × 0.02 = $24 mil/year;

- Regular market part: (1/2 × 300 × 3) × 0.02 = 9 million for 3 years, which is $3 million per year;
- Fixed Setup cost: 1 times in each T_{inv} (3 years), which is $0.66 million per year;
- Salvage cost: 300 × (–0.3) = –90 million dollars in 3 years, which is –30 million dollars per year;
- Variable cost:/min × v = 1200 × 0.2 = 240 million dollars for 3 years, which is $80 million per year.
- Fixed Setup cost: 1 times in each T_{inv} (3 years), which is $0.66 million per year;
- Salvage cost: 300 × (–0.3) = –90 million dollars in 3 years, which is –30 million dollars per year;
- Variable cost: $I_{\min}$ × v = 1200 × 0.2 = 240 million dollars for 3 years, which is $80 million per year.

The total annual cost of the proposed system is: 24 + 3 + 0.66–30 + 80 = $77.66 million.

13

A Bayesian Decision Model with Hurricane Forecast Updates for Emergency Supplies Inventory Management

S. Taskin[1] *and E. J. Lodree, Jr*[2]
[1]*Aselsan Inc., Ankara, Turkey; and* [2]*Auburn University, Auburn, Alabama, USA*

Hurricane forecasts are intended to convey information that is useful in helping individuals and organizations make decisions. For example, decisions include whether a mandatory evacuation should be issued, where emergency evacuation shelters should be located, and what are the appropriate quantities of emergency supplies that should be stockpiled at various locations. This paper incorporates one of the National Hurricane Center's official prediction models into a Bayesian decision framework to address complex decisions made in response to an observed tropical cyclone. The Bayesian decision process accounts for the trade-off between improving forecast accuracy and deteriorating cost efficiency (with respect to implementing a decision) as the storm evolves, which is characteristic of the above-mentioned decisions. The specific application addressed in this paper is a single-supplier, multi-retailer supply chain system in which demand at each retailer location is a random variable that is affected by the trajectory of an observed hurricane. The solution methodology is illustrated through numerical examples, and the benefit of the proposed approach compared to a traditional approach is discussed.

13.1 Background

Imagine the series of events triggered by the development of a tropical depression in the Atlantic or Pacific basin: increased activity in

Reprinted from *Journal of the Operational Research Society*, 62: 1098–1108, 2011, 'A Bayesian Decision Model with Hurricane Forecast Updates for Emergency Supplies Inventory Management', by S. Taskin and E. J. Lodree. With kind permission from Operational Research Society Ltd. All rights reserved.

state, regional, and national offices of the National Hurricane Center (NHC) with hurricane experts generating a host of reports related to the observed storm, which include track predictions, intensity predictions, and hurricane/tropical storm warnings; all of which will be updated on a regular basis throughout the storm's life cycle based on continuous monitoring. Local and state governments then use the information generated by the NHC to make important and costly decisions, namely if and when to mandate evacuation. Individual households also use hurricane forecasts and information updates to decide whether to evacuate voluntarily, whether they will adhere to a mandatory evacuation, and if they choose to evacuate, when they will do so. Also related to evacuation, non-government organizations (NGOs), such as the American Red Cross, use hurricane forecasts to make decisions such as the number of shelters to prepare, where to locate them, how to staff them, and how to distribute emergency supplies among them.

The above-mentioned chain of activities in response to an observed tropical cyclone demonstrate the importance of hurricane forecasts in making decisions related to evacuation. However, the NHC forecasts are useful for other types of decisions as well. For instance, hurricane forecasts can be translated into damage projections in monetary terms, which is of interest to insurance companies. Hurricane forecasts can also be used to estimate demands for emergency supplies such as non-perishable food items and gas-powered generators, which are used by households, first responders, and emergency evacuation shelters before, during, and after a landfall cyclone. These demand estimates can help emergency managers, NGOs, local retailers, and manufacturing facilities plan inventory levels for these emergency supplies, which can, in turn, help improve the responsiveness and efficiency of the humanitarian relief supply chain.

13.2 Problem description

This paper is motivated by a manufacturing facility whose demand for emergency supplies is influenced by the potential impact of a storm that has recently formed in the Atlantic or Pacific basin. Specifically, the manufacturer serves multiple retailers in a humanitarian relief chain including private sector retail stores, emergency evacuation shelters, NGOs, and government responders. These 'retailers' each place an order to the manufacturer in response to hurricane forecasts issued by the NHC, which are used to estimate each retailer's demand. The production manager must determine inventory levels for emergency supplies

(say, generators) in anticipation of the retailers' orders. During the earlier portions of the storm's development while it is still in the ocean, it is difficult for the production manager to gauge the demand for generators. If the storm eventually develops into a major hurricane and passes over land (particularly heavily populated areas), then there will be a demand surge for generators. If, on the other hand, the storm decides to remain in the ocean and does not threaten any populated areas, then there is not likely to be a demand surge for generators. In this situation, weather predictions regarding the observed storm's development will be of great service in helping the production manager make decisions about inventory levels.

The production manager is confronted with a trade-off. During the earlier stages of the storm's development, the facility has the capacity to reach potentially large inventory levels because of the hurricane's inherent slow onset. In other words, the manufacturer has several days (perhaps 3–5) to reach target inventory levels. However, the forecast accuracy in terms of storm development is limited during the earlier stages. On the other hand, the forecast accuracy is much better during the later stages after the storm has traversed a majority of its path. But in the event that the storm has developed into a major hurricane that will make landfall and affect a heavily populated area, it will be difficult or impossible for the manufacturing facility to respond to demand surge for emergency supplies (eg, generators) because the facility has limited capacity. Under these circumstances, the manufacturing facility's problem entails determining optimal inventory levels for generators, and during the storm's evolution this decision should be made.

The manufacturer's inventory management decisions are very important and challenging when it comes to responding to hurricane events. Low inventory levels could mean low service levels, which in turn propagates human suffering as demonstrated by the failed humanitarian logistics response associated with Hurricane Katrina in 2005. On the other hand, massive amounts of emergency supplies were unsold at the end of the inactive 2006 hurricane season because private sector firms were expecting an active season similar to the previous two years. From this perspective, the conflicting objectives between private sector organizations and humanitarian relief operations are evident. The goal of the private sector firm is to maximize profits, while that of the humanitarian responder is to minimize response time. These goals are conflicting because private sector organizations would have to assume the risk of holding large amounts of emergency supply inventories in order to

achieve quick response, thereby increasing expected costs. On the other hand, private sector organizations can reduce inventories, but this would increase the risk of prolonged humanitarian response time.

The humanitarian relief inventory management problem described above is addressed from the perspective of a private sector manufacturing firm. In particular, we formulate a sequential Bayesian decision model that will help the manufacturer minimize expected inventory control-related costs in response to an observed tropical cyclone. The Bayesian decision model embeds one of the NHC's official hurricane prediction models, namely the *tropical cyclone wind speed probabilities product* (NHC08). Wind speed probability forecasts were first used during the 2006 hurricane season, and represent one of the most recent and widely used developments of the NHC. These probabilistic forecasts are discussed in detail in the 'Probabilistic hurricane prediction model' section. But first, related research literature is reviewed in the 'Literature review' section followed by a detailed presentation of the Bayesian decision model and solution approach in the 'Bayesian decision model' section. Also, the solution methodology is illustrated through numerical examples in the 'Numerical examples' section. The benefit of the proposed approach compared to a traditional approach is discussed in the 'Model validation: an insurance policy framework' section based on the methodology introduced in Lodree and Taskin (2008) for assessing the value of preparing for potential disaster relief activities from an inventory control perspective. Finally, a summary and future research directions are discussed in the 'Summary' section.

13.3 Literature review

This paper is closely related to Lodree and Taskin (2009), who also incorporate hurricane predictions into a sequential statistical decision model. In particular, Lodree and Taskin (2009) formulate a Bayesian decision model that minimizes inventory-related costs for a manufacturing or retail organization that is in the process of preparing for a potential demand surge in response to an observed tropical disturbance. There are two differences between this paper and Lodree and Taskin (2009). First, Lodree and Taskin (2009) consider a single-supplier, single-retailer supply chain, whereas this paper considers a single-supplier, multi-retailer supply chain. Second, the information updates in Lodree and Taskin (2009) are based on an unofficial hurricane prediction model introduced by the authors that only forecasts storm intensity. However,

this paper incorporates an official hurricane prediction model used by the NHC that forecasts both storm intensity and track. Regnier and Harr (2006) and Regnier and Harr (2008) also incorporate hurricane predictions into a similar dynamic decision model, but their decision models address public evacuation decisions with information updates based on a Markov model that forecasts the evolution of an observed hurricane. The underlying theme of the above-mentioned papers is weather-sensitive decision problems based on decision models that use weather forecasts. More general discussions about weather-sensitive decision problems are discussed in Regnier (2005).

Our modelling framework is also similar to Choi *et al* (2004), who formulate a stochastic inventory control problem as an optimal stopping problem. Choi *et al* (2004) assume that demand is represented as a normal distribution and that demand information updates are based on market observations. More specifically, the problem addressed in Choi *et al* (2004) can be classified as a Bayesian inventory model with multiple delivery modes (eg, Choi and Li, 2003; Sethi *et al*, 2005). In general, Bayesian updating has consistently been an active area of research in the inventory control literature since the pioneering work of Dvoretzky and his colleagues (Dvoretzky *et al*, 1952) in 1952. Some recent examples include spare parts inventory management (Aronis *et al*, 2004), quick response inventory control (Choi *et al*, 2006), partially observed demand (Ding *et al*, 2002), supply chain contract design (Wu, 2005), quantity and pricing decisions (Zhang and Chen, 2006), and products with short life cycles (Zhu and Thonemann, 2004). For a more detailed review of the Bayesian and non-Bayesian approaches to information updating in inventory control theory, refer to Sethi *et al* (2005).

This paper is also related to humanitarian relief logistics. Specifically, humanitarian relief efforts do not involve only government entities such as the US Federal Emergency Management Agency (FEMA) and NGOs such as the American Red Cross. Effective humanitarian response also depends heavily upon cooperation and coordination with private sector firms as discussed in Phillips (2007), Townsend (2006) and Van Wassenhove (2006). From this perspective, the manufacturer's inventory control problem described in the 'Problem description' section is directly related to preparedness for potential humanitarian relief operations. In general, the humanitarian relief research literature addresses decision problems from the perspective of NGOs and other not-for-profit organizations, which leaves out the profit-driven interests of the private sector participants. One exception is Lodree and Taskin (2009), which is discussed at the beginning of this section, and another exception is

Lodree and Taskin (2008). Lodree and Taskin (2008) introduce a framework within the context of inventory control related to preparing for a potential humanitarian relief operation or supply chain disruption. The framework is analogous to buying an insurance policy, and the objective is to minimize expected inventory-related costs or provide a specified level of service. Beamon and Kotleba (2006) also address inventory control associated with humanitarian relief, but their context is long-term relief operations in contrast to our context, which is immediate response. Although the model in Beamon and Kotleba (2006) is introduced from the perspective of an NGO or government organization, it is also applicable to a profit-driven, private sector firm's perspective in a humanitarian relief situation. Other contributions to the humanitarian logistics literature include conceptual frameworks (Van Wassenhove, 2006; Kovács and Spens, 2007; Beamon and Balcik, 2008), facility location (Balcik and Beamon, 2008), and last mile distribution (Özdamar *et al*, 2004; Balcik *et al*, 2008).

13.4 Bayesian decision model

In this section, the single-supplier/multi-retailer supply chain inventory problem introduced in the 'Problem description' section is formulated as a sequential Bayesian decision problem, which entails formulating appropriate loss, risk, and Bayes risk functions. A solution to the model specifies the supplier's optimal production quantity (a one-time decision) and the optimal decision period (ie, the number of hurricane forecast updates that should be obtained before the production decision is made). Note that production quantity and decision period are the only two decision variables; optimal allocation of inventory among the retailers in the event of shortage is not considered.

13.4.1 Assumptions

The derivations of the loss, risk, and Bayes' risk functions are derived based on the following assumptions.

Assumption 1 *The production manager is allowed at most one opportunity to implement a production quantity decision. In addition, once the production quantity decision is made, it is irreversible.*

This assumption will allow the manufacturer's multi-retailer inventory control problem to be formulated as a sequential Bayesian decision problem with multiple delivery modes as in Choi *et al* (2004) and Sethi

et al (2005). Consequently, this modelling approach will facilitate the use of results from the Bayesian decision theory to help solve the inventory control problem with forecast updates. In general, this modelling framework allows the decision-maker to obtain sample data so that he can make an informed decision, but he is limited to making at most one decision during the planning horizon. Although allowing not more than one production decision can be construed as a limitation in practice, this assumption is applicable to certain real-world environments. For instance, consider a production process for emergency supplies characterized by time- and resource-intensive activities such as reconfiguring layouts in various parts of the facility and generating new production schedules. In addition, outsourcing may be required in order to access certain expertise or excess capacity. Under these circumstances, it is unlikely that a production decision can be reversed or altered after it is implemented.

Assumption 2 *The supplier's demand is a random variable X, where X is a function of the observed storm's intensity along its path.*

Let X_i, $i = 1, \ldots, n$, be a random variable that represents the demand of retailer i, where n is the number of retailer locations whose demand can potentially be affected by the observed storm. Then X is taken to be a convolution of the random variables X_i. Each X_i depends on the intensity of the storm surrounding location i, which includes the possibility that the storm does not threaten the location at all. Let H_i be a Bernoulli random variable that represents whether location i will experience hurricane force winds during the planning horizon. Then, Assumption 1 implies the existence of functions ϕ_i: $H_i \mapsto X_i$ so that $X_i = \phi_i(H_i)$, and $X = \sum X_i = \sum \phi_i(H_i)$.

Also note that the manufacturer's problem is interpreted as a single newsvendor problem with demand represented as a convolution of retailer demands X_i, as opposed to n separate newsvendor problems with respective demands X_i for $i = 1, \ldots, n$. In other words, centralized control is considered as opposed to decentralized control because it is known that centralized inventory control typically outperforms decentralized control, eg, (Eppen, 1979) and (Chen and Lin, 1989).

Assumption 3 *Two classes of demand are considered at each retailer location: demand associated with hurricane force winds and demand that corresponds to no hurricane force winds.*

Assumption 2 implies that the demand distribution at location *i* is one of two categories. Let Y_i denote demand at location *i* if no hurricane force winds are experienced and Z_i be demand at location *i* if hurricane force winds are experienced, where both Y_i and Z_i are random variables for each $i = 1, \ldots, n$. Then $X_i \in \{Y_i, Z_i\}$. This represents the initial approach to modelling the impact of an observed storm on each retailer's demand, and hence the supplier's demand.

Assumption 4 *Let* $\vec{H} = (H_1, \ldots, H_n)$ *be a multivariate Bernoulli random variable such that* $H_i = 1$ *indicates hurricane force winds at location i during the planning horizon, and* $H_i = 0$ *otherwise, where* $i = 1, \ldots, n$. Let $\vec{P} = (P_1, \ldots, P_n)$ *be the parameter vector associated with* $\vec{H}$, *where* $P_i = \Pr\{H_i = 1\}$, *and* $1 - P_i = \Pr\{H_i = 0\}$. *Also, let* $\vec{A}_t$ *denote a random vector of storm attributes (namely location, intensity, and radius) with realization vector* $\vec{\alpha}_t$, *where T is the fixed number of decision periods and* $t = 1, \ldots, T$. *Then*

1. $\vec{P}$ is a random vector with prior density $\vec{h}(\vec{p}) = [h_1(p_1), \ldots, h_n(p_n)]$.
2. Let $P_{it} = \Pr\{H_i = 1 | A_{t-1} = a_{t-1}, P_{i,t-1} = p_{i,t-1}\}$, and $\vec{P}_t = (P_{1t}, \ldots P_{nt})$. Then $\vec{P}_t$ is the posterior probability vector with respect to $\vec{P}_{t-1}, t = 1, \ldots, T$, and the corresponding posterior distribution is $\vec{h}_t(\vec{p} \mid \vec{\alpha}_{t-1}, \vec{p}_{t-1})$. Similarly, the likelihood probabilities are $\Pr\{A_{t-1} = a_{t-1}, P_{i,t-1} = p_{i,t-1} | H_i = 1\}$ with likelihood densities $\vec{h}_t(\vec{\alpha}_{t-1}, \vec{p}_{t-1} | \vec{p})$.
3. The likelihood and posterior densities, $\vec{h}_t(\vec{\alpha}_t | \vec{p})$ and $\vec{h}_t(\vec{p} \mid \vec{\alpha}_t)$, respectively, are based on hurricane prediction updates that are given every 12 h after a tropical depression or disturbance is initially observed. Thus, *t* progresses in 12-h intervals.

The NHC publishes certain hurricane forecasts, namely the *tropical cyclone wind speed probability product* (refer to the 'Probabilistic hurricane prediction model' section), typically in 12-h intervals as an observed storm evolves. Each forecast is generated based on an observation of $\vec{\alpha}_t$, given that $t - 1$ forecasts have been published before. Thus, each NHC forecast update represents a posterior probability calculation, which is represented mathematically based on the above notation as $\vec{P}_t = (P_{1t}, \ldots, P_{nt})$, which is calculated from the posterior density $\vec{h}_t(\vec{p} \mid \vec{\alpha}_t)$.

Assumption 5 *Demand realization happens at each location i exactly 5 days (120 h) after the storm is initially observed.*

Assumption 5 allows us to specify that the length of the supplier's decision horizon is known with certainty. That is, $T = 120$. In actuality, the length of the supplier's decision horizon is also uncertain. However, useful results from the sequential Bayesian decision theory are leveraged to model and solve the inventory control problem by assuming that T is known with certainty. This assumption also allows the manufacturer's demand to be represented as a convolution of the retailers' demands as described in Assumption 1. If retailer demand realizations occurred at different times, which is more representative of what happens in practice, then our approach would need some modifications. This assumption represents a shortcoming of our model that can be explored through future research.

Assumption 6 *Let c_t be the production cost associated with making a decision after t forecast updates. Then $c_t \leq c_{t+1}$ for all $t = 0, 1, \ldots, T-1$.*

Assumption 4 relates to the difficulty of implementing a production decision during the later stages of the planning horizon. In reality, the problem with waiting for very accurate storm information is that there may not be enough time to meet demands because of limited production capacity. If the decision about the target inventory level is determined earlier, then the supplier can schedule production over a few days without paying a premium for additional capacity. Define $C_j = c_{j+1} - c_j$ for $j = 0, 1, \ldots, T-1$ and $c_0 = 0$. Then C_j can be interpreted as the premium that the supplier has to pay for additional capacity in order to reach target inventory levels such that finished goods can be shipped to the disaster area immediately after the hurricane strikes.

13.5 Single period loss function

The proposed loss function is a generalization of the loss function introduced in Lodree and Taskin (2009), which is a generalization of the newsboy problem. Specifically, the loss function associated with period (ie, forecast update) t considers a newsboy problem with hurricane demand Z_i at location i with probability P_{it}, and a newsboy problem with demand Y_i for location i with probability $1 - P_{it}$ (see Assumption 1; also, recall from Assumption 3 that P_{it} is the posterior probability given by $p_{it} = \Pr\{H_{it} = 1 | A_{it} = a_{it}\}$).

Since demands at retailer locations are assumed to be independent random variables, the loss function, denoted by $L(Q)$, is then formed based on a convolution of these random variables. The probability of observing a specific demand convolution k during period t is defined as the *scenario*

probability, denoted by q_{kt}. These scenario probabilities are calculated using the probability vector $\vec{p}_t = \{p_{1t}, p_{2t}, \ldots, p_{nt}\}$, and the multiplication rule resulting in 2^n possible scenarios (refer to Table 13.1 for illustration with $n=2$). Note that exactly one out of the 2^n scenarios occurs during each update period *t*. The resulting expected loss function associated with period *t* is

$$L_t(Q_t) = \sum_{k=1}^{2^n} q_{kt} \cdot NB_{kt}, \tag{1}$$

where NB_{kt} is the expected cost function for the newsboy given by

$$NB_{kt} = Q_t c_t + h\int_0^{Q_t} (Q_t - x_{kt}) f_{kt}(x_{kt})\mathrm{d}x_{kt} + s\int_{Q_t}^{\infty} (x_{kt} - Q_t) f_{kt}(x_{kt})\mathrm{d}x_{kt}. \tag{2}$$

In Equation (2), c_t is the supplier's unit production cost during period *t*, *h* the unit holding cost for leftover inventory after demand realization, *s* the unit shortage cost for lost sales observed after demand realization, X_{kt} the convolution of Y_i and Z_j associated with scenario *k*, and $f_{kt}(x_{kt})$ is the (posterior) density of X_{kt}.

13.6 Bayes risk function

Since the loss function $L_t(Qt)$ given by Equation (1) corresponds to a single forecast period *t*, the dynamic evolution of the storm is not yet accounted for with respect to the supplier's decision. The *risk function* (eg, see Berger, 1985), denoted by $R(\vec{P}, d)$ where $\vec{P} \in \mathbb{R}^n$ and *d* is a sequential decision procedure (defined later), considers sequential information updates by introducing a *sequential random sample* $\mathbf{P}^T$, where

$$\mathbf{P}^T = \begin{pmatrix} P_{11} & P_{11} & \cdots & P_{1T} \\ P_{21} & P_{22} & \cdots & P_{2T} \\ \cdots & \cdots & \cdots & \cdots \\ P_{n1} & P_{n2} & \cdots & P_{nT} \end{pmatrix}. \tag{3}$$

Table 13.1 Demand scenario probabilities

Scenario k	Scenario probability q_{kt}	Retailer1	Retailer2	Convoluted demand X_k
1	$q_{1t} = p_{1t} \cdot p_{2t}$	Z_1	Z_2	$Z_1 + Z_2$
2	$q_{2t} = (1 - p_{1t}) \cdot p_{2t}$	Y_1	Z_2	$Y_1 + Z_2$
3	$q_{3t} = (1 - p_{1t}) \cdot (1 - p_{2t})$	Y_1	Y_2	$Y_1 + Y_2$
4	$q_{4t} = p_{1t} \cdot (1 - p_{2t})$	Z_1	Y_2	$Z_1 + Y_2$

Here, P_{it} is the updated probability during forecast period t that retailer location i will experience hurricane force winds during the planning horizon. The sequential sample $\mathbf{P}^t$ allows the decision-maker the opportunity to obtain more information about the random parameter vector $\vec{H}$ (see Assumption 3) before making a decision. However, in the Bayesian decision theory, sampling is not free. (Refer to Berger (1985) for background information about sequential decision processes.) Formally, let $\vec{P}_t = (P_{1t}, \ldots, P_{nt})^T$, which corresponds to a column vector in $\mathbf{P}^T$. Then the cost of observing $\vec{P}_t$ is C_t. (Note: it is assumed that the columns of $\mathbf{P}^T$ are independent.)

For our multi-retailer inventory problem, the stopping time t is interpreted as the number of forecasts such that sampling is stopped and a decision δ_t is made as opposed to observing the next sample $\vec{p}_{t+1}$. In this case, the cost due to sampling is $\sum_{j=1}^{T} C_j$.

Now define the *sequential decision procedure* (eg, Berger, 1985) as $d = (\tau, \delta)$. Taking the expectation of the loss function with respect to X yields the risk function given by

$$\begin{aligned} R(\vec{P}, d) &= E\left[L(\vec{P}, \delta_t, (\mathbf{P}^T), t)\right] \\ &= P(t=0) \cdot L(\vec{P}, \delta_0, 0) \\ &\quad + \sum_{T=1}^{t} \int_{\Lambda} L(\vec{P}, \delta_T(\mathbf{p}^T), T)\, \mathrm{d}H_T(\mathbf{p}^T | \vec{P}) \\ &\quad + \sum_{T=1}^{\infty} \sum_{j=1}^{T} C_j P(t=T). \end{aligned} \tag{4}$$

The prior density function $\vec{h}(\vec{p})$ associated with $\vec{P}$ is updated at each forecasting period t. The Bayes risk function of the problem is then obtained by applying the expected value operator with respect to $\vec{P}$ as follows:

$$r(\vec{h}_t, d, t) = \mathbb{E}[R(\vec{P}, d)], \tag{5}$$

where $\vec{h}_t$ is the posterior density of $\vec{P}$ after observing the sequential random sequence $\mathbf{P}^t$.

The sequential statistical decision problem is then formulated such that the optimal decision and stopping rule d can be determined:

$$r(\vec{h}_t, t) = \inf_{d} r(\vec{h}_t, d, t). \tag{6}$$

In order to determine the ordering quantity Q, the Bayes risk function of the problem $r(\vec{h}_t, d, t)$ is formalized by revising the loss function given in Equation (1) such that the posterior predictions of $\vec{h}_t(\vec{P}_t | \vec{\alpha}_t)$, and the

cost of observing a sequential sample $\mathbf{P}^t$ associated with $\vec{P}$ are incorporated into the function. As a result, $r(\vec{h}_t, d, t)$ is expressed as

$$r(\vec{h}_t, d, t) = \sum_{k=1}^{k=2^n} q_{kt} \cdot NB_k^t + \sum_{j=1}^{t} C_j \quad t = 1, \ldots, T. \tag{7}$$

Then, the sequential problem is bounded such that

$$r_T(\vec{h}_t, t) = \inf_d r(\vec{h}_t, d, t). \tag{8}$$

13.7 Solution methodology

We now describe the methodology used to determine the production quantity Q_t and production initiation period t^* that minimize the expected loss associated with production, overstocking, and understocking. Specifically, the optimal *optimal sequential decision procedure* is presented, which entails specifying the optimal decision rule and optimal stopping rule. Theorems 1 and 2 present the decision and stopping rules, respectively.

Theorem 1 (Lodree and Taskin, 2008) *Let $F_k(x_k)$ be the cumulative distribution function of X_k, where k = 1, ..., 2^n. Then the optimal decision $\delta_t = Q_t$ that minimizes the loss function $L_t(Q_t)$ with unit production cost c_t satisfies*

$$\sum_{k=1}^{k=2^n} q_{kt} \cdot F_k(Q_{t^*}) = \frac{s-c}{s+h}. \tag{9}$$

Theorem 2 (DeGroot, 1970) *Define $r_0(\vec{h}_t, t)$ as the minimum Bayes risk associated with making an immediate decision in period t, and $r_{T-t}(\vec{h}_t, t)$ as the minimum Bayes risk associated with T – t more observations. Then among all sequential decision procedures in which not more than T observations can be taken, the following procedure is optimal: If $r_0(\vec{h}_t, 0) \leq r_T(\vec{h}_t, 0)$; a decision δ_0 is chosen immediately without any observations. Otherwise, $\vec{P}_1$ is observed. Furthermore, for t = 1,2, ..., T−1, suppose the sequential random sample $\mathbf{P}^t$ has been observed. If $r_0(\vec{h}_t, t) \leq r_{T-t}(\vec{h}_t, t)$, a decision δ_t is chosen immediately without further observations. Otherwise, $\vec{P}_{t+1}$ is observed. If sampling has not been terminated earlier, it must be terminated after $\vec{P}_T$ is observed.*

Theorems 1 and 2 suggest the following approach to calculating the optimal decision procedure. First, determine the optimal production initiation period, t^*, based on the approach described in Theorem 2. Then apply the decision rule shown in Theroem 1 to obtain the optimal production quantity, Q_{t^*}.

13.8 Probabilistic hurricane prediction model

The scenario probabilities, q_{kt}, described in the previous section are based on the NHC's *tropical cyclone wind speed probability product* (NHC08). Specifically, the optimal stopping rule uses the scenario probabilities q_{kt} associated with an observed wind speed probability forecast issued by the NHC, but also requires an estimate of the wind speed probabilities associated with the ensuing information update, denoted by $q_{k,t+1}$. Therefore, it is not possible to apply the Bayesian decision framework by simply incorporating observed wind speed probabilities into a decision model; one has to anticipate the NHC's next wind speed probability prediction. For this reason, it is necessary to reproduce the NHC's wind speed probability forecasts, which is the subject of this section.

We digress for a moment to explain why the NHC's wind speed probability product is selected as the hurricane prediction model that is incorporated into the Bayesian inventory model formulated in the previous section. First, our intention is to improve upon the probabilistic hurricane prediction model that was used in an earlier paper, namely Lodree and Taskin (2009). Specifically, the hurricane prediction model in Lodree and Taskin (2009) only accounts for intensity and ignores the hurricane's track. In addition, their purely statistical prediction model was not developed by meteorologists, and it is not used by the NHC. Demands for emergency supplies are likely to be influenced by track and intensity, both of which are included in wind speed probability forecasts. Furthermore, the wind speed probability forecast is an official NHC product and represents one of their most recent and widely used developments (NHC08). These probabilities provide more meaningful hurricane forecasts compared to previous forecast products by considering the uncertainties associated with both intensity and size forecasts of cyclones. While the previous models forecast the track of the cyclone centre, the wind speed probabilities provide hurricane predictions for individual locations. In addition, the wind speed probabilities are predicted for a 5-day forecasting horizon, which is longer than those of the previous models.

Tropical cyclone wind speed probability products calculate the probabilities of sustained surface wind speeds of at least 74 mph (hurricane force) at different locations. More specifically, each graphic presents cumulative probabilities of hurricane force wind speeds across multiple locations during cumulative 12-h intervals (ie, 0–12 h, 0–24 h, ..., 0–120 h), and extends through a 5-day forecast. An example graphic of this product, which is readily available on the NHC website, is shown in Figure 13.1. Note that although the focus of this paper is hurricane

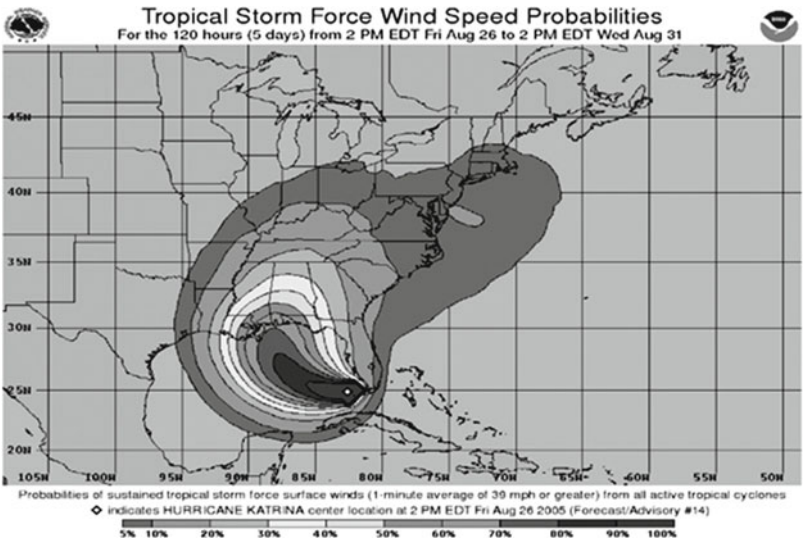


Figure 13.1 NHC's tropical cyclone wind speed probability product

force winds, the NHC also generates wind speed probability forecasts for tropical storm force wind speeds.

Wind speed probabilities are based on official track and intensity data, as well as their corresponding forecast errors issued by the NHC. These forecast errors are determined by comparing the official forecasts with the best track database, which is a record of the NHC's post-storm analysis of all historical storm data. These forecasts and forecast errors can be found from the NHC's public resources, that is, the Tropical Cyclone Forecast/Advisory and NHC Official Forecast Error Database. In addition to forecasting the centre position and intensity of a tropical cyclone, variability in size (ie, wind radii) is also incorporated into forecasts via the *climatological wind radii forecast model* introduced in Gross *et al* (2004). Specifically, the wind radii forecast model is given by

$$V(r,\lambda) = (V_m - \alpha)\cdot(r_m/r)^x + \alpha\cdot[\cos(\lambda - \lambda_o)]. \tag{10}$$

Here, $V(r, \lambda)$ is the wind speed threshold; r is the radius from the storm centre 'wind radius' (nmi); λ is the angle measured counterclockwise starting from a direction 90° to the right of the storm motion; V_m is the maximum wind (mph), r_m is the radius of maximum wind (nmi); x is a size parameter (non-dimensional); and α is an asymmetry parameter (kt)

that is a function of the storm speed of motion. Gross *et al* (2004) analyse the climatological model by fitting x, r_m, α, λ_o to the NHC wind radii forecasts for the 1988–2002 Atlantic storms. They determined that r_m and x can be estimated in terms of V_m and latitude θ, and also that α can be expressed as a function of the storm speed of motion c as shown below:

$$r_m = 35.37 - 0.111 \cdot V_m + 0.570 \cdot (\theta - 25), \tag{11}$$

$$x = 0.285 + 0.0028 \cdot V_m, \tag{12}$$

$$\alpha = 0.337 \cdot c - 0.003 \cdot c^2. \tag{13}$$

Wind radii samples are simulated using the error distribution of the x parameter. The initial value of the error in X is chosen as the difference between the value from the above climatological model and the best fit values to the observed radii at each storm quadrant. Various error values at forecasting times are then determined as a linear combination of the initial value and a random component to develop the forecast error distribution of wind radii. The wind speed probability maps are generated by simulating from these forecast error distributions. In addition, the wind speed probabilities are estimated as the fraction of the number of grid points representing the locations within the radius of hurricane force wind speeds. Then, the scenario probabilities q_{kt} are predicted by applying the multiplication rule to the combinations of wind speed probabilities.

The well-known Climatology and Persistence (CLIPER) model is used to predict the storm centre coordinates at each forecasting period. The CLIPER model uses two sets of regression equations to predict the storm's track, where each predictand is either the zonal or meridional displacement observed at time t. These regression equations are given by

$$\begin{aligned} Z = {} & \beta_0 + \beta_1 \cdot U + \beta_2 \cdot LAT + \beta_3 \cdot (LON \cdot V) \\ & + \beta_4 \cdot (LAT \cdot V) + \beta_5 \cdot (LON \cdot DAY) \beta_6 \cdot (LAT \cdot U) \\ & + \beta_7 \cdot (LAT \cdot LAT) + \beta_8 \cdot (DAY \cdot U) \\ & + \beta_9 \cdot (U \cdot V) + \beta_{10} \cdot (INT \cdot V), \end{aligned} \tag{14}$$

$$\begin{aligned} M = {} & \beta_0 + \beta_1 \cdot V + \beta_2 \cdot U + \beta_3 \cdot (INT \cdot U) \\ & + \beta_4 \cdot (LAT \cdot INT) + \beta_5 \cdot (INT \cdot V) + \beta_6 \cdot (U \cdot V) \\ & + \beta_7 \cdot (DAY \cdot U) + \beta_8 \cdot (LON \cdot DAY) + \beta_9 \cdot (LAT \cdot V). \end{aligned} \tag{15}$$

Here *LAT* is the initial latitude, *LON* is the initial longitude, *INT* is the initial intensity, *DAY* is the initial day number, *U* is the initial zonal motion, and finally *V* corresponds to the initial meridional motion.

13.9 Numerical examples

To implement the proposed solution methodology in practice, the production manager would simply observe the NHC's wind speed probability forecasts at the beginning of each period, and insert these probabilities and other relevant data (eg, mean demand, unit production cost, etc) into the Bayesian decision model. Although the NHC updates wind speed probability forecasts regularly during the 12-h intervals, an update can occur at any time if there is a major change in the storm's development. Once the wind speed probabilities and other relevant data are entered into the model, the result will indicate whether the production manager should initiate emergency supply production during the current 12-h period, or wait for the next hurricane information update. Of course, the output will also indicate the corresponding optimal production quantity. Note that the proposed model allows the production manager opportunities to initiate production each time an updated wind speed probability forecast is issued by the NHC. However, the model only allows the production quantity decision to be made during at most one forecast period, and the model does not allow the decision to be altered once production has been initiated.

13.10 Hurricane prediction and simulation

In this section, a fictitious storm is simulated to illustrate the proposed solution methodology. First, it is assumed that a storm develops at a hypothetical location, say an initial zonal (longitude) coordinate of $x = 75$ °W and an initial meridional (latitude) coordinate of $y = 25$°N. Next, the CLIPER model is used to simulate the fictitious storm's track by solving the CLIPER regression equations for a large sample size (65 years, which is justified in Aberson (1998) at each forecasting period). For illustrative purposes, it is assumed that forecasts are updated in 30-h intervals so that there are exactly four periods during the 5-day life cycle of the storm (refer to Assumption 5). Table 13.2 illustrates the simulated track produced by the CLIPER model (note that in practice, these outputs would be revealed sequentially).

The coordinates associated with each position shown in Table 13.2 are used to generate a wind speed probability forecast for each 30-h

Table 13.2 Simulated hurricane track based on the CLIPER model

Forecast time	Zonal displacement	Meridional displacement	Storm centre coordinate (°W, °N)
30	3.76	0.33	(78.76, 25.33)
60	6.30	–1.82	(85.06, 23.51)
90	3.27	–0.69	(88.33, 22.82)
120	4.50	–1.73	(92.83, 21.09)

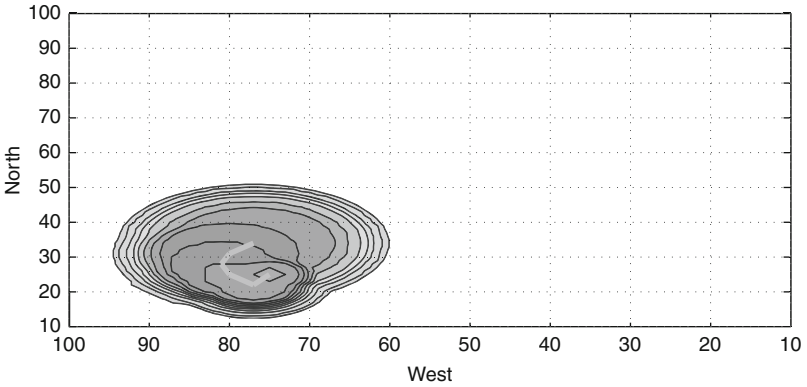


Figure 13.2 Wind speed probability map at $t = 0$ h

forecast period. In particular, a vector $\vec{p}_t = (p_{1t} \cdots, P_{nt})^T$ is computed for each period $t = 1, \ldots, 4$, which is equivalent to specifying the sequential sample $\mathbf{P}^4$ given by Equation (3) (again, each vector $\vec{P}_t$ would be disclosed sequentially in practice). For an illustration of the example problem 1, Figures 13.2 and 13.3 depict the wind speed probability forecasts in graphical form. Figure 13.2 corresponds to the initial wind speed probability forecast, whereas Figure 13.3 is associated with the last hurricane prediction. These graphs were created in Matlab™. Note that the light-coloured line corresponds to the forecasted storm track, and the dark-coloured line represents the realized storm path up to that forecasting period. Figure 13.2 shows only the forecasted track since it is formed during the initial forecasting period, $t = 0$. More specifically, it gives the prior wind speed probabilities, which are updated in 30-h intervals to predict the posterior wind speed probabilities maps, as more hurricane information becomes available. In fact, each wind speed probability

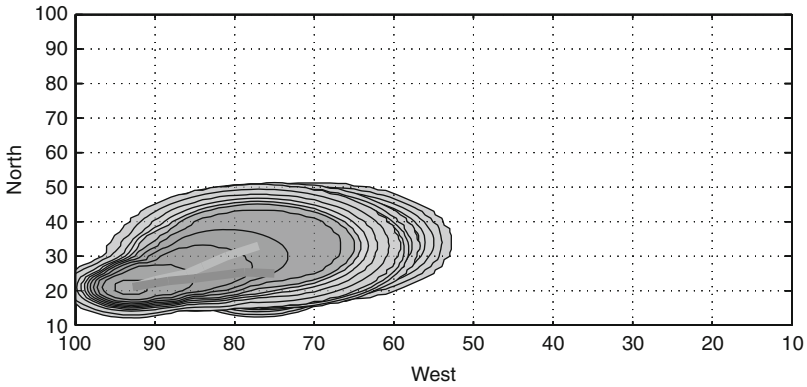


Figure 13.3 Wind speed probability map at $t = 120$ h

map corresponds to the prior distribution of its subsequent one. Similar maps can be created for the other example problems.

13.11 Example decision problems

The following data related to the supplier's inventory control problem are used to solve each of the 10 example problems corresponding to the hurricane data shown in Table 13.2: $c_0 = 20$, $c_{t+1} = c_t + 1$ for $t = 0, \ldots, 3$, $s = 100$, $h = 15$, and $n = 5$. In addition, the demand information at each retailer location is shown in Table 13.3.

The Bayesian decision process was implemented numerically using Mathematica 5.2™ in order to obtain decision rules (ie, decision period and production quantity) for each of the 10 example problems. The results, which are presented in Table 13.4, suggest that the production manager should postpone the production decision and wait for several hurricane forecast updates whenever wind speed probabilities indicate low probabilities for hurricane activity at multiple retailer locations. The optimal decision rules shown in Table 13.4 also suggest larger production quantities when wind speed probability forecasts indicate the possibility of hurricane force winds at all locations relative to the examples when hurricane activity is likely at only a few locations. These findings are intuitive: when hurricane activity is predicted for several locations and the optimal production decision requires large production quantities, then the corresponding optimal decision period is during the early part of the planning horizon. This makes sense because larger production quantities should be initiated early on, so

Table 13.3 Demand information

Location i	Regular demand Y_i	Hurricane demand Z_i
1	*N*(500, 25)	*N*(1000, 100)
2	*N*(550, 30)	*N*(1500, 125)
3	*N*(600, 35)	*N*(2000, 150)
4	*N*(650, 40)	*N*(2500, 175)
5	*N*(700, 45)	*N*(3000, 200)

Table 13.4 Example results

Example	Decision period t*	Stocking quantity Qt	Expected cost EC
1	1	8567.79	$198 995
2	1	6299.83	$140 188
3	1	7685.85	$203 198
4	1	8619.32	$174 569
5	2	6751.26	$196 656
6	3	5295.66	$155 180
7	3	7156.1	$132 302
8	3	5829.81	$122 755
9	4	4884.36	$112 849
10	4	6250.24	$145 707

that production can be spread out over several periods as opposed to starting production during the later periods and paying a premium for the excess capacity needed to achieve the specified production quantity.

13.12 Model validation: An insurance policy framework

Lodree and Taskin (2008) introduced an approach to preparing for demand surge caused by potential disaster relief activity or supply chain disruption within the context of inventory control. Their framework interprets proactive preparation as an insurance policy and they determine optimal premiums and payoffs in terms of production quantities associated with variations in the newsvendor problem. In this paper, we extend the insurance policy framework to account for the multi-location inventory control problem. The approach is based on the following definitions.

Definition 1 (Lodree and Taskin, 2008) The expected insurance premium associated with ordering quantities in the newsvendor problem with demand disruptions is the expected marginal cost incurred if the insurance policy is purchased and no disruption occurs.

Definition 2 (Lodree and Taskin, 2008) The expected payoff associated with ordering quantities in the newsvendor problem with demand disruptions is the expected marginal benefit realized if the insurance policy is not purchased and the disruption does occur.

The idea is to determine the expected benefit associated with optimal production quantity from the solution procedure described in the 'Solution methodology' section. Expected benefit is defined as the difference between expected payoff and expected premium. Specifically, let $f_1(x_1)$ correspond to the convolution Y_i, $i = 1, \ldots, n$ (ie, scenario k= 1 with regular demand at each location), and let Q_1 represent the corresponding optimal production quantity. Similarly, let $f_k(x_k)$ be the convolution of the random variables that correspond to the demand distribution under scenario $k \in \{2, 3, \ldots, 2n\}$. Then the above definitions suggest the following:

$$E[\text{Premium}] = h\int_0^{Q_D}(Q_D - x_1)f_1(x_1)\mathrm{d}x_1 - h\int_0^{Q_1}(Q_1 - x_1)f_1(x_1)\mathrm{d}x_1 + c(Q_D - Q_1),$$

$$E[\text{Payoff}] = s\int_{Q_1}^{\infty}(x_k - Q_1)f_k(x_k)\mathrm{d}x_k - s\int_{Q_D}^{\infty}(x_k - Q_D)f_k(x_k)\mathrm{d}x_k,$$

$$\text{where } k \in \{2, 3, \ldots, 2^n\}$$

$$E[\text{Benefit}] = E[\text{Payoff}] - E[\text{Premium}]. \tag{16}$$

Applying these concepts to the data associated with example problem 1 yields the results shown in Table 13.5. Note that the results shown in Table 13.5 are based on realization of the scenario in which regular demand is observed at location 1 and hurricane demand is observed at the other four locations. These computations suggest that the production manager can expect to lose \$194 597 if he prepares for hurricane demands and the hurricane does not affect any retailer locations (ie,

Table 13.5 Example 1 calculations based on the insurance policy framework

Model	Stock level		Financial		
	Q_D	% Increase	E[Payoff]	E[Premium]	E[Benefit]
Traditional	3006.77	—	—	—	—
Proposed	8567.79	184.95	556 102	194 597	361 505

Table 13.6 Impact of ρ and VaDR on production quantities

p	Q^*_D	VaDR	Q^*_D
0.15	10 028	50 000	4436
0.25	10 018	100 000	5865
0.35	10 010	175 000	8007
0.45	10 003.47	230 000	9579
0.50	10 000	250 000	10 150
0.55	9996	325 000	12 293
0.60	9993	375 000	13 722
0.65	9989	400 000	14 436

insurance premium). On the other hand, he can expect to be \$556 102 better off by preparing for hurricane demands if the storm affects the locations mentioned under the above scenario compared to taking no precautions and not preparing (ie, insurance payoff). The expected benefit is the manufacturer's expected increase in net worth given by \$556 102–\$194 597 = \$361 505. Similar conclusions can be inferred based on realizations of the other scenarios.

Lodree and Taskin (2008) also introduced the Value at Disruption risk (VaDR) concept, which is an extension of Value at Risk. VaDR is a measure of the expected financial loss associated with a specified service level and is defined as E[Premium] that corresponds to Q_D, where Q_D satisfies

$$p = \Pr\{X_n \geqslant Q_D | X_i = Z_i\} = 1 - F_n(Q_D), \quad \forall i = 1, \ldots n. \tag{17}$$

Here, ρ is the predetermined stockout probability associated with hurricane demands at each retailer location, and $F_n(Q_D)$ is the cumulative distribution function corresponding to hurricane demands at all n retailer locations. This approach generates the production quantity Q_D that achieves target service levels as opposed to maximizing expected benefit. Table 13.6 demonstrates the effect of ρ and VaDR on the production decisions. In particular, the calculations shown in Table 13.6 verify the intuitive result that inventory level decreases as the specified stockout probability increases, and also that VaDR increases as inventory level increases.

13.13 Summary

This paper addresses an inventory management problem encountered by a private sector production facility that prepares for potential humanitarian relief efforts in response to an observed hurricane. A sequential

Bayesian decision model that incorporates information updates associated with the NHC's wind speed probability forecasts is introduced to address the problem from the manufacturer's perspective. The solution to the decision problem specifies the facilities of optimal production quantity and the optimal number of forecast updates that should be observed before making this one-time decision. The effectiveness of the proposed approach is explored based on an existing framework for evaluating the expected costs and benefits associated with proactive preparation for potential disaster relief activities within the context of inventory control.

We believe that certain issues need to be addressed in future research so that the proposed approach could be effective in practice, namely allowing uncertainty in the number of forecast-updating periods, allowing adjustments in the inventory decision during the planning horizon, considering allocation of emergency supplies among multiple retailers, and considering transportation decisions and costs.

References

Aberson SD (1998). Five-day tropical cyclone track forecasts in the North Atlantic basin. *Weather Forecast* **13**: 1005–1015.

Aronis K-P, Magou I, Dekker R and Tagaras G (2004). Inventory control of spare parts using a Bayesian approach: A case study. *Eur J Opl Res* **154**: 730–739.

Balcik B and Beamon BM (2008). Facility location in humanitarian relief. *Int J Logist: Res Appl* **11**: 101–121.

Balcik B, Beamon BM and Smilowitz K (2008). Last mile distribution in humanitarian relief. *J Intell Transport Syst* **12**(2): 51–63.

Beamon BM and Balcik B (2008). Performance measurement in humanitarian relief chains. *Int J Publ Sec Mngt* **21**(1): 4–25.

Beamon BM and Kotleba SA (2006). Inventory modelling for complex emergencies in humanitarian relief operations. *Int J Logist: Res Appl* **9**(1): 1–18.

Berger JO (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Institute of Statistics and Decision Sciences, Duke University: Durham, NC.

Chen MS and Lin CT (1989). Effects of centralization on expected costs in a multi-location newsboy problem. *J Opl Res Soc* **40**: 507–602.

Choi T-M and Li D (2003). Optimal two-stage ordering policy with Bayesian information updating. *J Opl Res Soc* **54**: 846–859.

Choi T-M, Li D and Yan H (2004). Optimal single ordering policy with multiple delivery modes and Bayesian information updates. *Comput Opns Res* **31**: 1965–1984.

Choi T-M, Li D and Yan H (2006). Quick response policy with Bayesian information updates. *Eur J Opl Res* **170**: 788–808.

DeGroot MH (1970). *Optimal Statistical Decisions*. John Wiley and Sons, Inc: Hoboken, NJ.

Ding X, Puterman ML and Bisi A (2002). The censored newsvendor and the optimal acquisition of information. *Opns Res* **50**: 517–527.

Dvoretzky A, Kiefer J and Wolfowitz J (1952). The inventory problem: II. Case of unknown distributions of demand. *Econometrica* **20**: 450–466.

Eppen G (1979). Effects of centralization on expected costs in a multi-location newsboy problem. *Mngt Sci* **25**: 498–501.

Gross J, DeMaria M, Knaff JA and Sampson CR (2004). A new method for determining tropical cyclone wind forecast probabilities. *Hurricane Trop Meteor Soc* **26**: 425–426.

Kovács G and Spens KM (2007). Humanitarian logistics in disaster relief operations. *Int J Phys Distrib Logist Mngt* **37**(2): 99–114.

Lodree Jr EJ and Taskin S (2008). An insurance risk management framework for disaster relief and supply chain disruption inventory planning. *J Opl Res Soc* **59**: 674–684.

Lodree Jr EJ and Taskin S (2009). Supply chain planning for hurricane response with wind speed information updates. *Comput Opns Res: Special Issue on Disaster Recovery Planning*. **36**: 2–15.

NHC08 (2008). http://www.nhc.noaa.gov/aboutnhcprobs.shtml, accessed 3 August.

Özdamar L, Ekinci E and Küçükyazici B (2004). Emergency logistic planning in natural disasters. *Ann Opns Res* **129**: 217–245.

Phillips Z (2007). FEMA looks to private sector for disaster provisions. http://www.govexec.com/dailyfed/0807/082407mag2.htm, accessed 22 July.

Regnier E (2005). Doing something about the weather. *Int J Mngt Sci* **36**(1): 22–32.

Regnier E and Harr PA (2006). A dynamic decision model applied to hurricane landfall. *Weather Forecast* **21**: 764–780.

Regnier E and Harr PA (2008). Public evacuation decision and hurricane track uncertainty. *Weather Forecast* **54**(1): 16–28.

Sethi SP, Yan H and Zhang H (2005). Inventory and supply chain management with forecast updates. In: *International Series in Operations Research & Management Science*, Springer Science and Business Media Inc: New York.

Townsend F (2006). *The federal response to Hurricane Katrina: Lessons learned*. Report to the President of the United States.

Van Wassenhove LN (2006). Humanitarian aid logistics: Supply chain management in high gear. *J Opl Res Soc* **57**: 475–489.

Wu J (2005). Quantity flexibility contracts under Bayesian updating. *Comput Opns Res* **32**: 1267–1288.

Zhang J-L and Chen J (2006). Bayesian solution to pricing and inventory control under unknown demand distribution. *Opns Res Lett* **34**: 517–524.

Zhu K and Thonemann UW (2004). An adaptive forecasting algorithm and inventory policy for products with short life cycles. *Nav Res Log* **51**: 633–653.

14

Using Simulation to Improve the Blood Supply Chain

K. Katsaliaki and S. C. Brailsford
University of Southampton, Southampton, Hampshire, UK

This case study is concerned with analysing policies for managing the blood inventory system in a typical UK hospital supplied by a regional blood centre. The objective of the project is to improve procedures and outcomes by modelling the entire supply chain for that hospital, from donor to recipient. The supply chain of blood products is broken down into material flows and information flows. Discrete-event simulation is used to determine ordering policies leading to reductions in shortages and wastage, increased service levels, improved safety procedures and reduced costs, by employing better system coordination. In this paper we describe the model and present results for a representative medium-sized hospital. The model can be used by both the National Blood Service and by hospital managers as a decision support tool to investigate different procedures and policies.

14.1 Introduction

The demand for blood in the UK National Health Service is immense. Blood transfusions are frequently required for trauma victims, many types of surgery, organ transplants, in childbirth, and for patients receiving treatment for diseases such as cancer, leukaemia, and anaemia. With an aging population and advances in medical treatments requiring blood transfusions, the demand for blood continues to increase. Forecasts show an increase of 5% in the UK by 2008 (Wells *et al*, 2002),

Reprinted from *Journal of the Operational Research Society*, 58: 219–227, 2007, 'Using Simulation to Improve the Blood Supply Chain', by K. Katsaliaki and S. C. Brailsford. With kind permission from Operational Research Society Ltd. All rights reserved.

increasing within 20 years by 20% relative to supply (Currie *et al*, 2004). On the other hand, the introduction of precautionary measures to counteract the threat of new infections such as Creutzfeldt–Jakob disease (CJD) seriously restricts the pool of blood donors by ruling out existing or potential donors. These measures, together with the introduction of more sophisticated technologies for collecting blood, are constantly increasing costs within the blood supply system.

Currently only 6% of the UK population give blood, although 60% are eligible to donate (NHS Direct, 2004). The figures are even worse in the USA (5%) (Jones, 2003) and Canada (3.7%) (Canadian Blood Services, 2005; http://www.bloodservices.ca/, accessed 12 September 2005). Nationally, in the UK the demand for blood is closely balanced by the overall supply and occasionally exceeds it. However, poor coordination can result in a shortage in one area while units are wasted in another area. Since blood is so vital, a shortage can put life at risk, for example if an operation has to be postponed. Moreover, from a purely financial perspective, blood is a very costly product and its wastage is highly undesirable.

The overall aim of this study is a detailed analysis of a vertical section of the blood supply chain, to identify the factors involved in decision-making by the National Blood Service (NBS) centre and the hospital to which it supplies. This detailed, holistic perspective has not been taken in previous research in this area. By taking this view, we seek to obtain a better understanding of the system, and therefore to be able to manage it to benefit both parts of the supply chain. This should lead to improved service levels and reduced shortage and wastage to meet increased demand and/or reduced supply in the future.

Flows of material and information are included in the system as essential parts of the supply chain. Information flows are represented by placing orders and updating the status of delivery and record-keeping. Costs represent one of the performance measurements of the system for comparing different strategies. The problem is tackled using discrete-event simulation. We present results for a typical medium-sized hospital supplied by an NBS centre, and we recommend the best policy to be adopted for this scenario.

14.2 Description of the blood supply chain and issues

This study was carried out in collaboration with the National Blood Service for England and North Wales (NBS). The NBS consists of 15 Process, Testing and Issuing (PTI) Centres that serve 316 hospitals

across England and North Wales. Hence, each centre serves around 20 hospitals, geographically nearest to it.

The flow diagram (Figure 14.1) shows the processes of the blood supply chain.

14.2.1 Collections

Most blood is collected by NBS teams in 'Bloodmobiles' in public and industrial venues. A unit of 400 ml is the quantity taken from a donor in a single session. The blood is then transported to the Centres for processing during or after the scheduled sessions.

14.2.2 Processing

Blood is processed into several components such as red blood cells (RBCs), platelets, fresh frozen plasma and 115 other products, derived

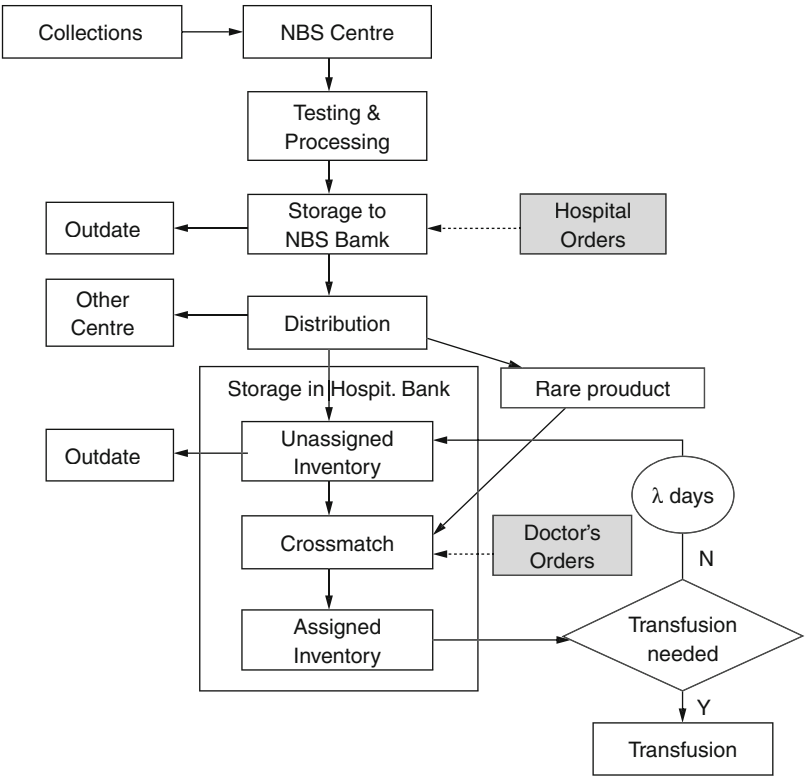


Figure 14.1 Flow chart of blood units and orders in the blood supply chain

from the main blood components after extra tests or processing, which change the shelf life of each product.

14.2.3 Testing and storage

During the processing stage, the donor's blood sample is tested for ABO and Rhesus grouping, as well as for infectious diseases or problematic antibodies. Any donated blood that fails these tests is discarded. After testing the units are ready for storage. Usually the oldest blood is given first (first in, first out FIFO discipline).

14.2.4 Delivery

The NBS collects blood and banks it locally. The majority is also used locally, but the NBS also moves blood between centres and zones to accommodate surpluses and shortfalls. Centres with excessive stock send 'middle-aged' RBC units to centres with deficient stocks once a week, if necessary. There are three kinds of delivery:

- routine scheduled deliveries, that are made on a daily basis and are free of charge to the hospital. Typically, the NBS transport fills the routine orders of several hospitals in one round trip at the same time of the day;
- emergency deliveries, prioritized for immediate dispatch;
- *ad-hoc* deliveries, additional to routine deliveries and made to an individual hospital if it places an order.

There is a single national charge of £15 for emergency and *ad-hoc* deliveries, regardless of distance or frequency. This rather low price arguably allows hospitals to misuse the service, enabling them to avoid developing a better ordering policy. Consequently, the NBS employs more drivers and staff than strictly necessary to deal with these orders. Hospitals are not legally allowed to return unused blood to the NBS Centres.

14.2.5 Hospital inventory management and demand

Hospital blood banks determine the optimal stock levels of RBC according to their own estimates of potential demand. An order is placed with the local NBS Centre when the inventory falls below a predetermined order point, or when rare products not held in stock are requested for particular patients. Hospital doctors are responsible for the quantity of blood products required and consumed for each patient in the hospital.

Following a request from a doctor for blood for a particular patient, *cross-matching* takes place in the hospital blood bank. This is a test to determine compatibility between donor's and recipient's blood. Crossmatched units are termed *assigned inventory*. However, these units are not always transfused because of over-ordering or postponed operations and in practice, only around half of all cross-matched units are actually transfused.

Any untransfused crossmatched blood remains assigned to that patient for a certain time in case complications arise, but is eventually returned to the hospital blood bank as *released inventory*. The elapsed time between assigning blood to a named patient and returning any unused units as released inventory is called the *crossmatch release period*. The duration of this period is determined by the hospital, but normally it is about 2 days. The longer this period, the less chance a unit has of being transfused, because of outdating. Typically, a unit is crossmatched four times before it is finally transfused or expires.

Another interesting issue in the transfusion process is the substitution or *mismatching*. Ideally, a patient should be transfused with the same blood group as his/her own, but this is not always possible. When a patient's blood group at the time of request is unavailable, a compatible blood group must be provided. For example, patients with group A positive can receive blood from group A negative, O negative or O positive. Table 14.1 shows all the possible combinations and also the frequency distribution of blood groups in the UK population. For clinical reasons, some compatible blood types are preferable to others. Nevertheless, in practice mismatching has also been observed in cases based on expiry time

Table 14.1 Mismatching donors against patients' group

		Patient							
Donor	**UK (%)**	**O Pos**	**O Neg**	**A Pos**	**A Neg**	**B Pos**	**B Neg**	**AB Pos**	**AB Neg**
O Pos	37		✓	✓	✓	✓	✓	✓	✓
O Neg	7	✓		✓	✓	✓	✓	✓	✓
A Pos	35				✓			✓	✓
A Neg	7			✓				✓	✓
B Pos	8						✓	✓	✓
B Neg	2					✓		✓	✓
AB Pos	3								✓
AB Neg	1							✓	

Source: BSMS (2003).
Pos, positive; Neg, negative.

Table 14.2 National Blood Service, Price List 2004/05

Components	Price per pack (in pounds)	Added value services
Whole blood	120.48	
Red cells RBC	120.22	£25
Platelets PLT	198.76	£122
Plasma FFP	30.90	
Cryoprecipitate	36.62	

Source: www.blood.co.uk/hospitals/communications/hl/0402/Prices9495.pdf.

in order to meet government targets on expiries. In general, mismatching is undesirable because it increases the risk to the patient.

14.2.6 Costs

There is a single national price for each blood product, with extra charges for additional testing (Table 14.2). Hospitals pay monthly invoices to the NBS for the purchase of blood components. Since 1999, when the national charges were established, prices have been increasing by more than 10% per annum, due to the introduction of new safety tests, more advanced processing methods, and so on. This money generates funding to enable the NBS to extend its product range and enhance its research and development. Further information on price lists over recent years can be obtained from http://www.blood.co.uk/hospitals/index.htm.

14.3 Previous modelling work in this area

Isolated parts of the supply chain were studied back in the late 1970s and early 1980s, mainly in the US. These studies focussed either on perishable goods in general, which only match some of the characteristics of blood processes (Nahmias and Pierskalla, 1973; Fries, 1975; Nahmias, 1976; Chazan and Gal, 1977; Goh *et al*, 1993) or on hospital blood bank inventory problems (Jennings, 1973; Pierskalla *et al*, 1977; Prastacos, 1984). These models were rather inflexible with numerous assumptions, but undoubtedly made a great contribution to understanding the theory of perishable inventory. Three more recent studies use simulation to model platelet inventory management (Sirelson and Brodheim, 1991; Blake, 2003; Kopach *et al*, 2003). Some qualitative research (Spens and Bask, 2002; Brandeau *et al*, 2004) described the supply chain of blood in other countries, but this did not provide a quantitative tool capable of modelling the product flow in the entire chain. Moreover, mismatching is not

captured in any of these models with an exception of Kopach's model (2003) which focuses on shortages. This is one of the key parameters that affects stock levels (BSMS, 2001; http://www.blood.co.uk/bsms/ comms/ reports/ips2001.pdf accessed 12 September 2005). Other factors such as daily demand, transfusion to crossmatch ratio and the crossmatch release period have been modelled successfully (Cohen and Pierskalla, 1979a; Jagannathan and Sen, 1991). Finally, no model has incorporated more than one blood product or varieties of products, due to the complexity of modelling the different shelf lives and of calculating the remaining shelf life when a unit is returned to the blood bank (Nahmias, 1982).

Thus, our model represents four advances from previous models:

1. we model an entire vertical section of the supply chain, from donor to patient;
2. we include mismatching;
3. we incorporate products with different shelf lives;
4. we include the time a unit spends in the assigned inventory stage when calculating the remaining shelf life.

14.4 The simulation model

A discrete-event simulation model was constructed, representing the operations shown in the flow diagram (Figure 14.1) in detail. Simulation was chosen as the most appropriate method to model this system, since both supply and demand are stochastic, the system is complex, and the processes and outcomes (such as outdates) are time driven as discrete events. This complex stochastic multi-product, multi-echelon perishable inventory problem is intractable by analytic techniques (Cohen and Pierskalla, 1979b).

The model was developed using the Simul8 software package (www.simul8.com), together with the support of Visual Basic for Applications. Modelling this large system required extensive data. Nineteen months' data from the Southampton NBS centre was provided and analysed using the NBS information system (PULSE). This gave details of blood collections and the products supplied to each hospital, by date, time, delivery type, quantity and blood group. Questionnaires were sent to the hospitals supplied by the Southampton NBS centre, and interviews conducted with NBS staff and hospital blood bank managers, for clarification and validation. Distributions were fitted to the PULSE data to describe the arrival of donors (collections) and doctors' requests. However, blood bank managers and doctors do not use well-defined procedures for placing

orders and mismatching blood groups, mainly because decision-making tends to be subjective and reliant on individual expertise. Therefore, assumptions about some of the current ordering procedures could not be avoided; nevertheless, the people working in the organizations considered these assumptions to be sensible and realistic. Moreover, procedures in the different NBS centres vary slightly, mainly in their working times. The current model describes the procedures at the Southampton NBS.

The model represents the processes of a hypothetical but representative medium-size hospital located 40 min away from the Southampton Centre. The Centre's inputs are reduced to cover the demand of this hospital, so that overall supply and demand are balanced. The database from PULSE provides the actual number of issues from the Centre to each hospital, by product, and thus only those issues to the chosen hospital were used as input. This equilibrium is maintained by moving the surplus stock once per week to other centres, so stock does not accumulate. Only middle-aged stock is transferred, which does not greatly affect the average age range of blood sent to the hospital. However, units of AB blood group are hardly ever transferred. These are normal processes of all NBS centres as observed in practice.

The baseline model incorporates common blood-ordering policies from several hospitals. The model does not represent an identifiable hospital, but these policies are all in use in one or more of the hospitals where data were collected. A snapshot of the model's interface follows (Figure 14.2).

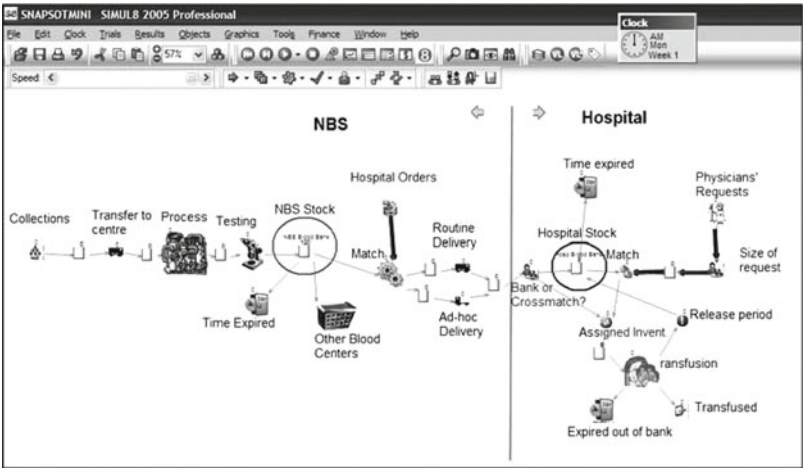


Figure 14.2 Snapshot of simulation model

14.4.1 Entities

There are two categories of entities in the model, one representing individual units of blood and the other orders for units of blood. The latter encompasses doctors' orders, which are satisfied from the units kept in stock in the hospital bank, and also hospital blood bank managers' orders, which are satisfied by the Blood Centre's stock. When an order is fulfilled the order and the actual unit combine to become a single item, as in a Kanban system.

We include the most common products, adult RBCs with a shelf-life of 35 days, and irradiated RBCs, which are derived from adult RBCs and have a life of 14 days. Together these capture 81% of the total units issued in hospitals. Frozen plasma represents 10% of issues, yet because of its long shelf life it does not cause problems and is always ordered together with other products. Therefore, it is excluded from our model. Platelets, which have a shelf-life of 5 days, represent 4% of issues and are included merely to define the way orders are placed with the NBS. Some RBCs are obtained through *ad-hoc* deliveries simply because platelets are needed and since a delivery is going to be made, small shortages of RBCs can be covered. This practice was identified because of high numbers of unexplained *ad-hoc* RBC orders. Wastage of platelets is not counted by the NBS or the hospitals, since the NBS does not plan production of platelets and usually hospitals do not keep stocks of platelets but just order them on request. Thus, the model captures 95% of blood product issues. The other 5% comprises 110 products with different shelf lives which for reasons of simplicity were not included in the model. Each blood product actually consists of eight sub-products representing the eight different blood groups, each of which is handled individually in the model.

14.4.2 NBS processes

Whole blood units arrive in the system from three entry points, representing the three collection modes, each described by a Poisson distribution. The frequency of blood groups is determined by a probability distribution following the UK population frequency of ABO and Rhesus blood types. The entry points represent the public, industrial and on-site collection sessions which operate at different times of the day. Units from the first two collection modes accumulate until they are transferred to the Centre for processing, twice per day. Whole blood for which the group has already been identified at collection is processed as RBC. A pack of whole blood gives a pack of RBC so there is no actual transformation in the model. The tests are released next day at 1400 hours and each unit is stored in the Centre's bank until an order

arrives matching its blood group. Then the order and the acquired unit merge to become a single entity that waits to be transported to the hospital bank (or straight to assigned inventory, if the order is for a rare or irradiated RBC). The time that the unit will be transported and the travel time depend on the order's status (routine, *ad-hoc*, emergency). A routine delivery is dispatched at 1000 hours every morning except Sunday. An *ad-hoc* or emergency delivery will be dispatched as soon as the units are prepared. If the order is for irradiated RBC, the age of the unit is changed to 14 remaining days of shelf life.

14.4.3 Hospital processes

There are five different types of blood requests from physicians. These include adult RBC units under normal circumstances, adult RBC units in emergencies, RBC units with extra group characteristics which are not held in stock and have to be specially ordered from the NBS, irradiated RBC and platelets. Demand for blood products is greater during working hours and is less during the evening and night and on Sundays. All request arrivals follow time-dependent Poisson distributions and represent a doctor's orders for an individual patient. Hence, each request is for a single blood group. The distribution of requested blood groups also follows the UK population frequency of blood types. The size of the request is described by a LogNormal distribution with a mean of 1.3 to 4.3 units, depending on the product, or with a mean of more than 10 units in cases of emergency requests which involve heavy bleeding. Other researchers in the past used the LogNormal distribution as well to describe this variable (Elston and Picrel, 1963; Rabinowitz and Valinsky, 1970; Brodheim and Prastacos, 1980). Each request is matched with the stock in the hospital bank and then proceeds to crossmatching as a single entity.

The hospital keeps almost 5 days of stock, but this is not spread evenly across all blood groups. For instance, there is no stock of B-negative since this is a very rare group. This is common practice in many hospitals. Approximately every hour, the hospital blood bank is checked for stock availability. Every Monday to Saturday at 0800 hours routine orders are placed whenever the RBC inventory is less than 80% of the ideal stock, and the difference for each blood group is ordered. Also, platelets or rare RBC units might be ordered if requested at that time by the doctors. This order will be delivered next day around 1100 hours.

The units are kept in FIFO order in the Centre and hospital blood banks. Ideally, the unused, returned cross-matched units should follow

a last in first out (LIFO) order if they are to be sorted according to the expiry date, but this is not always the case in practice. In the baseline model only a FIFO order is used. The model calculates the shelf life of each unit in both blood banks every simulated minute. When a unit is expired, it is removed from the model. This occurs on the 23rd day of its life in the Centre's bank, preventing very old blood from being given to hospitals where there are limited chances of using it, and on the 35th day in the hospital blood bank (or the 14th if it is for irradiated RBCs, after being irradiated). The Centres keep enough stock to satisfy expected demand for 4.5 days.

The model allows for mismatching if same-group blood will not be available within 2.5h. The selection of the compatible group is based on the following criteria: the oldest compatible blood in the blood bank will be given as long as there are more than four units of this type in the bank.

Finally, blood is crossmatched on average a day in advance of surgery, the time period being described by a LogNormal distribution. It goes through the transfusion process and if unused, it stays assigned to the patient for around half a day more (this duration is again described by a LogNormal distribution), and is returned to the bank at 10.00 the following morning. In all, 50% of the RBCs are transfused on average, which means that the batches ordered by doctors are usually larger than necessary.

While the model runs data are reported in an Excel file, such as the day and time of placing an order with the Centre, the type of order (routine, *ad-hoc* or emergency), the requested product and the amount by blood group.

Shortages are reported in the same way as orders. A shortage occurs when the required blood group is not available from the Centre at the time of ordering by the hospital. Shortages are rarely reported in the real system. However, this is not because there are always available units in the Centre blood bank, but mostly because hospitals change the orders placed with the NBS to match the Centre's available stock and possibly mismatch the units. Also records are kept for the age of blood by group when entering the hospital blood bank and all mismatches are reported by patient's and recipient's group.

These features helped the verification process by identifying errors in the modelling, and validation was performed by comparing the outputs of the model with the real data given from PULSE from similar-sized hospitals. At the end of the run more output measures are available from Simul8.

14.5 Simulation experiments and results

Since it is a discrete-event simulation the model runs in continuous time, but the model status is checked every minute. The model starts from a representative state to eliminate the need for warm-up. Figure 14.3

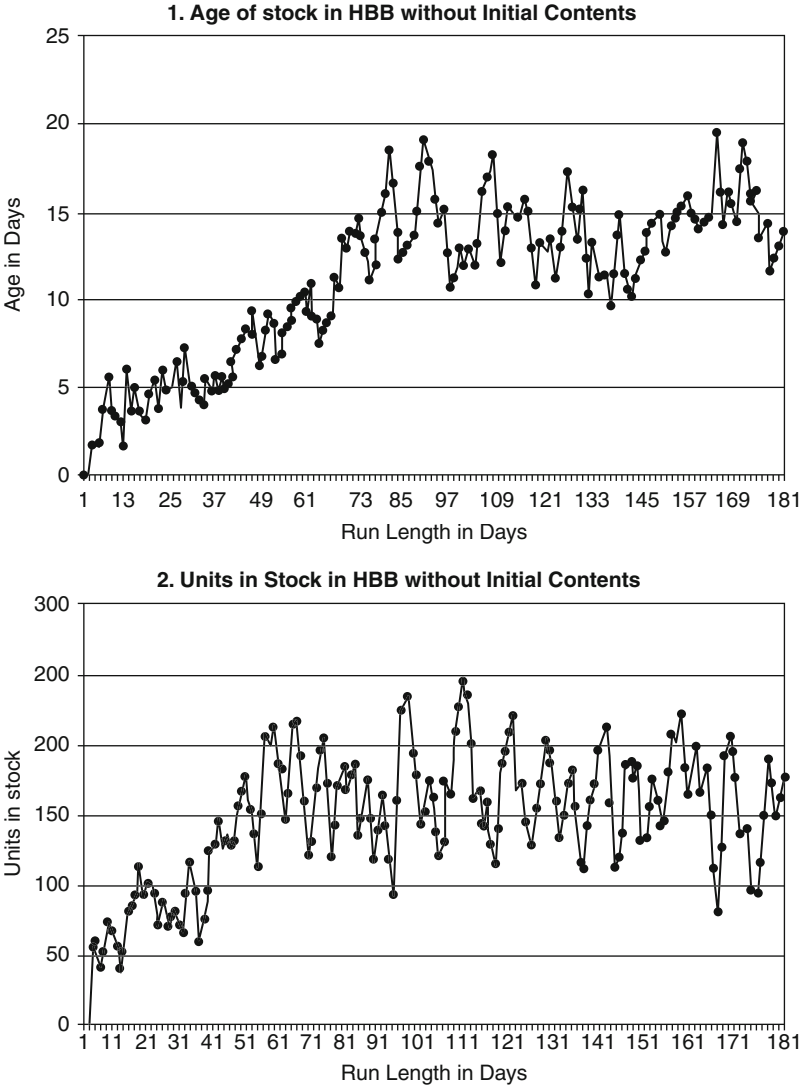


Figure 14.3 Phases of a simulation run with no initial contents. (a) Average age of all units in stock in the Hospital Blood Bank (HBB). (b) Number of units in stock in the HBB

shows the transient period of the model starting with zero stock; it can be seen that steady-state is reached after approximately 90 days. The initial status of the model is represented by the number of units in stock and the age of that stock. The initial stock of the Centre was calculated from the actual stock data of the Centre, adjusted to the size of the model. The hospital's initial stock is the ideal 5 days of stock (50 units of A+, 10 of A–, 5 of AB+, 5 of AB–, 12 of B+, 0 of B–, 60 of O+, 10 of O–) of the baseline scenario.

Monthly seasonality was not observed in the data set. However, daily fluctuation is evident and is captured in the model by altering the time-dependent input distributions of supply and demand at weekends. The model runs for 6 months, a sufficiently long period to capture all the required variability in the processes, bearing in mind that an RBC unit completes its cycle in the system at maximum 35 days. Averages are taken over five iterations for each policy to capture the possible variability of different runs. The key performance measure, actual average monthly outdates (12.5 units) were within the 95% confidence interval of average model values for this length of run and number of iterations. Table 14.3 shows some alternative run lengths and number of iterations; five iterations of 6 months duration (which took over 45 min to run) were chosen as a pragmatic tradeoff between model run-time and accuracy. Nevertheless, future work will include systematically testing the performance of the model under different run disciplines.

Before the experiments were performed, performance indicators were selected. However, we note that the number of outdates in the Centre depends on the number of outdates in the hospital. If the hospital manages its processes better and has fewer expired units, it can satisfy the same demand with smaller orders than before. In the current model, since the number of units in the system stays the same, this will result in fewer outdates in the hospital and more in the Centre (or more units moved to other centres). Nevertheless, the whole system will operate with less donations overall. On the other hand, if supply is more than demand the Centres will have piles of ageing stocks and the hospitals will receive old blood which has less chance of being used. So, both the

Table 14.3 Validated model results for wastage figures according to the length of run and number of runs

Run length	Runs	–95%	Average	95%
6 months	5	10.18	13.93	17.69
4 months	10	11.92	13.18	14.43
2 months	25	10.82	12.13	13.43

Centres and hospitals' outdates will increase, regardless of the good or bad practices of the hospital.

Therefore, the key performance targets for determining the cost-effectiveness and quality of the system had to be chosen carefully. We selected the following:

- number of outdates by group in the hospital only (the smaller the better—saves money);
- number of shortages (the smaller the better—saves lives);
- number of mismatches (the smaller the better—saves lives and money);
- number and percentage of routine/*ad-hoc*/emergency deliveries to the hospital (the smaller the better for the total number of deliveries, but routine deliveries are more cost-effective and well-organized from the Centre so the problem lies largely with *ad-hoc* and emergency deliveries).

The results of the baseline scenario were then compared with alternative policies proposed by hospital and NBS staff. In every run a different policy was tried one at a time. The model parameters and configuration were amended to reflect the new policy and the results compared with the baseline model using the above four criteria. The results did not necessarily show the simultaneous improvement of all the criteria, which complicated the recognition of a bad or a good policy. Certainly, the criteria do not have the same significance; for example, the risk to human life because of mismatch is far more important than cost savings. A weighted average could be a good solution to this problem. However, the combination of several policies had a different outcome overall. Hence, we attempted by trial and error to make realistic combinations of policies, following the suggestions of the NBS staff and blood bank managers. Only the best of the tested combinations is presented in this paper.

The actual results of the baseline model and the percentage change of the tested policies compared with the baseline scenario are given in Table 14.4. The numbers in bold indicate that the percentage difference was significant at the 5% confidence level.

- Policy 1, holding B-negative in stock too, radically decreases the cases of mismatches but slightly increases the expired units, as this group is infrequently requested as it is a rare blood group.
- Policies 2 and 3, keeping 4 and 3 days of stock, respectively, instead of 5 days, significantly reduces the number of expiries by 29 and

Table 14.4 Percentage change in the performance measurements achieved with different policies. The model runs for 26 weeks for a hospital with about 630 orders of RBC units per month

Policies	Expired units	Mis-matched units	Shortages in units	No. of all deliveries	No. of *ad-hoc* and emergence deliveries
(0) Baseline, actual results	333	159	147	292	199
(1) Keep stock of Bneg = 5	**+10%**	**−75%**	+10%	**+3%**	**+3%**
(2) 4 days stock (A+40, A−10, AB+4, AB−4, B+10, B−0, O+40, O−10)	**−29%**	−10%	+3%	**+8%**	**+8%**
(3) 3 days stock (A+32, A−6, AB+3, AB−3, B+7, B−3, O+34, O−7)	**−47%**	+7%	**−29%**	**+21%**	**+31%**
(4) Crossmatch release period 24 h	**−29%**	−9%	**−14%**	−2%	**−4%**
(5) 70% transfused, smaller batches of orders from physicians	**−51%**	**−38%**	**−46%**	**−8%**	**−14%**
(6) *Ad-hoc* order = 35% of ideal stock	+2%	−6%	+12%	−9%	**−15%**
(7) Blood Bank FIFO+LIFO for returns	**−70%**	−9%	−5%	−2%	−2%
(8) Order at 16.00	−2%	−3%	**+15%**	0%	−2%
(9) 2 routines (order at 1600 hours delivery at 1100 hours next day/ order at 900 hours delivery at 1500 hours same day	−2%	−7%	6%	**8%**	**−13%**

The numbers in bold indicate that the percentage difference was significant at the 95% confidence level.

47%, respectively, yet since less stock is kept more frequent ordering is required.
- Policies 4 and 5 look good for all the performance measurements. Here units stay in assigned inventory for a shorter time and have a reduced crossmatch to release period, and thus return faster to the blood bank if unused. Outdates are significantly diminished, and shortages and mismatched units are further reduced when the transfusion to crossmatch ratio rises to 70% as a result of more accurate requests from doctors.
- Policy 6, which uses a lower *ad-hoc* ordering point, obviously decreases the number of *ad-hoc* orders placed with the Centre (by 15%) but causes more shortages.
- Policy 7, following LIFO order for returned units in the hospital bank and FIFO for those entering for the first time, leads to a huge drop in the number of outdates.
- Placing orders at 1600 hours delivered at around 1100 hours next day (Policy 8), which implies a smaller lead time, hardly affects any of the criteria apart from shortages, which are increased.
- Finally Policy 9, introducing two routine deliveries (one in the afternoon to satisfy the orders placed in the morning, assuming there is enough time for the Centre staff to prepare the orders, and one in the morning to satisfy the routine order placed at 1600 hours the day before) decreases the *ad-hoc* and emergency deliveries by 13%, although it apparently increases the overall number of orders by 8% because routine deliveries have increased.

Table 14.5 illustrates the policies applied in the baseline and the improved policies, and Table 14.6 shows the results of the improved model in actual values and percentage improvement in the performance measurements of comparing the current practices of the example hospital with the best combination of policies recommended for adoption.

14.6 Discussion

In this study a computer simulation model was constructed to test in a risk-free environment the effects of several policies while managing the supply chain of blood products. The model captures all the activities from donation to transfusion and shows how some alteration in the current practices of the blood centres, the physicians and the medium-size hospitals' blood banks could improve service levels and the budget

Table 14.5 The baseline scenario compared with the improved policy

Baseline scenario	**Improved policy**
Hospital keeps 5 days of stock and no B-negative stock	4 days of stock of all groups
Routine orders at 08.00 only	Routine orders at 08.00 and 16.00
1 routine delivery, arrive 11.00 next day	Two routine deliveries at 11.00 and 15.00
50% of units transfused	70% of units transfused
48 h in assigned inventory	24 h in assigned inventory
Ad-hoc orders when either platelets or rare RBC requested, or RBC <50% of optimal level	The same, but RBC <35% of optimal level
All units FIFO order in bank	FIFO new units + LIFO returns

Table 14.6 Percentage improvements in the performance indices by applying the improved policy, compared with the baseline

			Savings	
Performance measurement	**Actual number**	**% improvement**	**£**	**%**
Expired units	36	89	£38 500	8.1
Shortages in units	78	47	Lives	
Mismatched units	19	88	Lives	
No. of all deliveries	293	0	£855	29
No. of *ad-hoc* and emergency deliveries	142	29		

of all the stakeholders, and how the system could easily cope with future threats of increased demand or shrunk supply.

The system can be improved by adopting the following policies:

- decrease the holding stock to 4 days,
- the introduction of two routine deliveries in working hours to better manage orders from the NBS and grouped transport to hospitals,
- a more insensitive *ad-hoc* ordering point for RBCs to 35% of the optimal stock level,
- the reduction of the total crossmatch release period (before and after transfusion) to 1 day,
- the increase of the transfusion to crossmatch ratio to 70%, either through stricter compliance with the recommended ordering system

leading to more accurate orders placed by doctors, or by applying multiple-crossmatching techniques
- strict adherence of the hospital blood bank when cross-matching units to a FIFO order for units coming from the Centre and LIFO for the unused, returned units from the other wards. In other words, compliance with a FIFO order according to the collection/production date of the unit.

In adopting these recommendations, the total system will gain:

- 89% (297 units) less RBC outdates and 8% (£38500) cost savings in the hospital budget;
- 47% (69 units) less shortages from the Centre;
- 88% (140 units) less mismatching resulting in avoidance of possible complications;
- 29% (57) less *ad-hoc* and emergency orders from the hospital with 29% savings (£855) on the money given for charged deliveries from this hospital alone, and uncounted savings for the NBS by reducing the number of occupied drivers, staff and transport.

The area is fruitful for future research. Taking the model forward, a small and a large hospital could be examined to identify better practices for different sized hospitals. Next, a more complex model with three hospitals of different sizes supplied by the same NBS centre could be built to illustrate how hospitals compete for supplies and whether the good policies identified in the individual models can also be effective in a combined model. The model can be further expanded to include all the hospitals served typically from a NBS Centre. Thus, general policies can be recommended and standardized processes could be applied. However, this raises issues of computation time because of the very large number of entities generated.

Furthermore, better information about the production of platelets by the NBS and their use by hospitals would lead to better understanding of the platelets supply chain. Details of this process could be incorporated in the model in order to take into account wastage of platelets and better management of this increasingly needed blood product.

Finally, the same methodology and modelling approach can apply to various supply chains of products with the similar characteristics, such as the food industry, horticulture and floristry, pharmaceuticals, transplants, and other perishable goods.

Acknowledgements

We thank Crispin Wickenden and Mike Northcott from the NBS for their time, advice and support, without which the project would not have been possible. We also thank several employees of the Southampton NBS Centre for their time and effort during the data collection, particularly Andrew Oliver for the provision of the complete database and Tracey Lofting from Southampton General Hospital for her enlightening comments. We express our gratitude to the members of the Blood Stocks Management Scheme for helping with the validation and potential use of the model.

References

Blake J (2003). Optimizing the platelet supply chain in Nova Scotia. In: Dlouhý M (ed). *Proceedings of the 29th Meeting of the European Working Group on Operational Research Applied to Health Services*. ORAHS, Prague, Czech Republic, pp 47–66.

Brandeau ML, Sainfort F and Pierskalla WP (2004). Operations research and health care: a handbook of methods and applications. *Series: International Series in Operations Research & Management Science*, Vol. 70, chapter 5, pp 103–146.

Brodheim E and Prastacos G (1980). *Demand usage and issuing of a blood at Hospital Blood Banks*. Technical Report, Operations Research Laboratory, The New York Blood Centre.

BSMS Blood Stocks Management Scheme (2001). Ad hoc exercises summer 2001. Preliminary Report, BSMS, PO Box 33910, London, NW9 5YH.

BSMS Blood Stocks Management Scheme (2003). *Inventory Practice Survey 2003*. BSMS, PO Box 33910, London, NW9 5YH.

Chazan H and Gal S (1977). A Markovian model for a perishable product inventory. *Mngt Sci* **23**: 512–521.

Cohen MA and Pierskalla WP (1979a). Target inventory levels for a hospital blood bank or a decentralised regional blood banking system. *Transfusion* **19**(4): 444–454.

Cohen MA and Pierskalla WP (1979b). Simulation of blood bank systems. *Simuletter* **10**(4): 1–8.

Currie CJ, Patel TC, McEwan P and Dixon S (2004). Evaluation of the future supply and demand for blood products in the United Kingdom National Health Service. *Transfusion Medicine* **14**: 19–24.

Elston R and Picrel JC (1963). A statistical approach to ordering and usage policies for a Hospital Blood Bank. *Transfusion* **3**: 41–47.

Fries B (1975). Optimal ordering policy for a perishable commodity with fixed lifetime. *Opns Res* **23**: 46–61.

Goh C, Greenberg BS and Matsuo H (1993). Perishable inventory systems with batch demand and arrivals. *Opns Res Lett* **13**(1): 1–8.

Jagannathan R and Sen T (1991). Storing crossmatched blood: a perishable inventory model with prior allocation. *Mngt Sci* **37**(3): 251–266.

Jennings JB (1973). Blood inventory control. *Mngt Sci (Application Series)* **19**(6): 637–645.

Jones RL (2003). The blood supply chain, from donor to patient: a call for greater understanding leading to more effective strategies for managing the blood supply. *Transfusion* **43**(2): 132–141.

Kopach R, Frances DM and Sadat S (2003). Models for predicting a critical blood product. In: Dlouhý M (ed). *Proceedings of the 29th Meetingof the European Working Group on Operational Research Applied to Health Services*. ORAHS, Prague, Czech Republic, pp 77–90.

Nahmias S (1976). Myopic approximation for the perishable inventory problem. *Mngt Sci* **22**: 1002–1008.

Nahmias S (1982). Perishable inventory theory—a review. *Opl Res* **30**: 680–780.

Nahmias S and Pierskalla WA (1973). Optimal ordering policies for a product that perishes in two periods subject to stochastic demand. *Naval Res Logistics Q* **20**: 207–229.

Pierskalla W, Deuermeyer B and Sassetti R (1977). Methods for reducing outdating without altering physician ordering patterns. Technical Report. Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Prastacos GP (1984). Blood inventory management; an overview of theory and practice. *Mngt Sci* **30**(7): 777–800.

Rabinowitz M and Valinsky D (1970). *Hospital Blood Banking—an evaluation of inventory control policies*. Mt. Sinai School of Medicine, City University of New York.

Sirelson V and Brodheim E (1991). A computer planning model for blood platelet production and distribution. *Comput Meth Programs Biomed* **35**: 279–291.

Spens KM and Bask AH (2002). Developing a framework for supply chain management. *Int J Logistics Mngt* **13**(1): 73–88.

Wells AW *et al* (2002). Where does blood go? Prospective observational study of red cell transfusion in north England. *BMJ* **325**: 803.

Index

AIDS, *see* HIV/AIDS epidemic, modelling treatment effects in
ambulance dispatch, 40–1
ambulance logistics, 36–50
  computational results of, 47–50
  dispatching support, 40–1
  dynamic ambulance relocation, 41–5
  operations simulator, 45–7
  preparedness, 39–40
ambulance relocation, dynamic, *see* dynamic ambulance relocation problem (DYNAROC)
Analytic Hierarchy Process (AHP), 88
assigned inventory, 357
Automated Situation-awareness Assessment Platform (ASAP), 53, 56

Balanced Scorecard (BSC), 87–8
  discrete-event simulation-based, Multi-Criteria Decision Analysis using, 85–116
    background of, 86–7
    Balanced Scorecard and simulation, integrating, 91–2
    business process modeling, 89–91
    data analysis, 100–2
    emergency department Balanced Scorecard, 97–100
    key performance indicators, selection of, 104–8
    literature review, 87–9
    model development and validation, 102–4
    process mapping, 93–7, 98
    real-time strategies, 108–15
    simulation modelling, 91
Bayesian decision model
  with hurricane forecast updates, for emergency supplies inventory management, 330–51
    assumptions of, 335–8
    background of, 330–1
    Bayes risk function, 339–41
    example problems, 347–8
    hurricane prediction and simulation, 345–7
    literature review, 333–5
    model validation, 348–50
    numerical examples, 345
    probabilistic hurricane prediction model, 342–5
    problem description, 331–3
    single period loss function, 338–9
    solution methodology, 341
Bayes risk function, 339–41
blood bank inventory problem, 251
blood supply chain, using simulation to improve, 353–70
  collections, 355
  costs, 358
  delivery, 356
  hospital inventory management and demand, 356–8
  previous modeling work, 358–9
  processing, 355–6
  simulation experiments and results, 364–8
  simulation model, 359–64
    entities, 361
    hospital processes, 362–4
    NBS processes, 361–2
  testing and storage, 356
business process modeling, 89–91

CICAS-Stop Sign Assist (CICAS-SSA), 55
climatological wind radii forecast model, 343–4
Climatology and Persistence (CLIPER) model, 344–6
continuous production model, 294–5
Cooperative Intersection Collision Avoidance System (CICAS) program, 55

cost decomposition, 301–6
crossmatching, 357
crossmatch release period, 357

data-driven patient grouping, 123–4
Dijkstra-based shortest time path, 59–63
  distinct weather regions, 61
  event-based congestion, 62–3
  'insensitive' versus 'sensitive' EMS vehicle, 60–1
discrete-event simulation (DES)
  -based Balanced Scorecard, Multi-Criteria Decision Analysis using, 85–116
    background of, 86–7
    Balanced Scorecard and simulation, integrating, 91–2
    business process modeling, 89–91
    data analysis, 100–2
    emergency department Balanced Scorecard, 97–100
    key performance indicators, selection of, 104–8
    literature review, 87–9
    model development and validation, 102–4
    process mapping, 93–7, 98
    real-time strategies, 108–15
    simulation modelling, 91
  and data-mining, combining, 119–35
  of patient treatment, 128–34
    concept realization, 129–31
    treatment-focused discrete event simulation, 128–9
    validation and verification, 131–2
  uncovering effective process improvement strategies using, 139–60
    ED modeling review, 142
    modeling scope and features, 146–51
    new doctor manpower schedules, 152–6
    new laboratory test processing times, 156–9
    process flow, 142–5
    what-if analyses, 152
disposal policy, 292
dynamic ambulance relocation problem (DYNAROC), 41–5
Dynamic Programming (DP), 251, 272–3, 275, 276, 277, 293

Economic Manufacturing Quantity (EMQ) model, 290, 291
  modified EMQ model, 297–301
    exact solution algorithm and complexity analysis, 310
    global optimality, 308–10
    local optimality, 307–8
  with perishable inventory management system, 295–7, 322
emergency department (ED)
  Balanced Scorecard, 97–100
  operations
    models of, 121
    nature of, 120–1
emergency responders, situationally aware routing for, *see* situationally aware routing, for emergency responders
emergency service vehicle location, reliability model for, 18
emergency supplies inventory management
  Bayesian decision model with hurricane forecast updates for, 330–51
    assumptions of, 335–8
    background of, 330–1
    Bayes risk function, 339–41
    example problems, 347–8
    hurricane prediction and simulation, 345–7
    literature review, 333–5
    model validation, 348–50
    numerical examples, 345
    probabilistic hurricane prediction model, 342–5
    problem description, 331–3
    single period loss function, 338–9
    solution methodology, 341
event-based congestion, 62–3, 76
expected gross attack rate (EGAR), 256

first-in-first-out (FIFO), 292, 294, 296, 299, 301, 322, 356, 362–3, 368, 370
fractional factorial designs, 168
fuzzy theory, 293

General Practitioners Post (GPP), 165–7, 169, 170, 174, 175, 178–86
gradient-based procedures, 168
greedy randomized adaptive search procedure (GRASP), 22–7
  pseudocode, 23
    construction phase, 23–5
    local search phase, 25, 26
    objective evaluation, 26–7, 28
    reactive, 25, 27
gross attack rate (GAR), 256, 260–2, 273, 275, 276

heuristics/approximations, 292–3
heuristic solution procedure, for queuing maximum coverage location allocation model, 22–7
Hierarchical Agglomeration algorithm, 125
highly active antiretroviral drugs (HAART), for HIV/AIDS, 225, 234, 242
HIV/AIDS epidemic, modelling treatment effects in, 224–44
  mathematical model, 226–9
  results of, 235–44
    extrapolations, 240–4
hospitalization model, 254–71, 281–7
  patient condition model, 254–61, 279–80
  pharmacy model, 261–71
humanitarian relief inventory management problem, 332–3
Hurricane Katrina, 332
hypercube model, 37

IDEF0, 97
influenza outbreak on hospital, impact of, 249–77
  hospitalization model, 254–71, 281–7
    patient condition model, 254–61, 279–80
    pharmacy model, 261–71
  literature review, 250–2
  numerical results, 271–6
    cost and demand parameters, 271
    sensitivity analysis, 275–6
  SIR infectious disease model, 250, 252–4, 280–1
integrated emergency post (IEP), 164–86
  approach, 173–5
  literature of, 167–9
  modelling for sustainability, 172–3
  results
    categorize interventions, 176–7
    category outcomes, 178–180
    formulate intervention sets, 180–1
    implementation, 184–5
    intervention sets, comparing, 181–4
    list interventions, 175–6
    sustainability, 185
  simulation model of, 170–2
Intelligent Transportation Systems (ITS), 52, 53
  information fusion approach to, 54
inventory management system
  assigned inventory, 357
  blood bank inventory problem, 251
  emergency supplies, *see* emergency supplies inventory management
  with one-way substitution, 251–2
  perishable inventory, *see* perishable inventory management system
  released inventory, 357
  two-echelon inventory system, 251
issuing policy, 291–2

key performance indicators (KPIs), 92, 173, 174, 185
K-PQL model, 17

last-in-first-out (LIFO), 363, 368, 370
locating emergency services
  ambulance logistics, 36–50
  queuing maximum coverage location allocation model, 15–34
  situationally aware routing, for emergency responders, 52–70
logistics, ambulance, *see* ambulance logistics

Manchester Triage System (MTS), 95, 101
marginal productivity of labour (MPL), 191
Markovian model, 293
mass immunization clinics (MICs)
  for pandemic influenza, improving staffing decisions at, 190–220
    baseline scenario, 209–10
    changes in staff productivity by hour, 204–6
    designed simulation experiment, 213–18
    facility closure, 199–200
    line tolerance, 199–200
    literature review, 192–4
    model overview, 194–208
    model validation, 208
    operating costs, 206
    patient arrival rates, 196–7
    patient characteristics, 197–9
    patient screening, 200
    patient wait-time tolerance, 199–200
    registration and vaccination stations, 200–3
    risk of disease transmission, 206–8
    sensitivity analysis, 210–13
maximum likelihood estimation (MLE), 230–5
metaheuristics, 168
M/G/1 nonpreemptive priority queuing model, 17
mismatching, 357
modified Dijkstra's algorithm pseudocode, 72–3
Multi-Criteria Decision Analysis (MCDA), 88, 92
  using discrete-event simulation-based Balanced Scorecard, 85–116
    background of, 86–7
    Balanced Scorecard and simulation, integrating, 91–2
    business process modeling, 89–91
    data analysis, 100–2
    emergency department Balanced Scorecard, 97–100
    key performance indicators, selection of, 104–8
    literature review, 87–9
    model development and validation, 102–4
    process mapping, 93–7, 98
    real-time strategies, 108–15
    simulation modelling, 91

new doctor manpower schedules, 152–6

one-median problem, 17
operational research (OR)
  definition of, 1
  application context, 3
  technique, 2–3
  application area, 3–4
Optimized Ambulance Logistics (OPAL), 37
ordering policy, 291

pandemic influenza
  mass immunization clinics for, improving staffing decisions at, 190–220
    baseline scenario, 209–10
    changes in staff productivity by hour, 204–6
    designed simulation experiment, 213–18
    facility closure, 199–200
    line tolerance, 199–200
    literature review, 192–4
    model overview, 194–208
    model validation, 208
    operating costs, 206
    patient arrival rates, 196–7, 210–11
    patient characteristics, 197–9
    patient screening, 200
    patient wait-time tolerance, 199–200, 211
    registration and vaccination stations, 200–3
    risk of disease transmission, 206–8
    sensitivity analysis, 210–13
patient acuity categorization, 145
patient condition model, 254–61, 279–80
  age and risk groups distributions of admitted patients, 255–9

disease simulation model, performance analysis of, 260–1
patient condition dynamics within hospital, 259–60
patient grouping
data-driven, 123–4
systems complexity through, reducing, 121–2
see also treatment-based patient groups
patient interarrival rates, 196–7, 210–11
patient wait-time tolerance, 199–200, 211
perishable inventory management system, 250–1
with minimum volume constraint, 288–320
computational experiments, 311–18, 328–9
continuous production model, 294–5
EMQ model with perishability, 295–7, 322–8
future research, 318–20
literature review, 291–4
modified EMQ model, 297–301
solution approach, 307–10
total cost evaluation and boundary conditions, 301–7, 322
pharmacy model, 261–71, 281–7
decision-making process, 283
inventory policies, 268–71
medication allocation process, 284
model notation and assumptions, 263–6
status update module, 285–7
total cost function, 266–8
preparedness, 39–40
pricing policy, 292
priority queuing covering location problem (PQCLP), 20–2, 33
directed choice environment, 21–2
user choice environment, 22

queuing maximum coverage location allocation model, 15–34
computation of
heuristic evaluation, 27–9
illustrative example, 29–33
formulation of, 18–22
priority queuing covering location problem, 20–2
problem definition, 20
results from queuing theory, 19–20
future research, 33–4
heuristic solution procedure, 22–7
literature of, 16–18

random search approach, 168
ranking & selection, 168
released inventory, 357
response surface methodology, 168
response time, *see* waiting period

sample path optimization, 168
self-organizing map (SOM), 125
sensitivity analysis, 192
EMQ model with perishable inventory management system, 315–18
influenza outbreak on hospital, impact of, 275–6
multi-criteria approach using simulation based Balanced Scorecard, 111, 114–15
pandemic influenza, improving staffing decisions at MICs for, 210–13
Simple Multi-Attribute Rating Technique (SMART), 88
simulation modelling, 91
single period loss function, 338–9
situationally aware routing, for emergency responders, 52–70
Dijkstra-based shortest time path, 59–63
distinct weather regions, 61
event-based congestion, 62–3, 75–6
'insensitive' versus 'sensitive' EMS vehicle, 60–1, 73–5
literature review, 55–6
motivation for, 53–5
simulation overview, 56–9
'real-time' data capture, 58–9
regular and EMS vehicle travel, 57–8

situationally aware routing, for emergency responders – *continued*
variance-constrained shortest path problem, 63–9
experiment design, 64–65
results of, 65–9
solution methodology, 77–82
SOS Alarm AB, 37
stochastic queue centre location model, 17
Strategic National Stockpile (SNS), 288–90, 294, 318
Susceptible/Infected/ Recovered (SIR) model, 250, 252–4, 280–1
sustainability, 172–3, 185

transportation network simulation model, 56–9
trauma resource allocation model, 17–18
treatment-based patient groups
concept realization, 124–5
data-driven patient grouping, 123–4
validation and verification, 125–8
treatment-focused discrete event simulation, 128–9
Triple Bottom Line (TBL), 165, 166
two-echelon inventory system, 251

uncovering effective process improvement strategies, using discrete event simulation, 139–60
ED modeling review, 142
modeling scope and features, 146–51
new doctor manpower schedules, 152–6
new laboratory test processing times, 156–9
radiological processing times, 159
process flow, 142–5
what-if analyses, 152

Value at Disruption Risk (VaDR), 350
variance-constrained shortest path (VCSP) problem, 52–3, 55, 56, 63–9
event-based congestion, 76
experiment design, 64–65
results of, 65–9
solution methodology, 77–82
Viscovery SOMine, 125

waiting period, 36–7
what-if scenarios/analysis, 91, 92, 152

zero first-order derivative point (stationary point) over total cost function, 292
zidovudine (AZT), for HIV/AIDS, 225, 239