Basic Engineering for Medics and Biologists

An ESEM Primer

Edited by

T. Clive Lee

Royal College of Surgeons in Ireland
and Trinity Centre for Bioengineering, Dublin, Ireland

and

Peter F. Niederer

Institute for Biomedical Engineering
Swiss Federal Institute of Technology (ETH)
and University of Zurich, Zurich, Switzerland

IOS Press

Amsterdam • Berlin • Tokyo • Washington, DC

ISBN 978-1-60750-526-6 (print)
ISBN 978-1-60750-527-3 (online)
Library of Congress Control Number: 2010923280

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

PRINTED IN THE NETHERLANDS

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

Preface

Making Engineering Accessible to Medics and Biologists

Research and development in bioengineering and medical technology, conducted during recent decades, have led to spectacular progress in clinical medicine. These achievements have triggered an enormous increase in the number of courses offered in the areas of bioengineering, clinical technology and medical informatics; nowadays, most major universities offer curricula oriented towards these fields. The majority of participants however come from engineering backgrounds and so modules dealing with basic biological and medical sciences have been included. These have been facilitated by the ready availability of textbooks in this area, many of which were specifically written for nursing A & P (Anatomy & Physiology) programmes.

In contrast, relatively few participants from medicine and biology have taken up courses in biomedical engineering, to the detriment of scientific exchanges between engineers and medical doctors. The reasons for this imbalance are many and may vary from country to country, but a consistent finding is the difficulty (or courage) in taking the first step. 'Introductory' texts in bioengineering tend to involve vector algebra and calculus early in Chapter 1. While in most countries entry to medical school is very competitive and requires, among others, high grades in mathematics, little more than arithmetic is required during the course itself, so numeracy undergoes disuse atrophy.

Furthermore, to paraphrase George Bernard Shaw, medics and engineers are separated by a common language. To the medic, stress and strain are both symptoms of anxiety, while to the engineer they are defined by equations involving symbols such as σ and ε and are related to forces and deformations. To the average medic this is, literally, all Greek.

ESEM is a bridge between medicine and engineering. It promotes cultural and scientific exchanges between engineers and medics and training courses in biomedical engineering. What better way to achieve these objectives than to address this problem and help medics and biologists to take this first step.

To this end, we herewith present a series of *First Step* chapters in book form aimed at medics and biologists to help bring them to the level where they can ***begin*** an MSc in biomedical engineering, or at medics who "simply" wish to understand a particular medical technology. Written by engineers who are leaders in their field, with input from medical colleagues, they will cover the basic engineering *principle*s underpinning biomechanics, bioelectronics, medical informatics, biomaterials, tissue engineering, bioimaging and rehabilitation engineering, and will include clinically relevant *examples*.

As Albert Einstein may have said 'Everything should be made as simple as possible, but not simpler'.

Dublin & Zurich, December 2009 *Clive Lee & Peter Niederer*

Cover Illustration

Stress trajectories in a curved Culmann crane (left) compared with a schematic representation of the trabecular pattern in the proximal human femur (right). From: Wolff J. Ueber die innere Architectur der Knochen und ihre Bedeutung für die Frage vom Knochenwachsthum (On the intrinsic architecture of bone and its significance with respect to the question of bone growth). Virchow's Arch. 1870; 50: 389–450.

Contents

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-1

Introduction to Chapter I: Biomechanics

Gijsbertus J. VERKERKE, Pascal VERDONCK and T. Clive LEE (eds.)

Biomechanics is the science that examines forces acting upon and within a biological structure and effects produced by such forces (*B.M. Nigg, Biomechanics of the musculoskeletal system, 1995*). As such, biomechanics is part of the world of physics and concentrates on forces that act on biological structures. These forces either originate from inside the human body, like muscle forces, or forces, created externally, like an external load or impact. The structures they act upon can be classified as solids, like skeletal parts or organs and by fluids, like blood and air.

The study on the effect of these forces can be classified as deformation and displacement.

Biomechanics concentrates on various situations. To apply biomechanics four subspecialisations are created, each focusing on a specific situation and simplifying the real world in a specific way to allow a practical model that can be analysed theoretically.

Statics deals with solid structures that will not move. In this way the forces that act upon a structure can be calculated. All structures are assumed to be undeformable.

It is based on Newton's third law, $\underline{\mathbf{F}}_{12} = -\underline{\mathbf{F}}_{21}$: when a structure exerts a force on another structure (for instance by its weight) that last body will exert an equal force in opposite direction to the first body. It is also based on Newton's first law:
$\underline{\mathbf{F}} = \underline{\mathbf{0}} \Leftrightarrow \underline{\mathbf{v}} = \mathbf{constant}$: If the resulting force $\underline{\mathbf{F}}$ on a structure is zero, then that structure will not change its velocity, so non-moving structures will remain motionless.

Mechanics of materials deals with solid structures that will not move, but are deformable.

Dynamics deals with solid structures that can move, but cannot deform. It is based on the second law of Newton, $\underline{\mathbf{F}} = \mathbf{m}\,\underline{\mathbf{a}}$: when a force $\underline{\mathbf{F}}$ acts on a body with mass $\mathbf{m}$, it will undergo an acceleration $\underline{\mathbf{a}}$.

And *fluid mechanics* deals with fluids that can move and that can deform.

These four different biomechanical subspecialisations and their applications will be discussed in this chapter.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-3

I.1. Statics

Gijsbertus J. VERKERKE[a,b] and T. Clive LEE[c]
[a]University Medical Center Groningen, University of Groningen
Dept of Biomedical Engineering, Groningen, The Netherlands
[b]University of Twente, Dept of Biomechanical Engineering, Enschede,
The Netherlands
[c]Royal College of Surgeons in Ireland, Department of Anatomy, Dublin, Ireland

Abstract. The forces that act on an object determine its dynamic behaviour and defromation. Analysis of all forces and moments is essential. A free-body diagram summarizes all forces and moments that act on an object. To calculate the magnitude of the forces we can use the static equilibrium of forces and moments.

Keywords. Force, moment, couple, static equilibrium, free-body diagram, internal force

Introduction

If we want to know, how an object will move (kinetics) or deform (mechanics of materials) we first have to know, which forces are acting upon those objects. Statics is the field of mechanics that studies forces on objects that don't move, so the forces acting upon them balance each other.

Since all objects on earth are subjected to gravity, forces are everywhere. Apart from gravity, we can distinguish spring forces, magnetic forces, electric forces, hydraulic and pneumatic forces and inertia forces. A force is defined by a magnitude, a direction and a point of action. A muscle force acts on its origin and on its insertion, has a distinct magnitude that can be derived (for isometric contractions) from the EMG and has a clear direction, the line between origin and insertion. The force of a cue acts on the surface of a billiard ball, in the direction of the cue and with a magnitude that is determined by the velocity of the cue. When you are sitting on a thumb tack, you experience the point of action, magnitude, and direction of a force.

A force creates either movement (of the billiard ball) or deformation (the muscle force you apply to grab a book will deform that book). By definition, force (F) is equal to mass (m) times acceleration (a):

$$F = m \bullet a \tag{1}$$

Acceleration is a change in velocity (Δv) over time (Δt):

$$a = \Delta v / \Delta t \tag{2}$$

Velocity (v) is a change in distance (Δs) over time (Δt):

$$v = \Delta s / \Delta t \tag{3}$$

So velocity is expressed in m/s; acceleration, a change in velocity, is expressed in (m/s)/s or m/s^2. So force is expressed in $kg.m/s^2$, often abbreviated as Newton (N).

Assumptions: To use forces, we have to do some modelling: we assume that a force is acting in a single point with infinitely small dimensions. In reality we study only forces that are acting on a surface, called pressure.

1. Resultant Force

When two forces with different lines of action are acting on the same point of an object (for instance two muscles that are active with the same insertion point but different line of action), then we can combine them into a single resultant force. The magnitude and line of action of this force can be found using the rules for adding vectors (Figure 1).

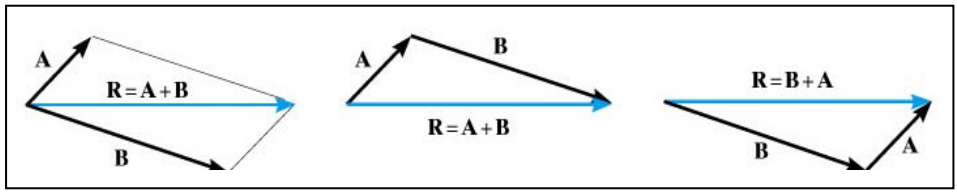


Figure 1. Two vectors **A** and **B** can be replaced by a resultant force **R** using the parallelogram construction (left) or putting them head-to-tail (middle and right).

A force can also be resolved into two different forces using the same rule. Often this is performed to simplify the effect of a force. If the seat of your chair is not horizontal, but inclined under an angle β, then the force F that is acting on this seat due to gravity can be replaced by two forces: a force N that is acting perpendicular to the seat (also called the normal direction) and a force V acting parallel to the seat (also called the shear direction). The force V lets you slide off the seat, while the force N determines whether or not the chair will be able to carry you.

You can find the force V by multiplying force F with sin β (sinus β). Force N is found by multiplying force F with cos β (cosinus β). The value of sin β and cos β can be found on most pocket calculators and using your Windows calculator (scientific view). The angle β is expressed in degrees. So:

$$V = F \bullet \sin \beta$$

$$N = F \bullet \cos \beta \tag{4}$$

1.1. Friction

Four main types of forces exist in and on the human body: gravity forces, muscle forces, friction forces and forces exerted by others or other objects, like a boxing glove. A friction force always acts perpendicular to a surface and prevents an object moving along that surface. So a friction force prevents you from sliding off the inclined seat.

Friction has some strange features:

First of all its magnitude is variable and only as large as is necessary to prevent movement. If a force V is acting upon you, because you are sitting on an inclined seat, then the friction force has the same size as V, but acts in the opposite direction.

Secondly there is a limit to the friction force, of course, because you will slide off a steeply inclined seat. The friction force, acting during motion is slightly smaller than the static friction force.

Thirdly, the friction force is linearly related to the normal force:

$$F_f = \mu_s \bullet F_N \tag{5}$$

with F_f = static friction force (N), μ_s = coefficient of friction and F_N = normal force (N). As you can derive from formula (5), μ_s is dimensionless. Its value depends on the two surfaces that are in contact with each other and can be found in handbooks.

Fourthly, the friction force is independent of the area of contact between the two objects. You can check this by placing a book on a board, start to incline the board and remember at which angle it starts to move. Then place the book on its side on the board and repeat the experiment. The angle will be the same. An explanation for this is that both surfaces are rough and in theory to keep the book stable it rests only at three contact points on the board. The coefficient of friction is determined by the nature of these three contact points and thus is not determined by the contact surface.

1.2. Moment and couple

Apart from the tendency to move an object along its line of action a force also has a second effect: think of a billiard ball that is hit by your cue right in the middle: it will move straight forward. Now hit it slightly off-centre. It will again move straight forward, but also start to rotate around its axis (Figure 2). This rotating effect of a force is called a *moment*. The effect is expressed by:

$$M = F \bullet d \tag{6}$$

with F the force (N), d the distance between the centre of the billiard ball and the line of action of the cue (m) and M the moment. The distance d, also called the *lever arm* of the force, must be perpendicular to the line of action of the force. Note that a moment always is related to the point of which you would like to know the rotating effect of the force, in this case the centre of the billiard ball.

Using formula (6), the unit of M is Nm. (Note that mN is not the correct notation of this unit, since it means milli-Newton)

A moment allows you to open a door, to unscrew bolts, and to spin a ball. To increase the effect of a moment you can either increase the force or the distance. Our body uses this effect: since the muscle force is limited, enlarging d is an additional strategy to increase the rotating effect, that is why you have a patella; it increases the

effect of your quadriceps muscle and makes it possible to extent your knee more powerfully.

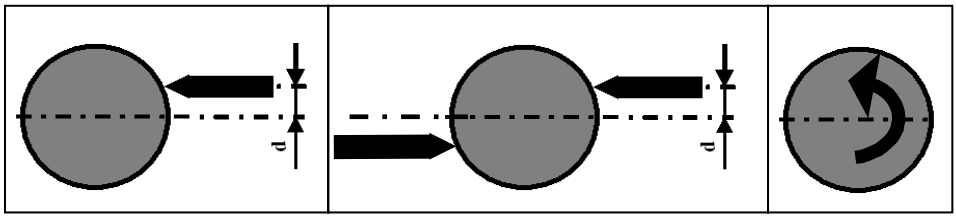


Figure 2. When a force is acting eccentrically over a distance *d*, a moment appears, resulting in a rotation (left). Two parallel forces in opposite direction are called a couple (middle), also denoted by a curved arrow (right).

A specific situation appears when two parallel and equal forces with an opposite line of action are acting on an object (Figure 2, middle). The net action of the forces in horizontal movement is zero, they only have a rotating action caused by their moment, which is twice the size of the moment when one force is acting. Such a pair of forces is called a couple and denoted by a curved arrow (Fig. 2, right), since its line of action (the rotation line) is pointing out of the plane. A couple has a magnitude and orientation and therefore is a vector.

In statics, moving a force along its line of action - translation - does not have any influence. It does not matter whether you push an object or pull it. However, moving a force perpendicular to its line of action does have an influence. Rotation is affected by it, since its arm is changing. Moving a couple over an object is allowed. Although the arm of one force is increasing, the arm of the other force is decreasing by the same amount, leaving no net change. Therefore a couple is called a free vector.

In practice the terms couple and moment are used inaccurately and often exchanged, which is not logical, since a couple is a free vector, but a moment is related to a point on an object. The term couple moment is often used instead of couple, which is also confusing.

2. Static Equilibrium

Now we know the effects of forces, either translation (movement along the line of a force) or rotation, we can apply Newton's First Law. According to this law, if an object is not subjected to a force, this object will not start to move or change its velocity. Since, in statics, we consider all objects to be at rest, no net forces should act on them. Also the sum of their moments should also be zero to avoid rotation. So in statics the sum of all forces should be zero and the sum of all moments in relation to a certain point on the object is also zero. This means that the object has no degree of freedom to move.

When we consider a two-dimensional situation, each force can be resolved into two components, a horizontal one, F_x, and a vertical one F_y. For static equilibrium all horizontal forces should be zero and all vertical ones should be zero.

This leads to three equations of equilibrium:

$$\begin{aligned} \Sigma F_x &= 0 \\ \Sigma F_y &= 0 \\ \Sigma M_p &= 0 \end{aligned} \tag{7}$$

where Σ is capital sigma and means 'the sum of'.

Note that when the sum of all moments is zero in relation to a certain point on the object, it will also be zero to another point on that object. So the reference point is a free choice, but must be used subsequently for all forces.

In the three-dimensional world, which we will not consider in this paper, six equations of equilibrium exist:

$$\begin{aligned} \Sigma F_x &= 0 \\ \Sigma F_y &= 0 \\ \Sigma F_z &= 0 \\ \Sigma M_x &= 0 \\ \Sigma M_y &= 0 \\ \Sigma M_z &= 0 \end{aligned} \tag{8}$$

3. Free-Body Diagram

In 2D, in order to apply the three equations of equilibrium, we have to know which forces are acting upon an object. We already found out, that not only obvious external forces are active, like gravity, but also hidden ones, like the force of a seat acting upon your bottom. To make this last type of force visible, we have to free the object from its surroundings. After that we have to draw all forces of the surroundings that act upon this object. These forces are unknown, but by applying the three equations of equilibrium we can calculate them and so complete our static analysis. Thus, all forces that are acting upon an object are known.

Then we can proceed to dynamics or mechanics of materials to find out the movement of this object or its deformation.

The main problem in making an object free from its surroundings is to determine the reaction forces from the surroundings. In Figue 3 a list of connections between the object and its surroundings (the supports) is shown, along with the replacing reaction forces and couples. As a general rule, if a support prevents translation of an object in a given direction, then a force is present on the object in the opposite direction. Similarly, if a rotation is prevented, a couple is exerted on the object.

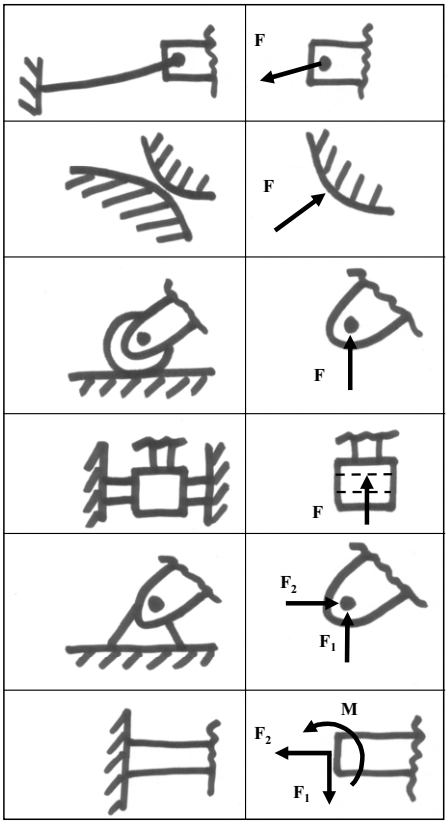


Figure 3. Overview of possible supports and their replacing forces and couples.

Making an object free from its surroundings is a very delicate action, since we have to decide which object we select. If we select a whole human body as our object, and we are interested in the forces acting in our shoulder joint when lifting a bag, we will not be able to calculate these forces, because they are hidden inside the object. So in that case we have to select the upper limb as our object and free it from the rest of the body. We again have to determine the reaction forces of the rest of the body. Most supports consist of joints, which in general cannot be characterised as regular hinge-type connections. We can distinguish 6 types of joints (Figure 4):

1. Ball and socket joints, like the hip joint, allow three rotations and no translations.
2. Condyloid (ellipsoid) joints, like the knee, allow in general rotation around one axis and some rotation around another axis. When the knee is extended there is no axial rotation possible, when it is flexed some axial rotation is possible.
3. Saddle joints, like the thumb (between the 1st metacarpal and trapezium), allow rotation around two axes.
4. Hinge joints, like the elbow (between the humerus and the ulna), act like a door hinge, allowing only flexion and extension around one axis

5. Pivot joints, like the elbow (between radius and ulna). This is where one bone rotates about another.
6. Gliding joints, such as in the carpals of the wrist (not shown in Figure 4). These joints allow a wide variety of movement, but not over a large distance.

A moment allows you to open a door, to unscrew bolts, and to spin a ball. To increase the effect of a moment you can either increase the force or the distance. Our body uses this effect: since the muscle force is limited, enlarging *d* is an additional strategy to increase the rotating effect, that is why you have a patella; it increases the effect of your quadriceps muscle and makes it possible to extent your knee more powerfully.

A moment allows you to open a door, to unscrew bolts, and to spin a ball. To increase the effect of a moment you can either increase the force or the distance. Our body uses this effect: since the muscle force is limited, enlarging *d* is an additional strategy to increase the rotating effect, that is why you have a patella; it increases the effect of your quadriceps muscle and makes it possible to extend your knee more powerfully.

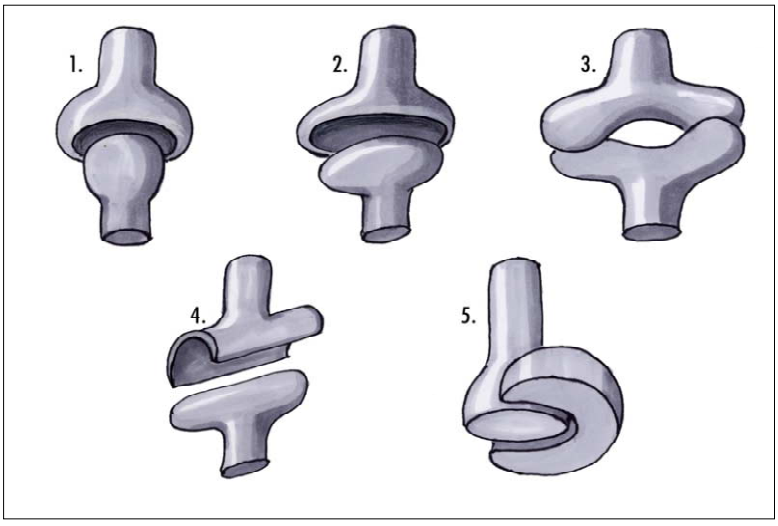


Figure 4. Schematic drawing of 5 joint types, ball and socket (1), condyloid (2), saddle (3), hinge (4) and pivot joints (5).

Apart from separating the joint surfaces we also have to cut through muscles to free our upper limb from the rest of our body. A muscle can be categorised as a cable and must be replaced by a single force with an orientation from origin to insertion.

Let us now summarize the procedure for determining the shoulder joint and muscle forces in case we think we need to lift some weights (of in total 100 kg, see Figure 5, left). First we must define a positive *x*-axis, *y*-axis and direction of moments (Figure 5, right). Then we need to free the object of interest, in this case the upper limb.

The upper limb is connected to the rest of the body by the shoulder joint, which can characterised as a ball and socket joint in 3D, and in 2D as a hinge. So the shoulder joint can be replaced by two forces, F_1 and F_2. The upper limb is also connected with muscles to the rest of the body. In this situation the deltoid muscle will be (very) active. A muscle can be replaced by a single force, pointing along its line of action, F_s. Now we can make the free-body diagram (Fig. 5, middle), including all forces acting on the upper limb: half of the weight of the dumb-bell (50 kg x 10 m/s^2 = 500 N), the weight of the upper limb (assumed to be 2.5 kg, resulting in 25 N), the unknown muscle force F_s, and the unknown joint reaction forces F_1 and F_2. To be able to use our equations of motion, we first have to split F_1, F_2 and F_s in a force, acting in x-direction and a force, acting in y-direction using formula (4).

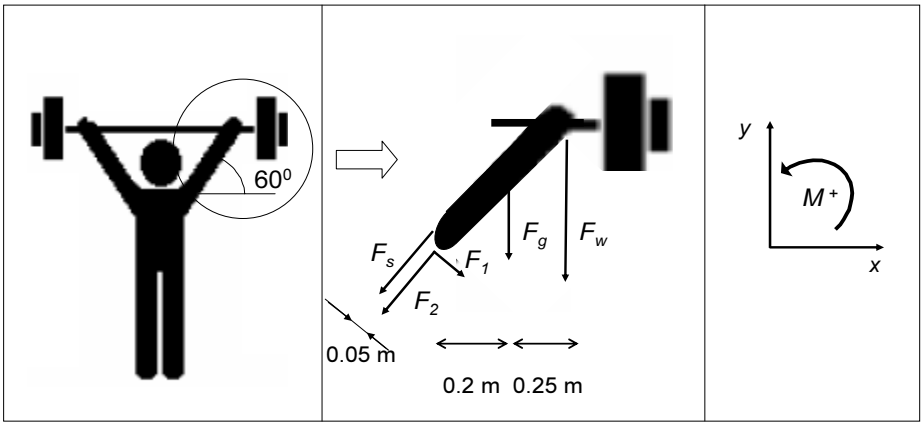


Figure 5. How to calculate shoulder joint and deltoid forces when lifting a weight (left): a free-body diagram is made of the arm (middle), including the orientation of the x- and y-axis and the positive direction of moments (right)

Finally we can calculate the unknown forces using the three equations of equilibrium:

$$\Sigma M_p = 0 \ (p = \text{shoulder joint centre}) \ - F_g \bullet 0.2 - F_w \bullet 0.45 + F_s \bullet 0.05 = 0$$

$$F_s = 4600 \text{ N} \qquad (9)$$

$$\Sigma F_x = 0 \qquad - F_s \bullet \cos 60^0 - F_1 \bullet \cos 60^0 + F_2 \bullet \cos 30^0 = 0$$

$$\boldsymbol{F_1 = \sqrt{3}\, F_2 - F_s} \qquad (10)$$

$$\Sigma F_y = 0 \quad \rightarrow \quad - F_s \bullet \sin 60^0 - F_1 \bullet \sin 60^0 - F_2 \bullet \sin 30^0 - F_g - F_w = 0$$

$$\rightarrow \quad F_1 = - F_s - \sqrt{3}/3 \ (F_2 + 1050) \qquad (11)$$

(9), (10), (11) → $\boldsymbol{F_2 = -262.5\ N}$ (12)

(9), (10), (12) → $\boldsymbol{F_1 = -5055\ N}$ (13)

Obviously we have chosen both F_1 and F_2 in the wrong direction, because the outcome is negative for both forces. So the direction of the unknown forces can be chosen arbitrarily, after the calculation will become clear, if the assumed direction is the proper one.

The magnitude of F_s and F_1 are considerable, corresponding with a mass of 460 and 505 kg, respectively! The main reason for this is the small arm of F_s compared to the arm of F_w; muscle arms are small due to geometric considerations. So muscles must be very strong to compensate for this short arm.

4. Internal Loads

Now we are able to determine the forces, acting on an object and to calculate the magnitude of these forces, next we can analyse two situations:

Firstly we can find out what will happen if an extra force or couple is applied and thus the equilibrium of forces and moments is disrupted. In the 'Dynamics' chapter we will calculate the resulting accelerations.

Secondly we can consider our object to be *not rigid*, so deformable and calculate the deformations that result from the forces and determine, if our object can resist the forces, and thus if it is strong enough. For this last action we need to do an extra step in statics, we have to calculate the *internal loads* and determine what location is the weakest point.

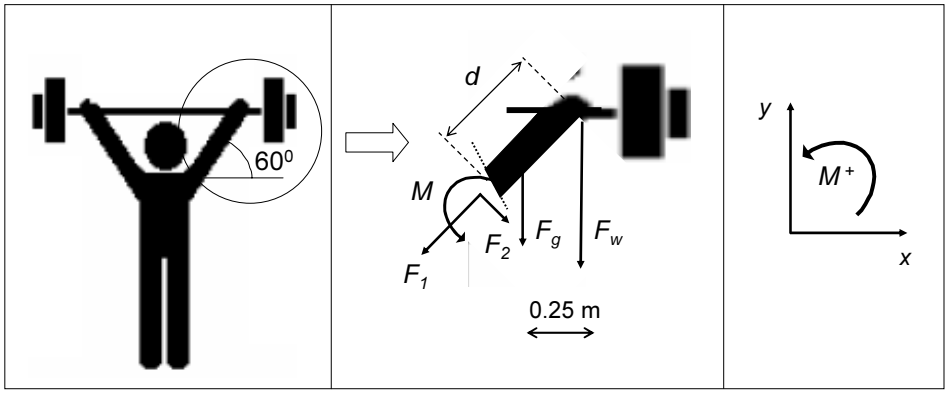


Figure 6. For the body builder on the left the internal forces of his arm are calculated (middle), using a co-ordinate system (right).

The first step we have to do is to isolate our object (Figure 6, left), define a positive *x*- and *y*-axis and direction of moments (Fig. 6, right) and calculate the external

reaction forces (in our case there are none). Then we have to cut our object in two parts and replace one part of our object (in our case the lower left part) by its reaction forces F_1, F_2 and M at the cutting edge (Fig. 6, middle).

The resulting free-body diagram (Fig. 6, middle) contains three unknowns, F_1, F_2 and M that can be calculated using the three equations of equilibrium. Note that the position of the cut is defined by the distance d between the cut and the upper right end of our object. By varying this distance we can scan the entire arm and thus calculate the internal forces at any spot. It then is possible to determine the site at which the highest loads will occur and thus the site where failure will occur first:

$\Sigma M_q = 0$ (q = cutting edge site)

$$\rightarrow - F_g \bullet d \bullet \cos 60^0 - F_w \bullet (d \bullet \cos 60^0 - 0.25) + M = M = 262.5 \bullet d - 125 \text{ Nm} \qquad (14)$$

$\Sigma F_x = 0$ $\rightarrow - F_1 \bullet \cos 60^0 + F_2 \bullet \cos 30^0 = 0$

$$\rightarrow \boldsymbol{F_1 = \sqrt{3}\, F_2} \qquad (15)$$

$\Sigma F_y = 0$ $\rightarrow - F_1 \bullet \sin 60^0 - F_2 \bullet \sin 30^0 - F_g - F_w = 0$

$$\rightarrow F_1 = - \sqrt{3}/3\, (F_2 + 1050) \qquad (16)$$

(15), (16) $\rightarrow \boldsymbol{F_2 = - 262.5\ N}$ (17)

(15), (17) $\rightarrow \boldsymbol{F_1 = - 4547\ N}$ (18)

Again, both F_1 and F_2 are acting in the opposite direction as was presumed.

Now we can find out, where the most critical spot in the upper arm is. Both F_1 and F_2 are constant, M is maximal for d = maximal. So the highest load will act where d is maximal, which is the lower left end of the upper arm.

References

[1] G.L. Lucas, F.W. Cooke, E.A. Friis, A Primer of Biomechanics, Springer-Verlag, 1998

[2] R.C. Hibbeler, Statics and Mechanics of Materials (Second edition), Pearson – Prentice Hall, 2004

[3] N. Özkaya and M. Nordin, Fundamentals of Biomechanics; Equilibrium, Motion and Deformation (Second edition), Springer Science+Business Media, 1999

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-13

I.2. Mechanics of Materials

Prashant K. SHARMA[a]
[a] *Department of BioMedical Engineering, University Medical Center Groningen, The Netherlands*

Abstract. Mechanics of materials is the science of forces applied on a body and response of the body in terms of deformation. Different type of loadings on bodies with different geometries or made of different material give rise to different deformations. Last but not the least, this science allows to predict the failure of a body under certain loading condition hence makes it possible to optimize the design for that particular condition.

Keywords. Stress, strain, stiffness, torsion, bending, ductile, brittle

Introduction

This field of science deals with the relation of external loads acting on a body to the induced internal forces experienced by the material of that body. Internal forces either result in deformation i.e. gross changes in the external shape or failure.

1. Force and Deformation

Assume a bucket of mass 1 *kg* hanging on a rope, then from intuition we know that the bucket is pulling the rope downwards i.e. exerting a ***force*** due to gravity. The magnitude of the force is equal to the mass of the bucket times the acceleration due to gravity, *g*, with a value of 9.8 $m\ s^{-2}$. Thus the force experienced by the rope ($F = m \times g$) is 9.8 $kg\ m\ s^{-2}$, or 9.8 *N* (≈ 10 *N*). In other words, one Newton of force will act if 102 (≈100) *g* of mass is hanging from the rope.

Let us now analyze a simple situation where mass *m* is hanging on a cylinder with radius of *r* (Figure 1, left) and a mass of *2m on* a cylinder with radius *2r* (Figure 1, right). *Can we say which cylinder will fail first?* A force of *mg* is acting on the left cylinder and twice the force *2mg* on the right cylinder. Cross-sectional area of the cylinder on which this force is active is also different, the left cylinder has an area of πr^2 (pi, a constant = 3.142 x radius squared) whereas the right cylinder has 4 times the area i.e. $4\pi r^2$. We have to invoke the concept of ***Stress***, denoted by σ, (sigma) in order to answer this question.

$$\sigma = \frac{F}{A} \tag{1}$$

where F is the force and A is the cross-sectional area on which the force is active. Unit of σ is N/m^2, also called Pascal (Pa)

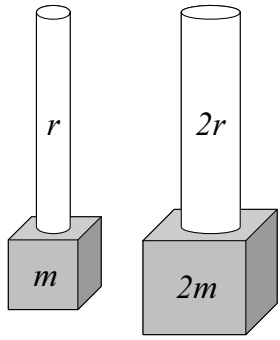


Figure 1. Loading situation of two cylinders

The left cylinder is experiencing a stress of $mg/\pi r^2$ whereas the right cylinder experiences half that much ($mg/2\pi r^2$), therefore we can now say that the left cylinder will fail first.

The situation we analyzed is called ***uniaxial state of stress*** because the cylinders are experiencing stress along one axis and the stress is called ***normal stress*** because the force is acting perpendicular to the cross-sectional area (Figure 2A). Normal stress always tries to either pull, called ***tension***, or push, called ***compression***, a body (Figure 2B).

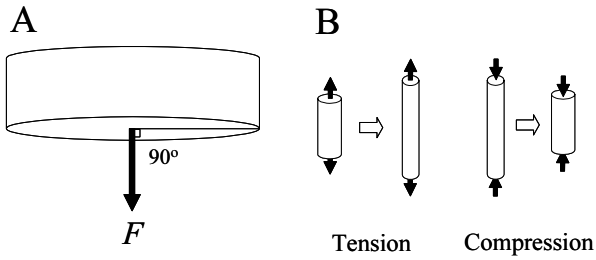


Figure 2. Normal stresses

Material inside the cylinders as shown in Fig 1 responds to this tensile stress and deforms by stretching itself resulting in increase in the length of the cylinder at the cost of a decrease in the diameter. This deformation is expressed in terms of ***normal strain***,

denoted by ε (epsilon). Strain is a concept where the deformation due to the applied stress is normalized by the initial dimensions of the body.

$$\varepsilon = \frac{\Delta L}{L} \tag{2}$$

where L is the original length and ΔL is the increase in length (Figure 3A). Strain is a dimension-less quantity since both ΔL and L are expressed in the same unit.

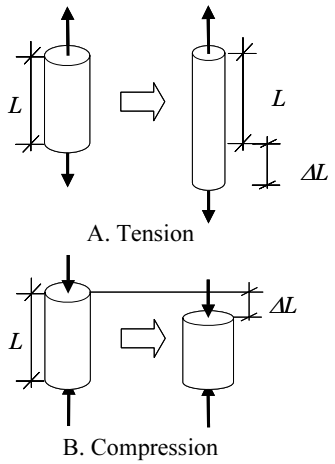


Figure 3. Normal strain

In case of compression (Figure 3B) the formula to calculate the strain remains the same but the meaning of ΔL changes into a decrease in height thus takes a negative sign.

Why can't we simply measure the applied force, in Newtons and deformation in centimeters instead of using stress and strain? The concept of stress and strain helps us to compare the mechanical properties of say stainless steel and aluminum cylinders without having to make them exactly of the same size and shape and applying exactly the same force. It helps us to compare our results with laboratories from the other side of the world and most importantly by measuring the mechanical properties of a small piece of metal in the laboratory we can predict the performance of a long beam of the same metal placed in a bridge.

2. Mechanical Properties of Materials

In day to day life we talk about the strength of different materials and compare them by saying that one is stronger, tougher or harder than the other.

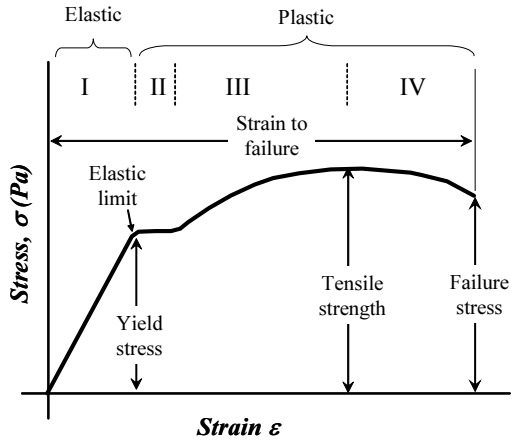


Figure 4. Stress – strain diagram

What do we mean by all these words- strength, toughness and hardness? Strength, toughness and hardness depend on the ability of a material to sustain load without undue deformation. If we keep on increasing the force on the left cylinder in Figure 1 then we know that the cylinder will deform, but do we know at what applied force it will fail and how long it will become? The answer is that it depends on the strength of material the cylinder is made of. Strength of a material can be determined by performing either a ***tensile*** or ***compression test*** and record the results in terms of stress and strain. During a tensile test, an increasing amount of stress is applied to a standard specimen and the resulting strain is measured, these stress and strain data are plotted on a so called ***stress-strain diagram*** (Figure 4). Starting from zero, as the applied stress is increased the strain increases linearly up to a limit called the ***elastic limit***. In this region I the material shows elastic behavior and follows *Hooke's Law* (Eq. 3), where the strain is linearly proportional to the applied stress and the constant of proportionality E is the Young's modulus having the same units as stress i.e. Pascal

$$\sigma = E\,\varepsilon \tag{3}$$

If the applied stress is removed within the elastic limit then the material regains its original size meaning that the strain reduces to zero and there is no permanent deformation in the specimen (Figure 5A). At the end of the elastic region, the ***plastic*** region starts where a slight increase in stress above the elastic limit i.e. ***yield stress*** results in deformation almost without an increase in stress, this process is called ***yielding*** (region II). In region III the applied stress again increases, resulting in an increase in strain. The point where the maximum amount of stress is supported by the specimen is called the ***tensile strength***. In region IV the supported stress continuously decreases leading finally to failure. If the applied stress is removed in the plastic region, the strain does not reduce to zero but some residual strain remains in the form of permanent deformation (Figure 5B). The start of the plastic region is difficult to determine, since the linear increase of the stress is gradually changing into the yielding phase. For practical reasons the start of the plastic region is defined by the point where

if the applied stress is removed the specimen keeps a small amount of permanent deformation ($\varepsilon = 0.002 = 0.2\ \%$); this stress is therefore called the yield stress $\sigma_{0.2}$.

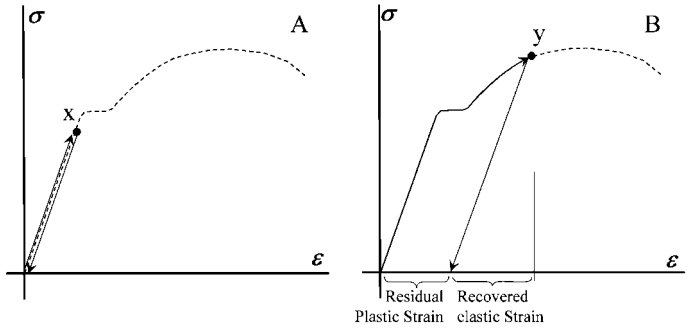


Figure 5. Elastic and plastic deformation

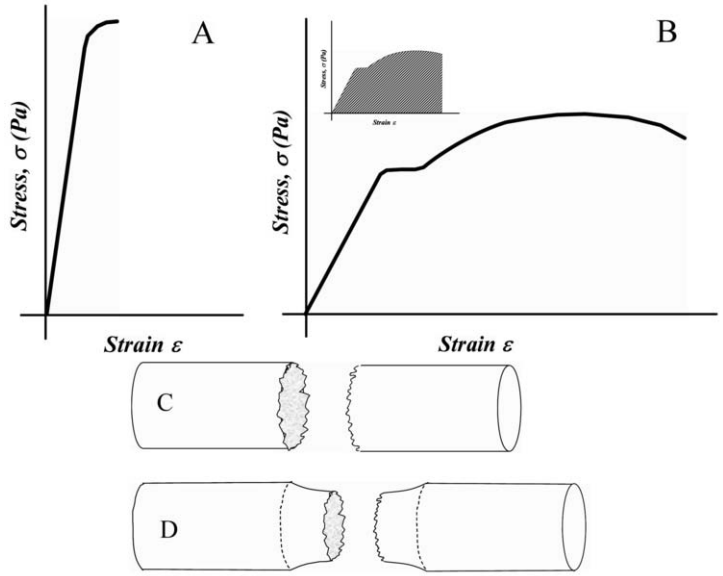


Figure 6. Different materials and their failure

Figure 6 shows two different type of materials; material A (Figure 6A) is ***strong*** because it can support very high stresses but this material is also *brittle*, meaning that it cannot be given any permanent (plastic) deformation, just after reaching the maximum stress the material fails giving rise to a flat fracture surface (Figure 6C). Material B

(Figure 6B) on the other hand is less strong, but ***tough*** and ***ductile*** because it deforms a lot before failure meaning that it can absorb a lot of energy by deforming permanently before failure. This amount of energy is represented by the area under the curve, Figure 6B inset.

For a tough material the cross sectional area reduces at a localized area and causes a large plastic deformation, called ***necking***, before the material fails (Figure 6D).

While constructing a bridge, engineers choose a tougher material because then they can see their bridge change its shape before it finally fails. If they use a strong and brittle material then the bridge will fail catastrophically without any prior warning. ***Hardness*** tells us the resistance of a material to scratching by another material, this property is important if we do not want our material to wear out during rubbing with other surfaces. This is the reason for use of hard materials as artificial joint surfaces and drill bits. Similarly cortical bone is stronger and harder to avoid wear and tear whereas trabecular bone is tougher to absorb energy during jumping and falling.

Till now we have talked about the *tensile strength* of a material, but the *normal stress* can also be compressive in nature. Materials like rubber can deform to up to 10 times their original size under tension where as concrete cannot (Figure 7). Concrete has a very low *tensile strength* but have very high *compressive strength;* that is why steel rods are embedded in concrete to bear the tensile part of the stress in a building column. In the human body most of the bone structure and teeth are good at bearing compressive stresses whereas ligaments are specialized structures meant to bear tensile stresses.

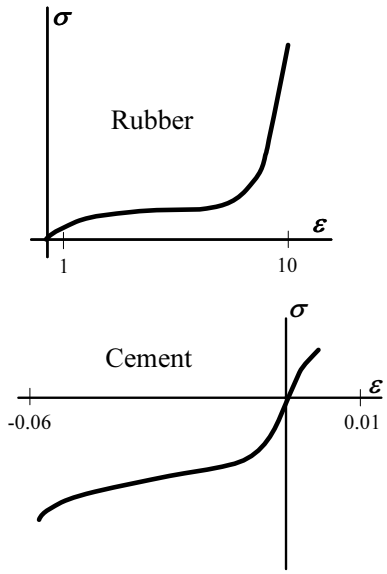


Figure 7. Materials to bear tensile and compressive stresses

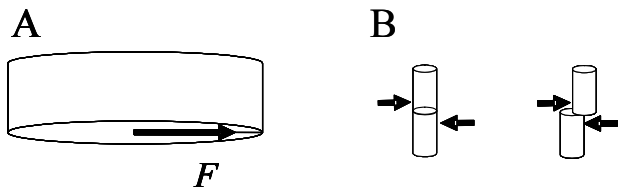


Figure 8. Shear Stress

3. Shear Stress and Strain

We rarely feel the presence of yet another type of stress, which is called ***shear stress***. This occurs when the force is active parallel to the cross-sectional area (Figure 8A) and tries to slide one part of the body in relation to the other part (Figure 8B). Shear stress is denoted by τ (tau) and also equal to the force acting per unit cross-sectional area with the unit Newton (N) (Eq.4).

$$\tau = \frac{F}{A} \tag{4}$$

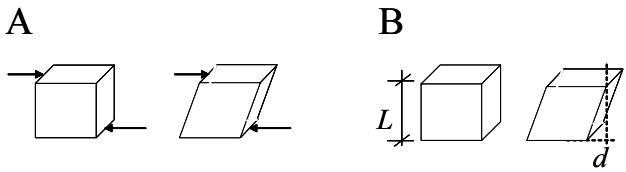


Figure 9. Shear Strain

The material reacts to this shear stress as shown in Figure 9A and the ***shear strain***, denoted by γ (gamma), is calculated using equation 5 (Figure 9B) and again it is a dimension-less number

$$\gamma = \frac{d}{L} \tag{5}$$

A material experiencing shear stress also demonstrates a ***shear stress-strain curve*** very similar to a normal stress strain curve, with an elastic region where it follows *Hooke's law* and shear strain is linearly proportional to shear stress and having a ***modulus of rigidity***, *G*, (Eq.6). Above the elastic limit the material deforms permanently

$$\tau = G\gamma \tag{6}$$

Normal stresses (both tensile and compressive) and shear stress are the only two types of stresses which can exist inside the material and hence the material manifests only normal and shear strain. All the different types of external forces acting on the body induce normal or shear stresses, although active in more than one direction.

There are two more ways in which external forces can be active on a body; one in which the cylinder is twisted called *torsion* (Figure 10) and second where the cylinder is bent called *bending* (Figure 14A). We shall now analyze how these external forces translate into internal stresses.

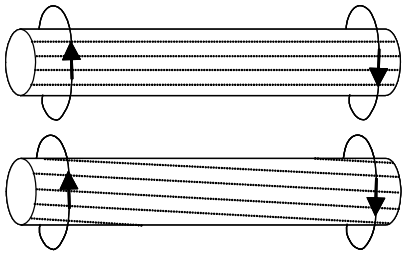


Figure 10. Torsion

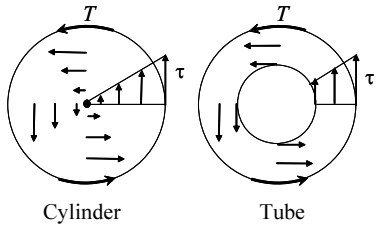


Figure 12. Shear stress distribution

If the cylinder surface is marked with a black square (Figure 11A) and external torsional moment, ***T***, is applied then the cylinder deforms in such a way that the square is transformed to a rhombus shape (Figure 11B&C). We have seen that this transformation takes place only when the forces are active parallel to the cross-sectional area i.e. shear stress (Figure 8, 9). Thus externally applied torsional forces induce shear stresses in the cylinder hence the cylinder shows shear strain ($\gamma=d/L$, Figure11B,C,D).

If we know the magnitude of the externally applied torsional moment, also called torque, on a cylinder of radius r then we can calculate the induced maximum shear stress at the cylinder surface using Eq.7.

$$\tau^{\max} = \frac{2T}{\pi r^3} \tag{7}$$

If we look at the cross section of the cylinder then we will see that the induced shear stresses vary linearly from zero in the center to a maximum value at the surface, whereas in a hollow tube the stress is a finite value at the inside diameter with maximum at the outer diameter (Figure 12). The maximum shear stress at the hollow tube surface with outside radius, r_o, and inside radius, r_i, is calculated using Eq. 8. For the same applied torque and outside tube diameter, as the inside tube diameter increases the maximum induces shear stress at the outside surface increases.

$$\tau^{max} = \frac{2Tr_o}{\pi(r_o^4 - r_i^4)} \tag{8}$$

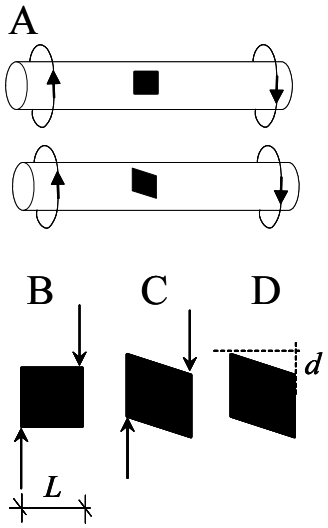


Figure 11. Internal shear stress induced due to external torsional forces.

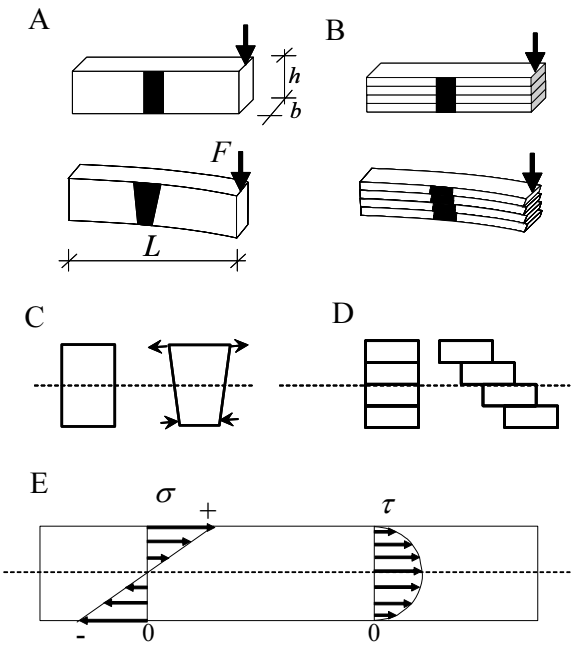


Figure 14. Normal and shear stresses induced by the external bending moment

If we know the material properties then we can calculate the shear strain at each point, this shear strain cumulatively results in twisting of the cylinder. The ***angle of twist***, θ (theta), (Figure 13) for a cylinder made of material with modulus of rigidity G and length L experiencing an external torque of, T, can be calculated using Eq. 9. The longer and thinner a cylinder greater is the angle of twist.

$$\theta = \frac{2TL}{\pi G r^4} \tag{9}$$

If a square cross-section beam is fixed at one end and a force is applied at the other end perpendicular to its axis then there is a ***induced bending moment M*** in the bar. A bending moment is the applied force multiplied by the distance of the force from the fixed end ($M=F.L$) and is thus reported in $N.m$ (Figure 14A).

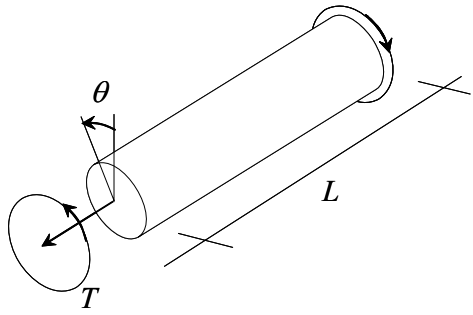


Figure 13. Angle of twist due to torsion

This bending moment induces both normal and shear stresses in the beam; Figure 14A shows how a rectangle drawn on the side of a beam changes to a trapezoid due to stretching of the beam above the central axis and compression below it. This means that tensile stresses are induces on the beam above and compressive stresses below the central axis (Figure 14C&E). The maximum tensile or compressive stress at the top or bottom surface can be calculated by Eq. 10.

$$\sigma^{max}_{\substack{Tensile\,/\,Compressive \\ Top\,/\,bottom}} = \frac{6M}{bh^2} \tag{10}$$

If we imagine the beam to be constructed with a pile of much thinner beams (Figure 14B) then by intuition, the same as bending a thick book along its binding axis, we know that these thin beams will slide and move relative to each other when a bending moment is applied (Figure 14D). This can only happen if there is a force active along the interface between the thin beams i.e. shear stress is being active. Even when the beam is not composed of thinner beams we can see that these shear stresses will be induced, the proof for this can be seen from the longitudinal fracture through the center visible in old wooden beams used to support the roof. Induced shear stresses are maximum at the central axis and drop to zero at the top and bottom surfaces (Figure 14E). Maximum shear stress at the central axis can be calculated by equation 11 where V is the internal shear force resulting from the applied bending moment.

$$\tau^{max}_{centralaxis} = \frac{3V}{2bh} \tag{11}$$

The deformation caused by a bending moment is deflection "v" and slope of the beam. Both parameters can be calculated at any point on the beam defined by distance x from the fixed end by solving a differential equation (Eq. 12), where the Young's modulus (E), moment of inertia (I) and bending moment at the point ($M(x)$) is known.

$$EI\frac{d^2v}{dx^2} = M(x) \tag{12}$$

Some typical solutions to the differential equation for a beam loaded with force, F (N), bending moment, M (Nm) and distributed load, Q (N/m) is presented in Figure 15.

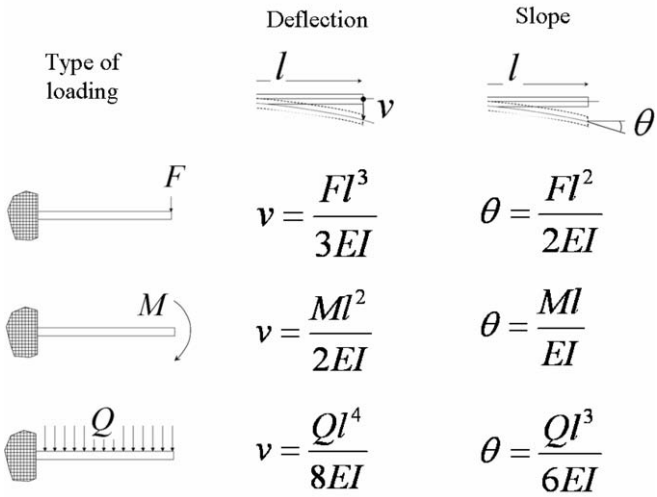


Figure 15. Deflection and slope of a beam under different loading situations

4. Failure

Failure can now be decided in terms of the maximum stress above which we consider our material to be rendered useless. While constructing a bridge or making a load bearing implant, we would never like that our structure deforms permanently thus we can say that the upper limit of any stress induced in the material should be the ***yield stress.*** In building and bridge construction a margin is given to be always on the safe side i.e. the calculated induced stresses in the material should be always, say, less than 70% of the ***yield stress*** and this is called the ***maximum allowed stress*** ($\sigma_{allowed}$, $\tau_{allowed}$).

In simple external loading situation like *tension*, *compression*, *torsion* or *bending* we can calculate the maximum internal stresses using equations 1, 4, 7, 8, 10 and 11 and if these stresses are higher than the allowed stresses then we know that the structure will ***fail***. *In real situation two or more of these external loading takes place simultaneously, what happens then?* One such simple example of combined loading is a spherical liquid storage tank, for example the heart. The tank is pressurized and if we look at a small square piece of material on the tank surface we can see that the square experiences tensile stresses on both sides as shown in Figure 16. To handle these situations with combined loading we use ***Mohr's circle***. Most of the materials have to be analysed 3-dimensionally but since they are geometric in nature, using the symmetry we can convert them to 2 dimensions like we did for the spherical storage tank (Figure 16).

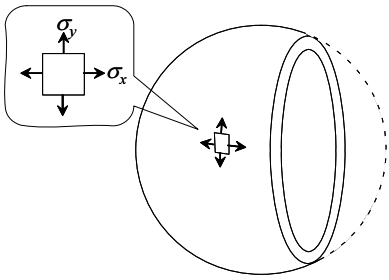


Figure 16. Spherical storage tank

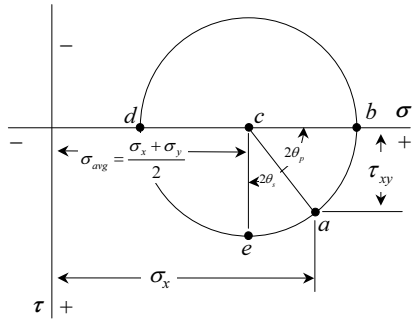


Figure 17. Mohr's circle

Mohr's circle is a plot between the normal stress (σ) on the horizontal or *x-axis* and the shear stress (τ) on the vertical or *y-axis* (Figure 17). Tensile normal stresses are positive and compressive are negative, for the shear stress we follow a sign convention shown in Figure 18 where a very small volume element inside the bulk of the material is shown. Coming back to Figure 17, the position of the center of the circle *c* is the average of the two normal stresses (σ_x *and* σ_y and *reference point a* can be plotted with τ_{xy} on the *y-axis* and σ_x on the *x-axis*. Now we can join points *c* and *a* to get the radius and then the circle can be drawn with a radius given by Eq. 13. From this Mohr's circle for a given situation we can find out the maximum normal stress, point *b*, and shear stress, point *e,* active in the material.

$$R = \sqrt{\left(\frac{\sigma_x - \sigma_y}{2}\right)^2 + \tau_{xy}^2} \tag{13}$$

If we take a simple situation where a tensile stress of 100 Pa is active on a square cross section beam then the *state of stress* at a very small volume element will

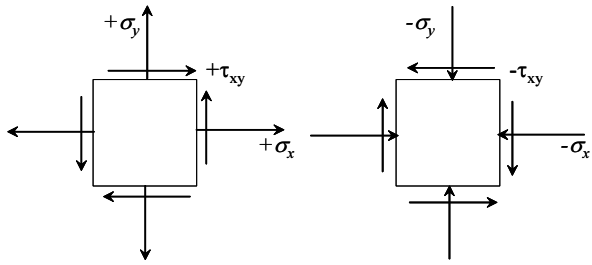


Figure 18. Sign convention

be as shown in Figure 19A. Surprisingly, the Mohr's circle (Figure 19B) of this volume element shows that there exists an orientation (45° clockwise, Fig19C) where

maximum shear stress of 50 Pa is active. Therefore for supporting 100 Pa of stress we have to choose a material for which the $\sigma_{allowed}$ is at least 100 Pa and $\tau_{allowed}$ is at least 50 Pa, all the material with $\tau_{allowed}$<50Pa cannot be used even when they can support the normal stress of 100 Pa and will deform (Figure 19D).

Looking at a more complex combined loading of compressive and torsional nature, Figure 20 shows a cylindrical beam loaded with 100 *Pa* compressive stress and a torque giving rise to a shear stress of -50 *Pa* on its surface. From point *d* in Mohr's circle we can see that the maximum normal stress induced is compressive and 120 *Pa* and point *e* tells us that the maximum shear stress is about 71 *Pa*. Therefore a material for this application should be capable of bearing a compressive stress > 120 *Pa* and a shear stress > 71 *Pa*.

From simple loading conditions Mohr's circle can be constructed and the highest normal and the highest shear stress can be determined and thus prediction of whether the construction will deform or totally collapse can be made.

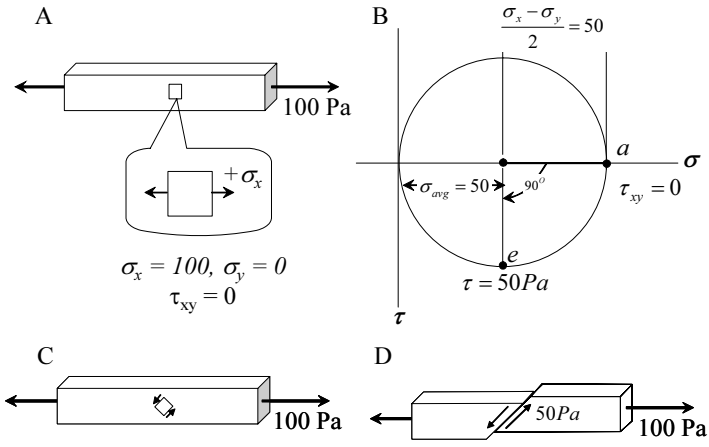


Figure 19. Mohr's circle for *uniaxial tension*

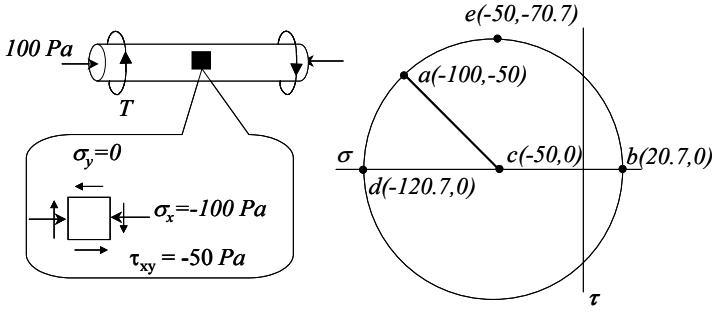


Figure 20. Mohr's circle for *combined compressive and torsional loading*

References

[1] R.C. Hibbeler, Mechanics of Materials. Printice Hall International, New Jersey, 2003
[2] N. Özkaya and M. Nordin, Fundamentals of Biomechanics: Equilibrium, Motion, and Deformation. Springer, 1999

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-27

I.3. Dynamics of Human Movement

Bart (H.F.J.M.) KOOPMAN[a]
[a]University of Twente
Dept of Biomechanical Engineering
Enschede
The Netherlands

Abstract. The part of (bio)mechanics that studies the interaction of forces on the human skeletal system and its effect on the resulting movement is called *rigid body dynamics.* Some basic concepts are presented: A mathematical formulation to describe human movement and how this relates on the mechanical loads acting on the skeletal system. These equations of motion depend on the mechanical properties of the skeletal system, such as dimensions and mass distribution. It is applied to describe and analyze human gait.

Keywords. Human movement, rigid body dynamics, gait analysis

Introduction

Human movement dynamics relates the applied forces or moments of force (i.e. the *kinetics*) to the movement (i.e. *kinematics*) of the skeletal system. In the control loop to displace body parts (Figure 1), the Central Nervous System (CNS) is the controller that activates the motors (muscles). The movement dynamics are represented by the skeletal system, and the resulting movement is measured and fed back into the CNS.

In this chapter we focus on the skeletal system only. The part of (bio)mechanics that studies this is called *rigid body dynamics.* Some basic concepts are presented: A mathematical description of human movement; how this depends on the mechanical loads acting on the skeletal system and how this depends on the mechanical properties of the skeletal system, such as dimensions and mass distribution. It will be applied to describe human walking.

1. Rigid Body Dynamics

The human body consists of a more or less rigid structure (the bones of the skeleton) to which soft tissues (muscles, fat, organs etc.) attach. A structure is considered rigid when under practical loading situations, the structure does not noticeably deform, i.e. the mechanical stiffness seems to be infinite. Since the soft tissues are not rigid and do deform in practical loading situations, an accurate mechanical description of human

movement involves the description of both soft tissues and skeleton. The description of the dynamics of the human body would involve very complex calculations, since not only movement, but also deformations have to be considered. This approach is only feasible in some very specific applications, for example to study the effect of car crash impacts on brain damage. As it is not possible to measure all soft tissue deformations, some *assumptions* have to be made to be able to study human movement.

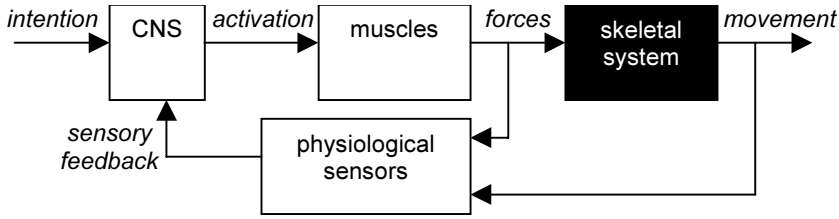


Figure 1. Schematic block diagram of the human motor control system

The most important assumption in rigid body dynamics is that movement only occurs in the joints. As a consequence of this assumption, the structures in-between the joints, the so-called segments, are assumed to be rigid, so they have an infinite large stiffness. All passive soft tissue structures are considered to be rigid as well. The muscles are replaced by the forces they produce. So it is assumed that the human body behaves like a segments model or a linked system of rigid bodies. Apart from the enormous reduction of complexity, this viewpoint has other advantages as well: The rigid body equations of motion, when applied to each segment, are sufficient to describe the dynamics of the entire system.

Rigid body dynamics describes the kinematics (movement) and kinetics (forces) of a given segments model. The equations of motion are the link between these two, which is a set of two first-order nonlinear differential equations for each movement component (or, alternatively, a single second order equation that includes a dependency on the second time derivative, the acceleration). For a given segments model the forces and movements provide interchangeable information: The forces can be calculated when the movements are known (i.e. inverse dynamics approach) and the movements can be calculated when the forces are known (i.e. direct or forward dynamics approach).

2. Kinematics

2.1. Medical Motion Description

Two bones can move with respect to each other by virtue of the joint in between. In the human body, only the synovial joints allow for large movements. In principle, a bone has six Degrees-of-Freedom (DOF) of motion with respect to the other bone: three rotations and three translations. The motions of the joint are limited by passive structures like the articular surfaces and the ligaments. These passive structures pose

restraints to the joint motions: Though motions in the direction of the restraint are still possible, these motions will be very small.

For example, cartilage can be compressed a few millimeters. Most of the time these small motions are neglected, the motion is said to be *constrained*. For each constraint the number of DOF diminishes by one. As soon as constraints come into the picture, one has already started modeling the joint. In Figure 2 all combinations of (constrained) rotations and translations are shown. Many of them are merely hypothetical, and will not be found in the human body.

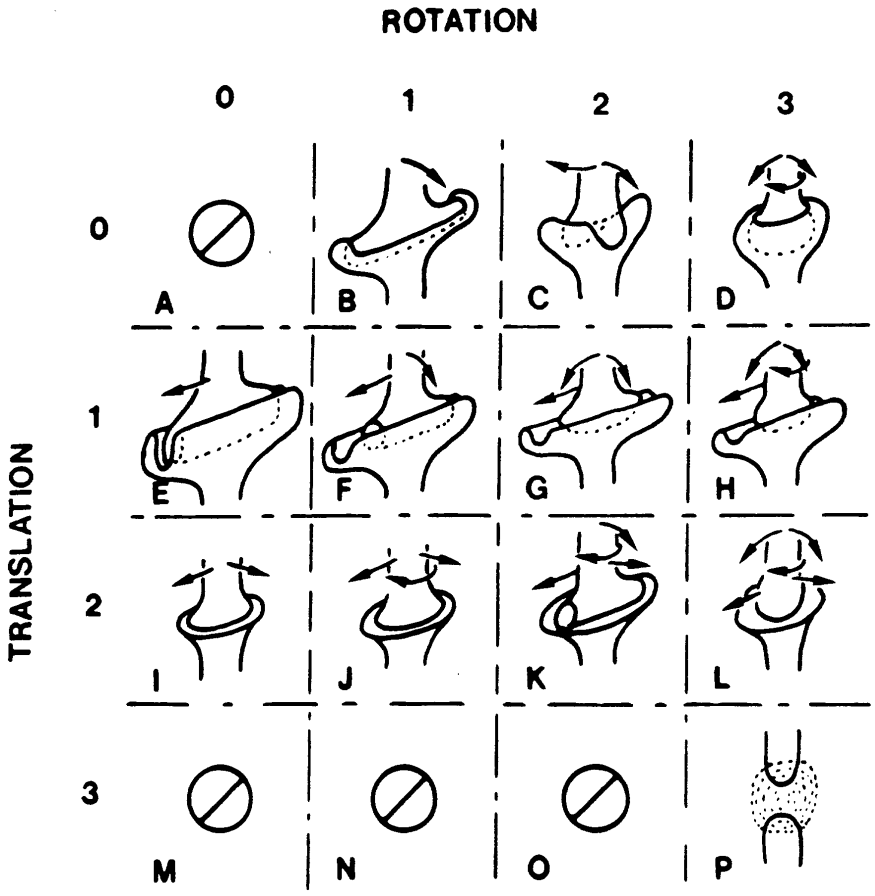


Figure 2. Combinations of rotations and translations in joints. Many joints shown are hypothetical and are not found in the human body.

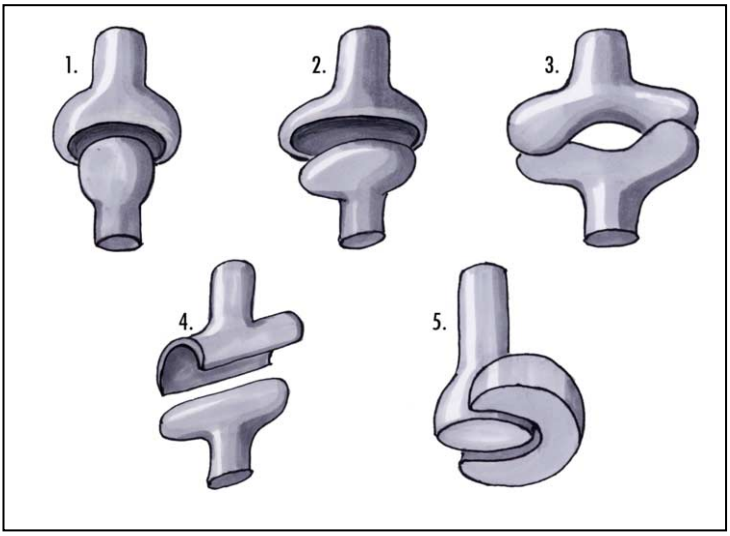


Figure 3. Schematic drawing of 5 joint types, ball and socket (1), condyloid (2), saddle (3), hinge (4) and pivot joints (5).

Traditionally, joints have been studied (e.g. by Fick [4]) by comparing them with standard revolute joints: Hinges, spherical joints, ellipsoidal or condyloidal joints, saddle joints (see Figure 3). The shoulder joint and the hip joint behave approximately as spherical joints only permitting three rotational DOF. The elbow joint and finger joints can be regarded as hinge joints with one rotational DOF. The first metacarpo-phalangeal (thumb) joint and the ankle joint resemble a saddle joint, having two non-intersecting rotational axes. But often one cannot derive the potential motions from the shape of the articular surfaces.

Since anatomists were the first to study joint motions, medical definitions still dominate the way joint motions are described. The goal of such medical definitions is to distinguish between pathological and normal motion, and to evaluate the outcome of treatment: Is there improvement in the range of motion or not.

As a starting position the *anatomical position* (Figure 4) is used. From this position the motion is defined for each single rotation. For spherical joints, the rotation axes are defined along the axis of the global coordinate system: A vertical axis and two horizontal axes, pointing backward-forward (antero-posteriorly) and side-side (medio-laterally, from the centre of the body to the left or right side). Commonly used terms are flexion-extension, abduction-adduction, and medial and lateral rotation. A problem occurs if the motion is not about just one of the standardized axes, but is a combination of rotations. The order of rotations is not defined, and it is not clear whether the rotation axes move with the bone or not. This causes much confusion about rotation angles, and makes comparison between studies often impossible.

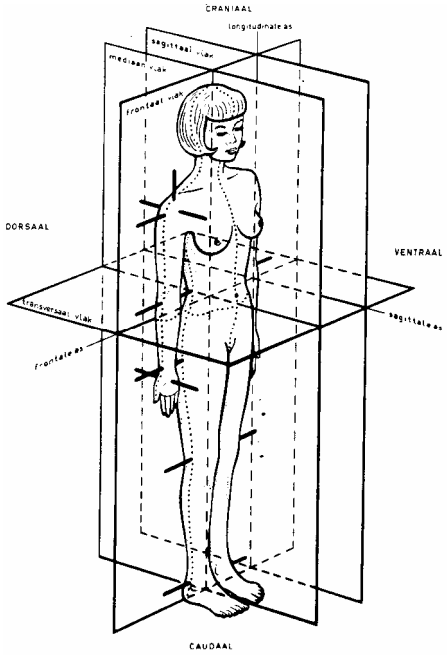


Figure 4. The anatomical positions

2.2. The Segments Model

All models to analyze or simulate human movement are based on segments. These are coupled systems of rigid bodies, with dimensions and inertial properties that are deduced from the dimensions and inertial properties of the human body. Initially, these segments models were used to dimension the dummies used in car crash experiments [3] and were therefore based on averaged data of human bodies. With the introduction of the segments models to computer simulations, a wider variety became possible. The segments models became scalable, with inertial properties usually depending on local segment dimensions and total body weight. In this way, the segments model could be matched to fit each individual [2].

The choice of the number of segments should be large enough to simulate the movement adequately. Too large a number of segments, however, would lead to unnecessary complexity and larger computational efforts. To simulate walking, segments models varying from 3 segments [10] up to 17 segments [6] have been proposed. To further reduce complexity, symmetry between right and left leg is often assumed [1] and the movement is often restricted to the sagittal plane only.

In a model for normal walking, there are segments for the thighs, shanks and feet. The head, arms and trunk (HAT) and the pelvis are modeled as two separate segments. The segments are connected to each other at the joints (Figure 5). Although it is possible to have more than two joints in a segment (e.g. the pelvis), each joint is

connecting just two segments. To define the position of the segments in space, an absolute or reference frame is attached to the floor, with the x-axis pointing in the walking direction, the y-axis pointing upward and the z-axis perpendicular to the xy-plane in the lateral direction. Figure 5 shows an 8-segmental model with numbering of the segments, the joints and definition of the reference frame. It should be noted that the shape of the segments is of no importance as long as the positions of the joints and the mass properties are well defined. In each segment, a local frame is defined, standardized with the help of some bony landmarks on the segment.

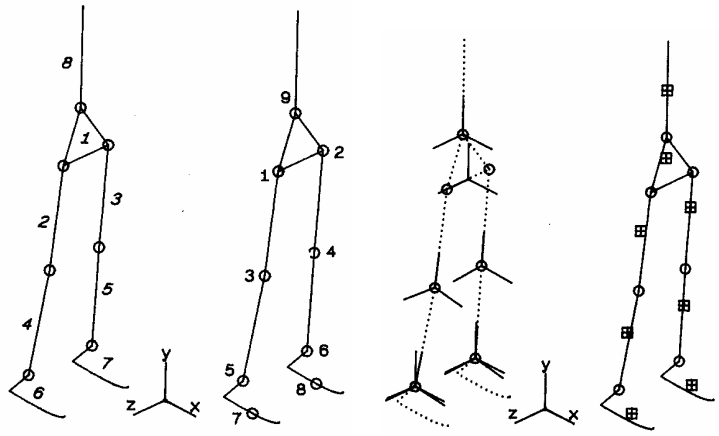


Figure 5. The segments model for the body at rest. From left to right: Numbering of segments and joints, definition of the reference frame, definition of local frames and positions of the centers of mass.

The segments model must include the following mechanical properties, which are dependent on individual characteristics: The dimensions and positions of the segments in space, the positions of the joints, and the mass properties of the segments (mass, position of the center of mass and moment of inertia tensor). Other properties, such as the ranges of motion of the joints, the maximal moments of force that can be exerted on the joints, and muscle models, may also be included [8].

2.3. Mathematical Motion Description

The segments model is the basis for a sound technical description of human movement. In general, a movement of a rigid body is defined as the displacement and orientation of the local frame relative to the reference frame or another local frame. Therefore, any description of a (body or joint) rotation implicitly assumes that some kind of segments model is defined, otherwise the rotation has no unique meaning. The anatomical position is usually taken as the offset position for the segments model where the relative rotations are defined as zero.

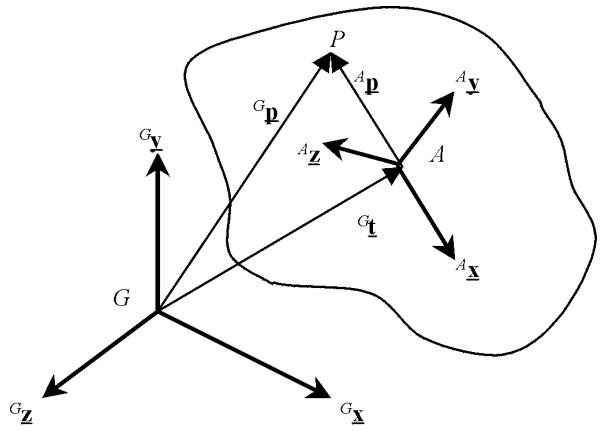


Figure 6. Transformation of position vector $^{A}\underline{\mathbf{p}}$ from local coordinate system A to position vector $^{G}\underline{\mathbf{p}}$ in the global coordinate system G.

Mathematically, the motion of a body with respect to another body is described unambiguously by a [3x3] *rotation matrix* **R** and a *translation vector* $\underline{\mathbf{t}}$ [5], see Figure 6. This results in a linear equation for an arbitrary position vector $\underline{\mathbf{p}}$. The rotation matrix can be viewed as some multi-dimensional function that, when it operates on the vector $\underline{\mathbf{p}}$, it changes the orientation of $\underline{\mathbf{p}}$. Note that $\underline{\mathbf{p}}$ has 3 components, and each component may change with respect to each of the 3 dimensions, so the rotation matrix represents 9 (or [3x3]) dependencies.

$$\begin{aligned} (^{G}\underline{\mathbf{p}} - {}^{G}\underline{\mathbf{t}}) &= {}^{GA}\mathbf{R} \cdot (^{A}\underline{\mathbf{p}} - {}^{A}\underline{\mathbf{t}}) \\ {}^{G}\underline{\mathbf{p}} &= {}^{G}\underline{\mathbf{t}} + {}^{GA}\mathbf{R} \cdot {}^{A}\underline{\mathbf{p}} \end{aligned} \tag{1}$$

Bold capitals are used to denote matrices; underlined lowercase characters are used to denote vectors. $^{G}\underline{\mathbf{p}}$ is the position vector of point *P* in the global coordinate system *G*, $^{A}\underline{\mathbf{p}}$ is the position vector of point *P* in the local coordinate system *A*. $^{G}\underline{\mathbf{t}}$ is the translation vector from the origin of *A* expressed in coordinate system *G*. Obviously $^{A}\underline{\mathbf{t}}$, the origin of *A* expressed in coordinates of *A* equals the zero vector $\underline{\mathbf{0}}$. $^{GA}\mathbf{R}$ is the rotation matrix that describes the rotation from the local coordinate system *A* to the coordinate system *G* (the global coordinate system).

Element-wise, equation (1) looks like

$$\begin{bmatrix} {}^{G}p_x \\ {}^{G}p_y \\ {}^{G}p_z \end{bmatrix} = \begin{bmatrix} {}^{G}t_x \\ {}^{G}t_y \\ {}^{G}t_z \end{bmatrix} + \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^{A}p_x \\ {}^{A}p_y \\ {}^{A}p_z \end{bmatrix} = \begin{bmatrix} {}^{G}t_x + r_{11}\,{}^{A}p_x + r_{12}\,{}^{A}p_y + r_{13}\,{}^{A}p_z \\ {}^{G}t_y + r_{21}\,{}^{A}p_x + r_{22}\,{}^{A}p_y + r_{23}\,{}^{A}p_z \\ {}^{G}t_z + r_{31}\,{}^{A}p_x + r_{32}\,{}^{A}p_y + r_{33}\,{}^{A}p_z \end{bmatrix} \tag{2}$$

in which the elements of the rotation matrix are cosines of the angles between the axes of the global and local coordinate system (see Figure 6):

$$
{}^{GA}\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} \cos(\theta_{11}) & \cos(\theta_{12}) & \cos(\theta_{13}) \\ \cos(\theta_{21}) & \cos(\theta_{22}) & \cos(\theta_{23}) \\ \cos(\theta_{31}) & \cos(\theta_{32}) & \cos(\theta_{33}) \end{bmatrix}
$$

$$
\theta_{11} = \angle({}^{A}\underline{\mathbf{x}}, {}^{G}\underline{\mathbf{x}}), \qquad (3)
$$

$$
\theta_{21} = \angle({}^{A}\underline{\mathbf{x}}, {}^{G}\underline{\mathbf{y}}), \quad \textit{etc.}
$$

Although the rotation matrix contains 9 elements, there exist dependencies between them because we assume rigidity of the segment it represents. Ultimately, there are only three *independent* variables present. Together with the three elements of the translation vector, these six variables describe the potential 6 DOF motion of the segment or between two bones. To better visualize motion and comparison between subjects, the rotation matrix is reduced to three successive rotations about well-defined axes, so-called *Euler angles*. This can be viewed as three hinge joints in succession. Rotation of the first hinge will change the orientation of the second and third, rotation of the second hinge will change the orientation of the third hinge, and rotation about the third hinge will result in the desired orientation of the local coordinate system.

In Biomechanics, usually *Cardan angles* are used, which are Euler angles with successive rotations about the local *x*-, *y*- and *z*-axes. Instead of the order *x*-*y*-*z,* any other combination of these rotations could have been chosen. Since matrix multiplication is not a commutative operation (a different order of the rotations, e.g. y-x-z instead of x-y-z, will give different outcomes), each combination of rotations will result in other values of the angles. Other parameterizations for the orientation may be used as well. For example, a well-known parameterization for rigid body movement is the Finite Helical Axis (FHA) definition, which utilizes *helical angles*.

3. Dynamics

The next step after defining the movement of rigid bodies is to derive the equations of motion that define the relations between the acting forces and torques and the movement. A straightforward way to do this is to apply the Newton-Euler approach. Although alternative formulations exist that have specific applications, such as the Lagrange approach, the Newton-Euler approach is simplest because:

The method is identical for each segment.

It is based on the Free Body Diagram of the segment. With this Free Body Diagram there is a simple recipe to obtain the equations of motion.

The equations of motion for the entire body is the sum of the equations for each segment.

Newton formulated the equations of motion for systems of mass particles. Euler recognized that a rigid body is a special case for such a system: The positions of the particles are constrained with respect to each other. This leads to the notion that the internal forces (the forces acting between the particles) do not perform work and do not contribute to the equations of motion for the entire system, the rigid body. Since a rigid body has six *degrees of freedom (DOF)*, there must be six equations describing the relation between forces and motion. This leads to the formulation of the Newton-Euler equations of motion for each segment:

$$\underline{\mathbf{F}}_{CM} = m\frac{d^2\underline{\mathbf{p}}_{CM}}{dt^2}$$
$$\underline{\mathbf{M}} = \frac{d(\mathbf{J}_{CM}\,\underline{\boldsymbol{\omega}})}{dt} \qquad (4)$$

Where $\underline{\mathbf{F}}_{CM}$ is the *resulting* external force acting on the center of mass (*CM*) of the rigid body with mass *m* and $\underline{\mathbf{p}}_{CM}$ is the position of *CM*. The first derivative with respect to time of $\underline{\mathbf{p}}_{CM}$ reflects the change of position in time, or the velocity of *CM.* Likewise, the second derivative of $\underline{\mathbf{p}}_{CM}$ (as in equation 4) reflects the change of velocity in time, or the acceleration of the center of mass. Any force, acting on the rigid body, can be divided in a force, acting on *CM*, and a moment of force $\underline{\mathbf{M}}$. The second vector equation in (4) resembles the first, as the *resulting* torque equals the product of rotation inertia tensor $\mathbf{J}_{CM}$ and angular acceleration (or the first time derivative of the angular velocity vector $\underline{\boldsymbol{\omega}}$ (omega). However, unlike the mass *m*, $\mathbf{J}_{CM}$ is in general not constant in each coordinate system. This leads to considerable complications in 3-D; in 2-D on the other hand $\mathbf{J}_{CM}$ reduces to a single constant component.

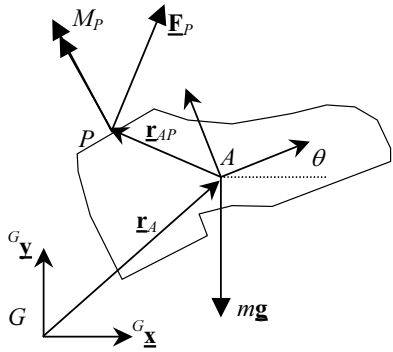


Figure 7. Free body diagram of a rigid body A with joint P.

Consider now the 2-D situation of a (foot) segment *A* with mass *m* and moment of inertia J_A during the swing phase of walking. A local frame is attached to the center of mass with an angle θ to the global (inertial) frame *G*. The segment connects to another segment in joint *P*, the ankle joint (Figure 7). Of course, the lower leg segment applies some connection forces on the foot in the ankle joint, expressed by the joint force $\underline{\mathbf{F}}_P$ and joint moment of force M_P. Finally, a gravity force applies to the foot, thus completing the free body diagram. When all vectors, which have both magnitude and direction, are expressed in global coordinates, equation (4) reduces to 3 scalar equations – scalars have magnitude only:

$$
\begin{aligned}
&{}^{G}F_{Px} = m \cdot \frac{d^2\,{}^{G}r_{Ax}}{dt^2} \\
&{}^{G}F_{Py} - m \cdot g = m \cdot \frac{d^2\,{}^{G}r_{Ay}}{dt^2} \\
&M_P + ({}^{G}r_{APx} \cdot {}^{G}F_{Py} - {}^{G}r_{APy} \cdot {}^{G}F_{Px}) = J_A \cdot \frac{d\omega}{dt}
\end{aligned}
\tag{5}
$$

Note that g is the acceleration of gravity constant. The expression between brackets in the moment of force equation is the vector product of force and moment arm. Note also that the vector ${}^{A}\underline{\mathbf{r}}_{AP}$ (expressed in the local frame A) is a constant but ${}^{G}\underline{\mathbf{r}}_{AP}$ (The same vector but now expressed in global frame G) depends on the angle θ (and thus on time) as defined by equation (3). The vector ${}^{G}\underline{\mathbf{r}}_{A}$ may be viewed as the translation vector in equation (1). With equations (1), (2) and (3), equation (5) may be expressed as three differential equations in the (time-dependent) degrees of freedom r_{Ax}, r_{Ay} and θ.

When the movement of the segment is known the second derivatives of the position and rotation coordinate may be calculated. Equation (5) may then be used to calculate the (unknown) joint forces and moments of force. Likewise, when additional known forces would be present, equations similar to (5) may be derived to compute the unknown forces at the ankle. For example, in an inverse analysis of the walking movement, the measured ground reaction forces are known. These would then contribute to the resulting forces and torques acting on the foot segment, and the equations of motion (4) are used to solve for the ankle force and torque. From these and the equations of motion of the shank segment, the knee forces are calculated, and so on.

4. Application to Human Gait

Bipedal walking is a complex movement, a balancing performance against the constant pull of gravity. Numerous definitions of walking are made, such as:

> *"In bipedal locomotion, man is continuously preventing a fall by placing one foot in front of the other."*
>
> *"The two basic requisites of walking are the continuing ground reaction forces that support the body and the periodic movement of each foot from one position of support to the next in the direction of progression"*

Apart from these formulations, walking can be quantified with a number of parameters. These are shown in the next sections: The step-parameters to describe the timing of the movement; kinematics for the movement itself (the joint rotations) and dynamics to describe the forces and moments of force that cause this movement.

The two ways to apply the equations of motion are usually referred to as the inverse dynamics and the direct (or forward) dynamics approach. In the inverse dynamics approach, the movement is assumed to be known and the forces and moments of force, needed to bring about that movement, are calculated (e.g. Koopman [9]). Inverse dynamics is applied in gait analysis, the equations of motion are usually derived with the Newton-Euler formulation. The estimated internal forces can be

further processed in muscle models to estimate the distribution of muscle forces, which allows for a validation with measured EMG patterns.

In the direct dynamics approach, the movements of the segments are calculated by integrating the equations of motion, usually based on a Lagrangian formulation. This is only possible when the joint moments of force are known or assumed to be zero. The latter is the case in ballistic walking [10]. The joint moments of force can be found by trial and error such that a normal walking pattern results [12], from estimations of the muscle forces [11], as the result of an inverse dynamics model or by optimization techniques. This direct dynamics modeling will not be discussed further.

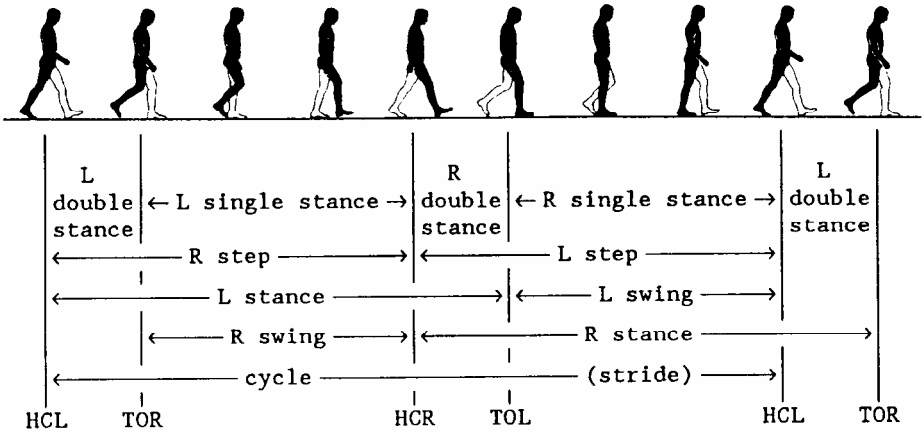


Figure 8. The walking cycle

4.1. Step-Parameters

The step-parameters are used for a general characterisation of the walking pattern. They are determined by the points of heel contact (HC) and toe-off (TO) of the left and right feet. For a walking cycle beginning with left heel contact (HCL), this is followed by right toe-off (TOR), right heel contact (HCR) and left toe-off (TOL). A cycle is completed with HCL again (Figure 8).

These points divide the walking cycle in four different phases. For the left leg, the stance phase is from HCL to TOL and the swing phase (or the single stance phase of the right leg) is from TOL to HCL. There are two double stance phases where both feet are on the floor. The left double stance phase is from HCL to TOR. One stride consists of two steps: The left step is usually defined from HCR to HCL.

For the step-parameters, a distinction is made between the time-parameters and the distance-parameters. With the cycle beginning at HCL (t_{HCL}=0), four other time-parameters will suffice to define the points of HC and TO. These may be normalized with the stride or cycle time T to make these time parameters dimensionless for comparison purposes. Likewise, the distance parameters can be made dimensionless with the stride length S. For symmetrical walking, the step-parameters are not independent any more.

Two step-parameters are derived from S and T, which are the average forward velocity v and the step ratio r. These are calculated from:

$$v = S/T \tag{6}$$

$$r = 1/4\, S \cdot T \tag{7}$$

The step ratio is usually defined as the step length divided by the step frequency. In contrast to this, definition (7) does not depend on differences between right and left leg for asymmetrical walking. The reason for introducing the step ratio is that this parameter is shown to be reasonably constant for a wide range of walking velocities and for different subjects [13]. For normal walking, r ranges from 0.39 to 0.44 m·s for men and from 0.34 to 0.40 m·s for women [7, 15]. For a constant r, S and T are determined by the forward velocity only:

$$T = 2\sqrt{r/v} \tag{8}$$

$$S = 2\sqrt{r \cdot v} \tag{9}$$

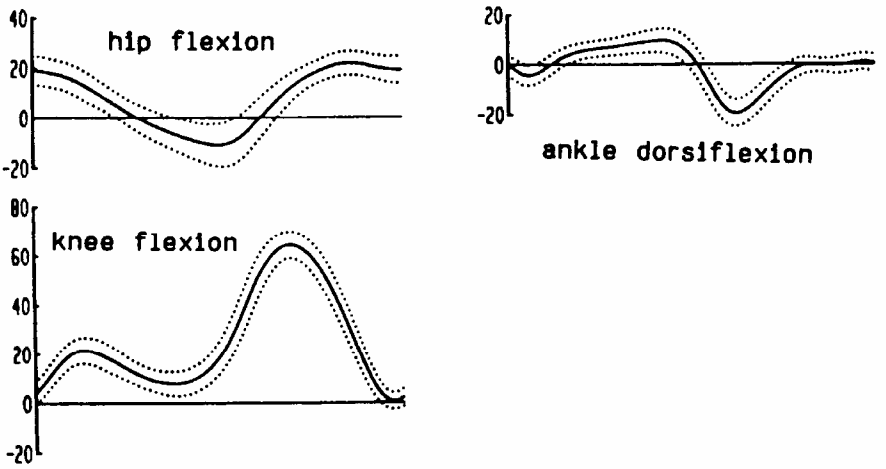


Figure 9. Joint rotations in the sagittal plane for a cycle beginning with heel contact (in degrees). Hip flexion, knee flexion and ankle dorsiflexion; average values for 19 subjects with measuring deviations. From (Winter, 1991).

4.2. Kinematics

A more detailed description of the walking movement is obtained when the joint rotations are measured at each point of time of the walking cycle. The hip, knee and ankle flexion are well documented in the literature, so measured data can be compared

with "standard" data. As an example, the rotations that are shown in Figure 9 are average values for normal walking, measured by Winter [16].

The other rotations that also contribute to the walking movement (e.g. hip adduction, pelvic rotations) are less well documented; accepted average values have yet to be established. The kinematics of the walking movement could also be described with the displacements of the joints as time functions or with a combination of displacements and rotations. The choice of one of these possibilities mostly depends on the measuring system that is available.

Note that a joint angle (e.g. hip angle) describes the orientation of one bone relative to another bone, or more precisely, the orientation of one local frame with respect to the other local frame of two connecting rigid bodies. A segment angle, on the other hand (e.g. the foot angle) describes the orientation of a local frame with respect to the global frame.

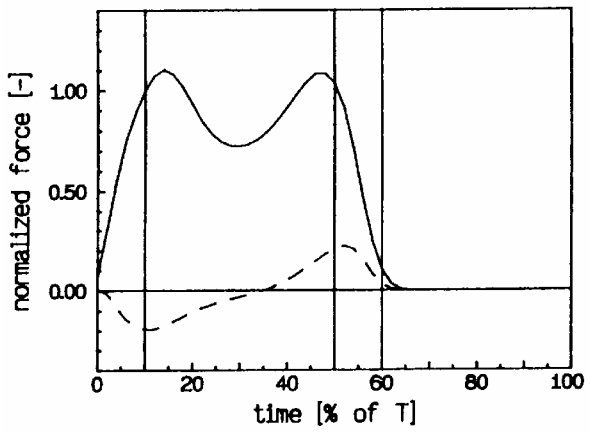


Figure 10. Forward (- - -) and vertical (—) ground reaction forces, normalized with body weight, ensemble average from 19 subjects. From Winter (1991).

4.3. Kinetics

An alternative description of the walking movement is provided with the internal and external forces and moments of force, acting on the body segments. Of these forces only the ground reaction forces can be measured directly. The average ground reaction forces, as measured by Winter [16], are shown in Figure 10.

Note that the average vertical ground reaction force should equal body weight. Variations from this average must result from accelerations of the total body centre of mass, according to equation (4).

The joint moments of force are determined by applying the Newton-Euler equations of motion (4). This always implies that some modeling assumptions, concerning the mechanical properties of the human body, have to be made. The joint moments of force are computed from the measured ground reaction forces and

accelerations. Where the variation in the ground reaction forces between different subjects is reasonably small, due to the assumptions there may be a larger variation in the joint torques. Figure 11 shows some of the joint moments of force, as determined for four different subjects [7]. The joint moments of force can be viewed as the net result of all muscular, ligament and frictional forces acting on the joint. For example, the push-off with the calf musculature is clearly visible in the ankle torque of Figure 11. For some applications, the muscular forces can be related to electromyographic (EMG) signals, which may validate the modeling assumptions.

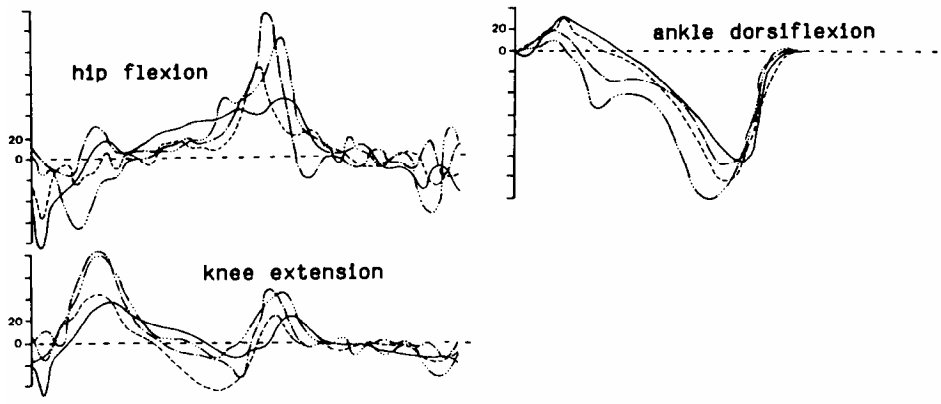


Figure 11. Joint moments of force for four different subjects (in Nm).
From Inman et al. (1981).

4.4. Energy Expenditure

A distinction is usually made between mechanical energy and metabolic energy expenditure. Mechanical energy can only be determined by modeling and is based on computed kinetic and potential energies of segments or on joint powers.

Metabolic energy expenditure is measured by oxygen uptake during walking and is usually expressed in energy per unit time (E_w) or energy per unit distance walked (E_m). Ralston [14] first showed that E_w is proportional to the square of the walking velocity:

$$E_w = a + bv^2 \tag{10}$$

This relation is confirmed by various investigators [7]. When E_w is expressed in watt per kg body mass and v is in m/s, the experimental values for the constants a and b are a = 2.24 W/kg and b = 1.26 Hz. E_m is defined by

$$E_m = \frac{E_w}{v} = \frac{a}{v} + b \cdot v \tag{11}$$

E_w and E_m are shown in Figure 12.

The optimal velocity v_{opt} is defined as the velocity where E_m is minimal. Differentiating equation (11) with respect to v and equating to zero yields v_{opt} = 1.33

m/s. The comfortable walking velocity v_{comf} is the velocity a person tends to adopt in a natural walk. v_{comf} is found to have an average value of 1.39 m/s [7], which differs from

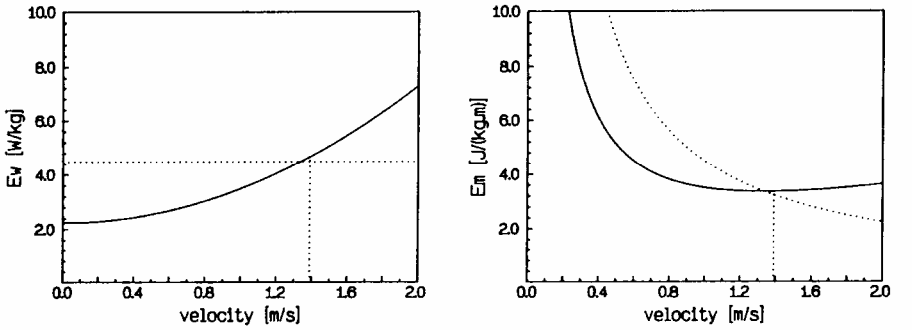


Figure 12. Metabolic energy expenditures E_w and E_m for normal walking at different velocities.

v_{opt} by only 4 %. Inman et al. called this an example of a fundamental feature of human motor behavior: *"In freely chosen rate of activity, a rate is chosen that represents minimal energy expenditure per unit task".* In the case of walking a speed is adopted, with a specific stride length S and cycle period T, such that each meter is covered as cheaply as possible. The energy expenditure E_m can in this sense be interpreted as inversely proportional to the efficiency of walking.

In Figure 12 is seen that E_m varies only little for a wide range of velocities: The sensitivity of the efficiency to the velocity is small in this range. When the velocity is enforced, for example in a treadmill, the stride length, cycle period and joint rotations can still be chosen freely. The choice is such that the efficiency of the movement is maximized.

4.5. Work Balance

By applying the equations of motion to the moving segments model in an inverse dynamics approach, the internal forces and moments of force are calculated. The product of the joint moment of force and the angular velocity equals the power output at the joint (Figure 13). A positive power reflects energy generation; a negative power reflects energy dissipation at the joint. On a muscular level, this is comparable to concentric (i.e. the muscle shortens while pulling) and eccentric (i.e. the muscle lengthens while pulling, force opposite to movement) muscle contractions respectively. These joint powers finally result in an increase (or decrease) of the mechanical energy (kinetic or potential) of the segments. However, since walking is a cyclic movement, the total power output of all the joints together in one cycle must equal zero for level walking. If for example the power output of one cycle would be positive, than the total mechanical energy of the segments model would be increased during one cycle. An increase of kinetic energy or walking speed is contradictory to the assumption of a cyclical movement; an increase of potential energy is only possible for walking uphill.

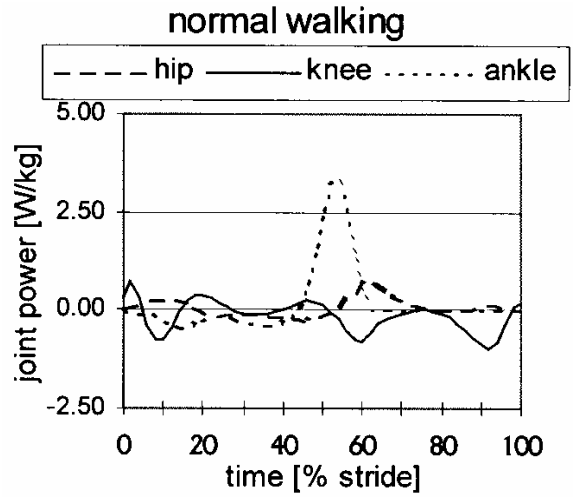


Figure 13. Joint powers (in W/kg). Average values for 19 subjects. Obtained from Winter (1991).

To be able to tell something about the work done at the joints, usually a distinction is made between the positive and the negative work (W_j^+ and W_j^-) at joint j with power P_j:

$$W_j^+ = \int_0^T P_j^+ dt; \qquad P_j^+ = \begin{cases} P_j & (P_j > 0) \\ 0 & (P_j \le 0) \end{cases}$$
$$W_j^- = \int_0^T P_j^- dt; \qquad P_j^- = \begin{cases} 0 & (P_j \ge 0) \\ P_j & (P_j < 0) \end{cases} \tag{12}$$

This definition results in a work balance as shown in table 1. From the table it is obvious which joints in general generate energy and which joints dissipate energy. In most cases, the hip joints generate as much as they dissipate, the knee joints are mainly dissipaters (in the beginning of the stance phase to accept body weight) and the ankle joints are mainly generators (at the end of the stance phase to push off, also visible in Figure 12). The total work for all joints during one cycle should equal zero. In table 1, the total $W_j^+ + W_j^-$ equals –2 J/cycle. This value reflects some inaccuracies of the calculation, especially the fact that the properties of the segments model do not perfectly match the properties of the human body.

The last column shows the absolute amount of mechanical work done. When it is assumed that mono-articular muscles do all the work, no energy is stored in passive structures, and the efficiency equals 1 for concentric contractions and –1 for eccentric contractions, this number can be related to the metabolic energy consumption. The metabolic energy consumption for a person of 80 kg with a stride length of 1.5 m/cycle is about 3.4×80×1.5=408 J/cycle (see Figure 11).

Table 1. Typical work balance for normal level walking.

joint	normal walking [J/cycle]			
j	W_j^+	W_j^-	$W_j^+ + W_j^-$	$W_j^+ - W_j^-$
R hip	18	-14	-4	32
L hip	18	-14	-4	32
R knee	17	-32	-15	49
L knee	17	-32	-15	49
R ankle	13	-3	10	16
L ankle	13	-3	10	16
total	96	-98	-2	194

About half this amount is needed for basal metabolism, the energy consumption at zero velocity. This leaves about 200 J/cycle spendet on walking alone. This value is close to the estimated mechanical work (table 1).

However, if more realistic values for the efficiencies are assumed (i.e. 0.3 and -1.0 for concentric and eccentric contractions respectively, see McMahon [10], the total mechanical work predicts a metabolic energy consumption of 418 J/cycle. This implies that energy transfer between joints through biarticular muscles and energy storage in passive structure are important mechanisms to reduce the over-all metabolic energy consumption. The work balances are especially useful when analyzing situations that deviate from normal.

References

[1] R.A. Brand, R.D. Crowninshield, C.E. Wittstock, D.R. Pedersen, C.R. Clark and F.M. van Krieken, A model of lower extremity muscular anatomy. *J. Biomech. Eng.* **104** (1982), 304-310.

[2] R.F. Chandler, C.E. Clauser, J.T. McConville, H.M. Reynolds and J.W. Young, *Investigation of the inertial properties of the human body.* Report DOT HS-801430, National Technical Information Service, Springfield Virginia 22151, U.S.A, 1975.

[3] C.E. Clauser, J.T. McConville and J.W. Young , *Weight, volume, and center of mass of segments of the human body*. Aerospace Medical Research Laboratory TR-69-70 (AD 710 622), Wright-Patterson Air Force base, Ohio, 1969.

[4] R. Fick, *Handbuch der Anatomie und Mechanik der Gelenke (Handbook of joint anatomy and mechanics)*, Gustav Fischer, Jena. 1911

[5] H. Goldstein, *Classical mechanics* (second edition). Addison Wesley Publishing company, Reading, Massachusetts. ISBN 0-201-02969-3, 1980.

[6] H. Hatze, Quantitative analysis, synthesis and optimization of human motion. *Hum. Movem. Sc.* **3** (1981), 5-25.

[7] V.T. Inman, Ralston, H.J., Todd, F., *Human walking*, Baltimore, Williams & Wilkins, 1981.

[8] B. Koopman, Grootenboer H.J., Jongh H.J. de, An inverse dynamic model for the analysis reconstruction and prediction of bipedal walking. *J. Biomech* **28** (1995), 1369-1376.

[9] H.F.J.M. Koopman, *The three-dimensional analysis and prediction of human walking*, Ph.D. Dissertation, University of Twente, Enschede, 1989.

[10] T.A. McMahon, *Muscles, reflexes, and locomotion.* Princeton University Press, Princeton, New Jersey, 1984.

[11] S.J. Olney and D.A. Winter, Predictions of knee and ankle moments of force in walking from EMG and kinematic data. *J. Biomech.* **18** (1985), 9-20.
[12] M.G. Pandy and N. Berme, A numerical method for simulating the dynamics of human walking. *J Biomech* **21** (1988), 1043-51.
[13] R.H. Rozendal, P.A.J.B.M. Huijing, Y.F. Heerkens, R.D. Woittiez, *Inleiding in de kinesiologie van de mens.* (Introduction in the human kinesiology) Culemborg: Educaboek, 1990.
[14] H.J. Ralston, Energy-speed relation and optimal speed during level walking. *Int. Zeitschrift für angewandte Pysiologie* **17** (1958), 277.
[15] R.L. Waters, B.R. Lunsford, J. Perry, R. Byrd, Energy-speed relationship of walking: standard tables. *J Orthop Res* **6** (1988), 215-22
[16] D.A. Winter, *The Biomechanics and motor control of human gait*, Waterloo (Canada), University of Waterloo Press, 1991.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-45

I.4. Biofluid Mechanics & the Circulatory System

Pascal VERDONCK[a] and Kris DUMONT[b]
[ab] *Institute Biomedical Technology, Ghent University, Belgium*

Abstract. A fluid is a medium which deforms, or undergoes motion, continuously under the action of a shearing stress and includes liquids and gases. Applying biofluid mechanics to the cardiovascular system requires knowledge of anatomy and geometry, pressure data and blood flow, volume and velocity measurements. A good example is the assessment of the haemodynamics of biological and mechanical heart valves.

Keywords. Pressure gradient, Bernoulli, Doppler measurement, Effective orifice area and performance index, heart valve, regurgitation, laminar flow, turbulent flow, Reynolds number

Introduction

The performance of a heart valve depends on the flow of blood passing through it. Cardiac output is the amount of blood pumped *by each ventricle* in one minute. It is a function of heart rate (HR), the number of heart beats per minute, and stroke volume (SV), the amount of blood pumped out by each ventricle per minute. The stroke volume (SV [ml]), times the number of beats per minute (heart rate, HR), equals the cardiac output (CO [l/min]; equation 1). Cardiac output in humans varies between 4-6 l/min at rest to 20-35 l/min during exercise, with heart rate variations from 40-70 bpm at rest to 170-200 bmp during exercise and SV from 60-125 ml during rest to 100-200 ml during exercise. A well trained person develops a larger SV and so CO can be achieved with a lower HR. [13].

$$CO = SV \cdot HR \tag{1}$$

Changes in either stroke volume or heart rate will alter cardiac output. The SV can be calculated from velocity measurements over the valve (equation 2):

$$SV = VTI \cdot A \tag{2}$$

with VTI, the velocity time integral (VTI [cm]) and A the valve area in [cm2]. A velocity-time curve is obtained using Doppler ultrasound and the VTI (= $\int v \cdot dt$) is the area under the velocity curve.

1. Pressure Gradient

The resistance of the valve against blood flow can be quantified by pressure/energy losses over the valve.

1.1. Pressure gradient from catheterization

A dynamic systolic gradient is found between left ventricular and aortic pressures (Figure 1a). Assuming proper alignment of the pressure tracings, a series of instantaneous pressure differences can be measured, and the mean systolic gradient ΔPmean can be calculated. The gradient ΔPpeak between the peak systolic left ventricular pressure and the systolic aortic pressure can also be determined, even though these two peaks are non-simultaneous. Note that the maximum instantaneous gradient ΔPmax is always larger than the peak-to-peak gradient. Similar gradients can be determined for the mitral valve during ventricular filling or diastole [16].

1.2. Pressure gradient from Doppler measurements

The pressure gradient can be calculated using the simplified Bernoulli equation. This equation is derived from the Bernoulli equation which is based on the assumptions that flow through the stenosis is laminar and inviscid (viscous forces negligible). Bernoulli's hydraulic formula states the conservation of mechanical energy between two points along a streamline (Figure 1b) [22]; equation 3:

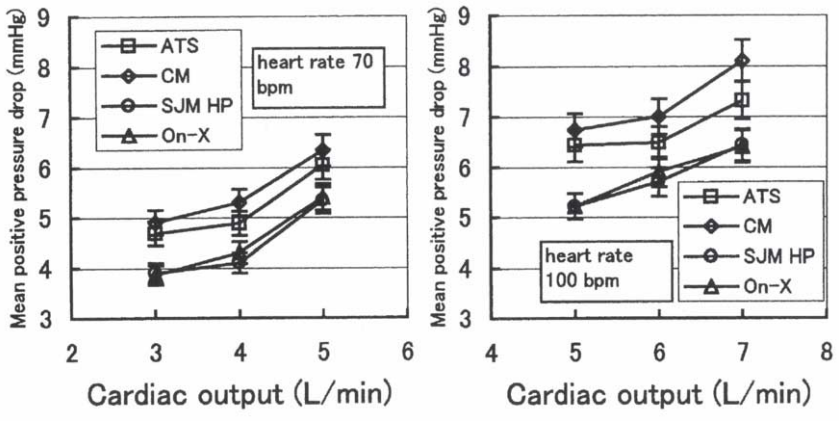
$$P_1 - P_2 = \frac{1}{2}\rho(v_2^2 - v_1^2) + \rho\, g(z_2 - z_1) + \rho \int_1^2 \frac{dv}{dt} ds \qquad \text{(3, Bernoulli's equation)}$$

where P_1 [Pa], v_1 [m/s], z_1 [m] are the pressure, velocity, and height from a reference level at the upstream location, and P_2 [Pa], v_2 [m/s], z_2 [m] are the pressure, velocity, and height at the downstream location. The integral term accounts for the flow acceleration or deceleration between the two locations. If the equation is applied at peak systolic velocity, this term becomes zero, because the rate of change of velocity, the derivative, becomes zero. When further applied on two points along a stream line located at the same height, one finds that [22]; equation 4:

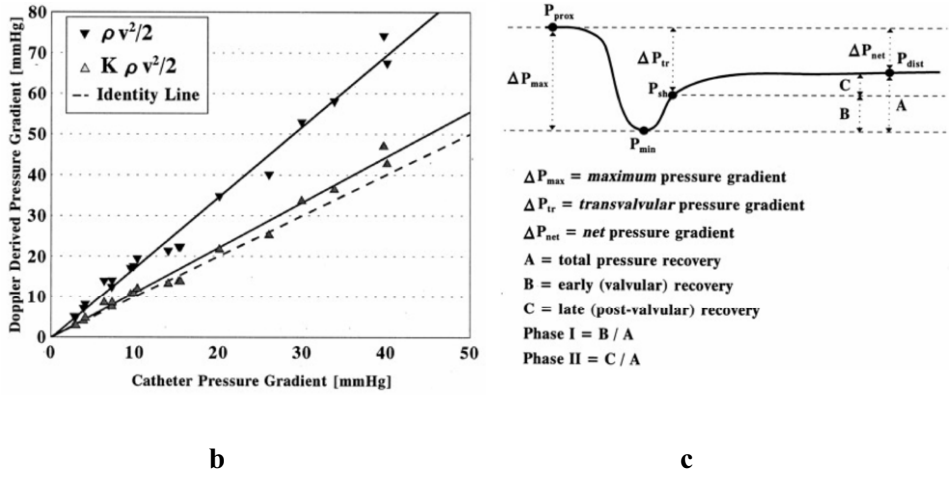
$$P_1 - P_2 = \frac{1}{2}\rho(v_2^2 - v_1^2) \qquad (4)$$

When the downstream velocity is much higher than the upstream velocity ($v_2 >> v_1$), v_1 can be neglected. Filling in the value for the density of blood at 37°C (1060 kg/m^3) ρ (rho), and converting pressure from Pa to mmHg (133 Pa = 1 mmHg), results in a factor of about 4, in the so-called simplified Bernoulli's equation [22]; equation 5:

$$\Delta P_{Doppler} \quad [mmHg] = 4 v_2^2 \qquad \text{(5, simplified Bernoulli's equation)}$$



a. Example of pressure gradient measurements [5].



b **c**

Figure 1. Pressure gradient.

where ΔP is in [mmHg] and v_2 is in [m/s] [22]. This equation has found wide application clinically in determining pressure drops across severe stenoses from noninvasive velocity measurements using Doppler ultrasound. Bernoulli's equation is not valid in cases in which viscous forces are significant, such as in long constrictions with a small diameter, or in flow with separation regions. In addition, it does not account for turbulent energy losses in cases of severe stenosis. Such losses should be taken into account, because they reduce the energy content of the

fluid [22]. In normal individuals, there is a very slight (1-2 mmHg) pressure difference between the left ventricle and aorta during ejection that helps drive the blood across the aortic valve. In contrast, very high pressure gradients are observed in patients with a stenotic valve. Patients with a pressure gradient ΔP of 100 mmHg, in the case of severe aortic stenosis, and of 16 mmHg in the case of mitral stenosis, are referred for clinical assessment..

2. Pressure Gradient in Experimental and Clinical Studies

Results from *in vitro* experiments by Feng et al [5] are displayed in Figure 1a. The SJM HP (St. Jude Medical Hemodynamic Plus) and On-X valves produced the lowest mean positive pressure drop, and the CM (CarboMedics) valve produced the highest under every condition. The differences in these pressure drops were related to the geometric orifice diameters and the degrees of valve opening. That the On-X valve produced the smallest pressure drop is mainly due to its larger internal orifice diameter and the parallel opening of its leaflets. The SJM HP valve also benefits from a larger internal orifice diameter and a large opening angle. The ATS valve has a lower mean positive pressure drop than the CM valve in pulsatile flow, despite the findings that its maximal opening angle is less than that of the CM valve. Furthermore, it can be observed from Figure 1a, that the pressure drop over the valves increases with the cardiac output. In the clinical setting, this maximum pressure gradient cannot be measured invasively because it is not possible to position a catheter across or between mechanical prosthetic valve leaflets in patients [18]. Doppler gradients across the central orifices are significantly higher than the transvalvular and net catheter pressure gradients measured across the valve (Figure 1b). These differences are due to downstream pressure recovery (Figure 1c) [6, 18], which gives the pressure at a certain time along the axis of the valve. The pressure shows a drop, but recovers at a certain distance downstream the valve, due to conversion of kinetic energy into potential energy upon deceleration of the flow through the valve.

3. Effective Orifice Area and Performance Index

When blood is flowing through an orifice with cross section area A [m^2], convergence of streamlines make that only part of the cross section is effectively used for flow passage. This is particularly important in valve dynamics where the effective valve area should be maximized to avoid high velocities, turbulence, shear, and associated high pressure drops. The effective orifice area (EOA) is the standard parameter for the clinical assessment of the severity of the valve stenosis. It is also a parameter that can be used to compare performance of different mechanical and biological valve prostheses.

As demonstrated in Figure 2, the flow through the valve is narrowing, using only the EOA, i.e. a fraction of the total geometric orifice area (GOA). The effective orifice area (EOA [cm^2]) is calculated - based on the continuity equation - as the ratio of forward stroke volume and VTI; equation 6:

$$EOA_{Continuity} = SV / VTI \quad (6)$$

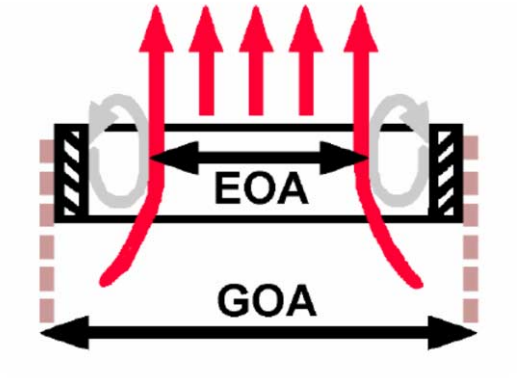


Figure 2. Geometric and effective orifice area (GOA and EOA).

Besides the continuity equation, the Gorlin equation provides an alternative way to calculate the effective orifice area (EOAGorlin). With Q the flow rate [m3/s] and v the maximal velocity [m/s] in the *vena contracta* (narrowed vessel), the EOA is given by

$$\text{EOA} = Q / [\, C_1 v \,] \tag{7}$$

where C_1 is a coefficient to adjust for the velocity profile in the *vena contracta*. Through the orifice, there is a conversion of potential energy to kinetic energy, and this conversion is described by [22]:

$$v = C_2 \sqrt{2 \cdot g \cdot \Delta h} \tag{8}$$

with the pressure head across the orifice $\Delta h = \Delta z + \Delta P/(\rho g)$ and where C_2 is a coefficient to adjust for loss in the conversion of energy from potential to kinetic energy. With $\Delta z = 0$ and assumption of $C_1 \cdot C_2 = 1$, the formula can be written as:

$$\text{EOA} = \frac{Q}{v} = \frac{Q}{\sqrt{2\dfrac{\Delta P}{\rho}}} = \frac{Q}{\sqrt{2\dfrac{\rho_{Hg} \cdot g \cdot \Delta h_{Hg}}{\rho}}} \tag{9}$$

The commonly used SI units [m^2], [m^3/s], [Pa] are converted into clinical used units [cm^2], [ml/s], [mmHg] for the area, the flow and pressure respectively.

$$\begin{aligned} &\text{EOA } [cm^2] = 104 \cdot \text{EOA } [m^2], \\ &Q \text{ [ml/s]} = 10^6 \cdot Q \text{ [m}^3\text{/s]}, \\ &\Delta P \text{ [mmHg]} = 10^3 \cdot \Delta h_{Hg} \text{ [mHg]} \end{aligned} \tag{10}$$

$\rho Hg = 13600$ [kg/m3]
$g = 9.81$ [m/s^2] (11)

Combining equations 9 and 10 results in:

$$EOA \cdot 10^{-4} = \frac{Q \cdot 10^{-6}}{\sqrt{2\frac{\rho_{Hg} \cdot g \cdot 10^{-3} \cdot \Delta P}{\rho}}} \quad (12)$$

And thus finally the Gorlin equation can be written as:

$$EOA_{Gorlin} = \frac{Q}{51{,}6\sqrt{\Delta P}} \quad (13, \text{Gorlin equation})$$

with EOA in [cm^2], ΔP in [mmHg] and the mean forward flow $Q_{fwd,mean}$ in [ml/s], calculated as the ratio of the "positive" area under the flow curve (forward stroke volume SV) divided by the duration of forward flow. EOA_{Gorlin} is a fraction θ (theta), of the available geometric orifice area GOA, so that $EOA_{Gorlin} = \theta \cdot GOA$. The performance index (PI [no dimension]) of the valve is calculated as the ratio of the effective orifice area and the geometric orifice area (GOA):

$$PI = EOA / GOA \quad (14)$$

3.1. Effective Orifice Area

Many clinical and experimental studies are done in order to estimate the EOA and calculate the performance index PI which enables clinicians and engineers to compare different types of heart valves. The higher the PI, the better the hydrodynamic performance is of the valve. This means the contraction of flow through the valve is minimized and the flow utilizes a high percentage of the available geometric orifice area. Bech-Hanssen et al. showed similar effective orifice areas *in vitro* for St Jude Medical and Omnicarbon valves (Figure 3) [1].

4. Regurgitation

Regurgitation is reversed flow through the 'one way' valve. The percentage of regurgitation (% reg) in relation to the total SV is calculated as

$$\% \text{ reg} = V_{reg} / (V_{reg} + SV) \quad (15)$$

Vreg [ml] is the total volume of regurgitation and can be split into V_{close} and V_{leak}. V_{close} is the volume of regurgitation due to the closing of the valve. V_{leak} is the volume of regurgitation due to leakage of the closed valve (Figure 4).

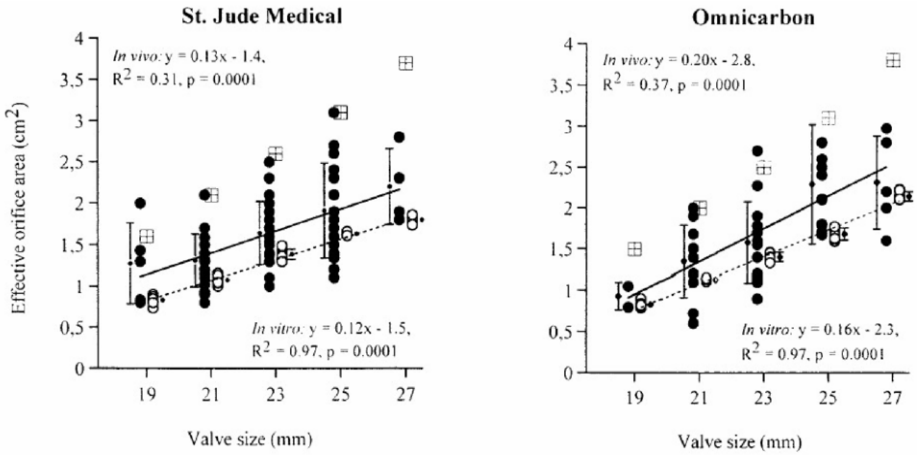
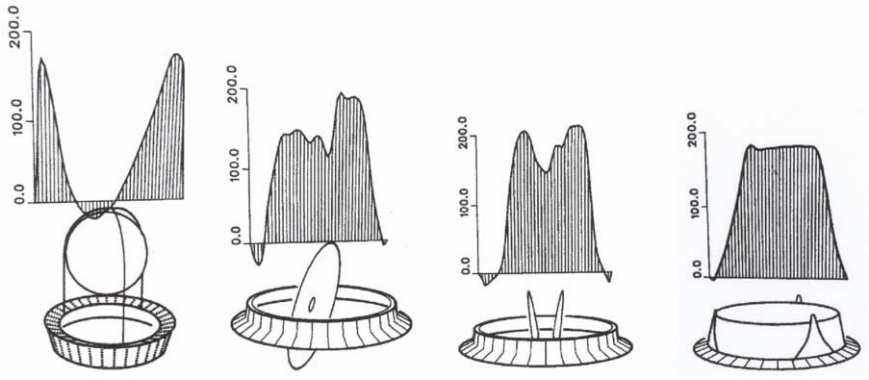


Figure 3. Effective orifice area of two different mechanical valve types [1]. Results are shown for the *in vivo* (black o) and *in vitro* (open o) studies. Bars indicate mean ± SD. Boxes show the geometric orifice area.



(a) Caged ball valve. (b) Tilting disc valve. (c) Bileaflet valve. (d) Bioprosthesis.

Figure 4. Velocity profiles in cm/s based on *in vitro* measurements on 27 mm aortic valve designs. Measurements were done downstream the valve on a centerline at peak systole, at a cardiac output of 6 l/min and a heart rate of 70 beats/min [21].

4.1. Regurgitation

Figure 5 shows regurgitation data from different types of bileaflet valves [5]. As reported by Wu et al. [20], the closing volume, expressed as a percentage of the forward flow volume, increased with decreasing cardiac output. The closing volumes

of the valves, shown in Figure 5, are within the acceptable range (<8%) [5]. The SJM HP valve under low cardiac output condition at 100 bpm showed a somewhat higher closing volume of 11.4%. Closing volume is believed to be proportional to opening angle of the mechanical heart valve [5].

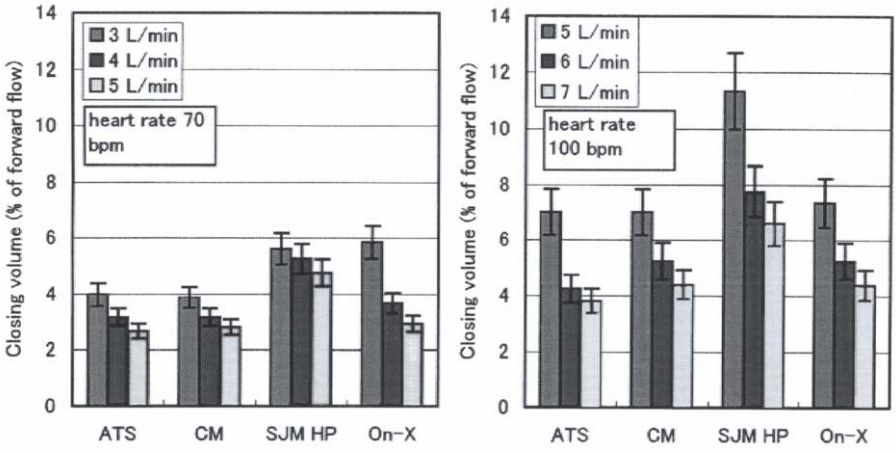


Figure 5. Closing volumes of different bileaflet heart valves (ATS: Advancing The Standard, CM: Carbomedics, SJM HP: St. Jude Medical Hemodynamic Plus, On-X) studied in mitral position in an experimental setup at 70 beats/min and at 100 beats/min [5].

5. Flow Patterns and Shear Stresses

5.1. Laminar and Turbulent Flow

The Reynolds number (Re) is a quantity that engineers use to assess whether fluid flow is laminar or turbulent. This is important, because increased mixing and shearing occur in turbulent flow. This results in increased energy losses which affects the efficiency of heart valves. A good example of laminar and turbulent flow is the rising smoke from a cigarette. The smoke initially travels in smooth, straight lines (laminar flow) then starts to ”wave” back and forth (transition flow) and finally seems to randomly mix (turbulent flow). The dimensionless Reynolds number Re [-] is defined as:

$$\mathrm{Re} = \rho \frac{U \cdot D}{\mu} \tag{16}$$

with ρ (rho) the density of the fluid (kg/m^3), U the velocity of the flow (m/s), D the diameter of the vessel (m) and μ (mu) the dynamic viscosity, or internal friction, of the fluid (mPa·s). Figure 6 demonstrates the different flow regimes in function of Re for the flow in a straight tube. Generally, a fluid flow is laminar in a stiff tube from Re = 0

to some critical value (Re < 2000) at which transition flow begins. Transition flow is fluctuating.

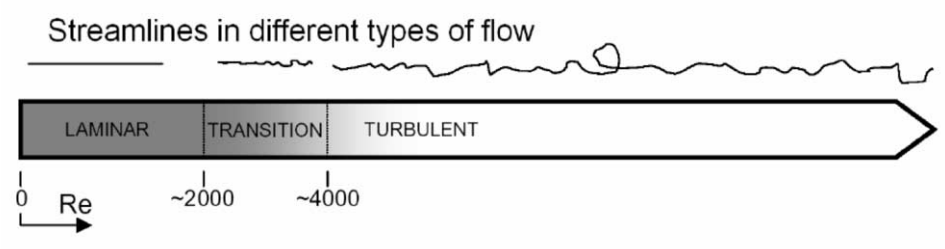


Figure 6. Diagram of flow regimes in pipe flow. between laminar and turbulent flow (2000 < Re < 4000).

Fluid flow becomes unstable for higher Reynolds numbers (Re > 4000). In turbulent flow, there is increased mixing that results in viscous losses, which are generally much higher than those in laminar flow [19]. The Reynolds number can reach up to 4500 at peak flow in the normal aortic valve. The Strouhal number gives an indication of a steady or transient regime of the flow. The Strouhal number is defined as

$$\mathrm{Sr} = \frac{D}{T_p \cdot U} \tag{17}$$

with T_p the period of time of the observed flow phenomenon. A very small Strouhal number (Sr << 1) is considered quasi-steady. For a Strouhal number close to one, the flow is considered transient.

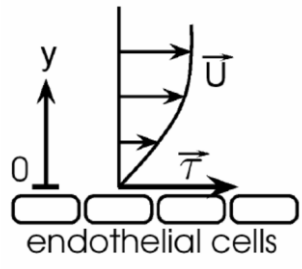


Figure 7. Shear stress represents the frictional force exerted by the flowing blood on the endothelial surface of the wall.

5.2. Shear Stresses

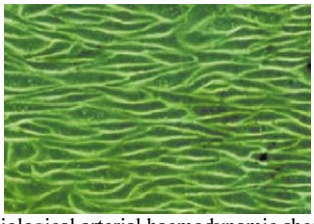
As blood flows through a vessel or valve, it exerts a physical force on the vessel wall or the valve leaflet. This force can be resolved into two principal vectors (Figure 7). Shear stress, being tangential to the wall, represents the frictional force exerted by the flowing blood on the endothelial surface of the wall. The shear stress on the vessel wall or on the valve leaflet is normally called wall shear stress or WSS. Normal stress, or pressure, is perpendicular to the wall. In the case of laminar flow shear stress is calculated with equation 18.

$$\vec{\tau}_{laminar} = \mu \frac{\partial \vec{U}}{\partial y} = \mu \cdot \dot{\gamma} \qquad \text{[Pa] or [N/m2]} \tag{18}$$

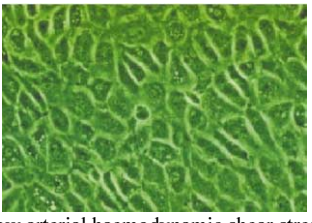
with τ (tau), the shear stress, μ (mu) the dynamic viscosity of the fluid, *U* the velocity vector and $\dot{\gamma}$ the shear rate. Turbulence in the blood system increases resistance to flow, resulting in higher pressure gradients. Turbulent shear stresses are calculated in a slightly different manner, as shown in equation 19 [3]:

$$\vec{\tau}_{turbulent} = \mu \frac{\partial \vec{U}}{\partial y} - \rho \overline{u'v'} \tag{19}$$

with *u'* and *v'* the turbulent fluctuations of the velocities *u* and *v* respectively. The vascular and valvular endothelial cells are subjected at all times to shear forces that act on the vessel surface as a result of the flow of viscous blood. Fluid shear stress regulates endothelial phenotype by altering its gene expression profile, including growth factors [11,12]. Fluid shear stress transforms polygonal, cobblestoneshaped endothelial cells of random orientation (Figure 8a) into fusiform endothelial cells aligned in the direction of flow (Figure 8b). High shear stress (> 40 N/m^2) can cause direct endothelial injury [11]. Low and oscillating shear stress regions (< 0.4 N/m^2) can lead to atherosclerotic lesions or plaques, mainly at arterial bifurcations and in the coronary arteries [11].



a. Physiological arterial haemodynamic shear stress (τ > 1.5 N/m^2) [1].



b. Low arterial haemodynamic shear stress (τ ~ 0-0.4 N/m^2) [1].

Figure 8. Transformation of endothelial cell morphology by fluid shear stress: aortic endothelial cells exposed to low shear stress are randomly oriented (a), while those exposed to physiological shear stress (1.5 N/m^2, right panel) for 24 hours align in the direction of blood flow [11,12]

High shear stresses in the blood may create platelet activation [10,15] leading to thrombosis [23], and the subsequent risk of embolism. Blood platelet damage starts to occur at shear stress values of 10 N/m2 [8]. Furthermore, the magnitude, exposure time, and spatial distribution of the shear stresses coincide with damage to red blood cells. Shear stresses above 200 N/m2 will cause hemolysis [2,4,14]. Hemolysis can also occur at lower shear stress values if there is a long exposure time [3,7]. Therefore recirculation zones in prosthetic heart valves are to be avoided, unless they are washed out by the next heart beat [9], so that the residence time of particles remains limited. *In vitro* studies of aortic prosthetic heart valve designs have revealed shear stress values sufficiently high to cause lethal or sublethal damage to blood cells [17,24]. Shear stress cannot be measured directly *in vitro* nor *in vivo* on a moving heart valve leaflet. For that reason computational fluid dynamics is used to estimate the shear stress.

6. Flow Patterns and Shear Stresses

6.1. The Caged Ball Valve

Natural heart valves allow blood to flow straight through the center of the valve. This property, known as central flow, minimises the work done by the heart. With non-central flow, the heart works harder to compensate for momentum lost to change in direction of the fluid. Caged ball valves completely block central flow (Figure 9a); blood therefore requires more energy to flow around the occluder. In addition, collisions with the occluder ball caused damage to blood cells. Caged-ball valves are also notorious for stimulating thrombosis, requiring patients to take lifelong prescriptions of anticoagulants.

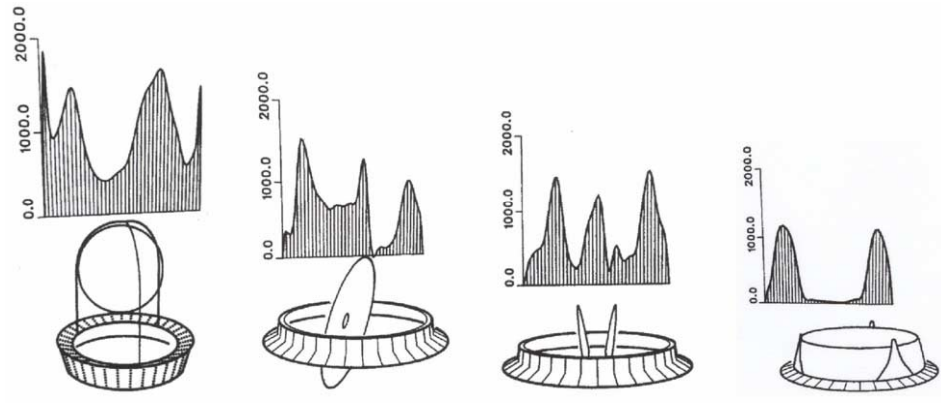
High turbulent shear stresses are observed at the edges of the jet (Fig 9a). The intensity of turbulence during peak systole does not decay very rapidly downstream of the valve. Elevated shear stresses occurred during most of systole [21]. This type of valve is no longer used clinically.

6.2. The Monoleaflet Tilting Disc Valve

This valve design allows more central flow while still preventing backflow. However, flow is not completely central (Figure 9b), because blood has to make a significant directional change to flow around the occluder disc. Newer models of monoleaflet valves improve this shortcoming. In the major orifice region (Figure 9b), high turbulent shear stresses are confined to narrow regions at the edges of the major orifice jet. During the acceleration and deceleration phases, the turbulent shear stresses are relatively low [21]. High turbulent shear stresses are more dispersed in the minor orifice than those in the major orifice regions (Figure 9b)

6.3. The Bileaflet Valve

These valves provide the closest approximation to central flow yet achieved in a mechanical heart valve (Figure 9c). The two leaflets block flow to some extent, leading to a three-part flow pattern. High turbulent shear stresses occur at locations immediately distal to the valve leaflets (Figure 9c).



(a) Caged ball valve. (b) Tilting disc valve. (c) Bileaflet valve. (d) Bioprosthesis.

Figure 9. Turbulent shear stress profiles in 0.1·N/m2 based on *in vitro* measurements on 27 mm aortic valve designs. Measurements were taken downstream of the valve on a centerline at peak systole, at a cardiac output of 6 l/min and a heart rate of 70 beats/min [21].

6.4. Biological Valves

Biological valves, optimally mimic native heart valves and provide central flow (Figure 9d). Turbulent shear stress measurements during the deceleration phase show low shear stresses spread out over a wide region [21]. During peak systole, the high turbulent shear stresses were confined in a narrow region (Figure 9d).

7. Conclusions

This chapter illustrates the importance of biofluid mechanics in understanding hemodynamics in general and the flow across heart valves in particular. Also the impact of biomechanical loading, both normal and shear stress, on heart valve flow is discussed.

References

[1] O. Bech-Hanssen, K. Caidahl, I. Wallentin, P. Ask and B. Wranne, Assessment of effective orifice area of prosthetic aortic valves with doppler echocardiography: An in vivo and in vitro study, *The Journal of Thoracic and Cardiovascular Surgery* **122**(2), (2001) 287–295.

[2] P. Blackshear, F. Dorman and E. Steinbach, Shear, wall interaction and hemolysis, *Transactions of the American Society of Artificial Internal Organs* **12**, (1966) 113–120.

[3] D. Bluestein, Y. Li and I. Krukenkamp, Free emboli formation in the wake of bi-leaflet mechanical heart valves and the effects of implantation techniques, *Journal of Biomechanics* **35**(12), (2002), 1533–1540.

[4] J.T. Ellis, T.M. Wick and A.P. Yoganathan, Prosthesis-induced hemolysis: mechanisms and quantification of shear stress, *Journal of Heart Valve Disease* **7**(4), (1998) 376–386.

[5] Z. Feng, T. Nakamura, T. Fujimoto and M. Umezu, In vitro investigation of opening behavior and hydrodynamics of bileaflet valves in the mitral position, *Artificial Organs* **26** (2002), 32–39.
[6] D. Garcia, J. Dumesnil, L.G. Durand, L. Kadem and P. Pibarot, Discrepancies between catheter and doppler estimates of valve effective orifice area can be predicted from the pressure recovery phenomenon: Practical implications with regard to quantification of aortic stenosis severity, *Journal of the American College of Cardiology (JACC)* **41** (2003), 435–442.
[7] L. Goubergrits and K. Affeld, Numerical estimation of blood damage in artificial organs, *Artificial Organs* **28**(5), (2004) 449–507.
[8] J. Hellums and R. Hardwick, *The Rheology of Blood, Blood Vessels and Associated Tissues, chapter Response of Platelets to Shear Stress - a Review*, 160–183, Sijthoff & Noordhoff, Alphen aan den Rijn, 1981.
[9] S. Kelly, P. Verdonck, J. Vierendeels, K. Riemslagh, E. Dick and G.V. Nooten, A threedimensional analysis of flow in the pivot regions of an ATS bileaflet valve, *International Journal of Artificial Organs* **22**(11), (1999) 754–763.
[10] M. Kroll, J. Hellums, L. McIntire, A. Schafer and J. Moake, (1996) Platelets and shear stress, *Blood* **88**(5), (1996) 1525–1541.
[11] A.M. Malek, S.L. Alper and S. Izumo, Hemodynamic shear stress and its role in atherosclerosis, *Journal of the American Medical Association* (JAMA) **282**(21), (1999) 2035–2042.
[12] A. Malek, and S. Izumo, Mechanism of endothelial cell shape change and cytoskeletal remodeling in response to fluid shear stress, *Journal of Cell Science* **109**, (1996) 713–726.
[13] K. Matthys, *Assessment of Vascular Hemodynamics: Investigation of non-invasive and minimally invasive methods for assessment of vascular function at rest and during cardiovascular challenge*, Ph.D. Dissertation, Ghent University, 2004.
[14] A. Nevaril, E. Lynch, C. Alfrey and J. Hellums, Erythrocyte damage and destruction induced by shearing stress, *Journal of Laboratory and Clinical Medicine* **71**(5), (1968) 781–790.
[15] Z. Ruggeri, Mechanisms of shear-induced platelet adhesion and aggregation., *Thrombosis and Haemostasis* **70** (1), (1993) 119-123.
[16] K. Schmailz and O. Ormerod, ed., Ultrasound in Cardiology, Blackwell Science, 1998.
[17] W. Tillmann, H. Reul, M. Herold, K. Bruss and J. van Gilse, In vitro wall shear measurements in aortic valve prostheses, *Journal of Biomechanics* **17**(4), (1984) 263–279.
[18] P.M. Vandervoort, N.L. Greenberg, M. Pu, K.A. Powell, D.M. Cosgrove and J.D. Thomas, Pressure recovery in bileaflet heart valve prostheses: localized high velocities and gradients in central and side orifices with implications for doppler-catheter gradient relation in aortic and mitral position, *Circulation* **92** (1995), 3464–3472.
[19] J. Welty, C. Wicks, R. Wilson and G. Rorrer, *Fundamentals of momentum, heat, and mass transfer, Wiley Text Books*; 4 edition, ISBN 0471381497, 2000.
[20] Z. Wu, B. Gao, J. Slonin. and N. Huang, N., Bileaflet mechanical heart valves at low cardiac output, *ASAIO Journal* **42**(5), (1996) 747–749.
[21] A.P. Yoganathan, *The Biomedical Engineering Handbook, Volume I: Second Edition., chapter Cardiac Valve Prostheses*, CRC Press LLC, ISBN 0-849-38594-6, 2000.
[22] A.P. Yoganathan and G. Chatzimavroudis, *PanVascular Medicine: Integrated Clinical Management, Chapter 7: Hemodynamics*, Springer, 2002.
[23] A.P. Yoganathan, T. Wick and H. Reul, H., *Current issues in heart valve disease, chapter Thrombosis, Embolism and Bleeding*, London: ICR, 1992.
[24] A.P. Yoganathan, Y.R. Woo and H.W. Sung, Turbulent shear stress measurements in aortic valve prostheses, *Journal of Biomechanics* **19**(6), (1986) 433–442.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-58

I.5. Biomechanics of Implants

Jan G. HAZENBERG [a], Johannes SCHMID [b], T. Clive LEE [c] and Gijsbertus J. VERKERKE [d,e]

[a] *IVAX Pharmaceuticals, Waterford Industrial Estate, Waterford, Ireland*
[b] *University of Applied Sciences, dept of Mechanical Engineering, Regensburg, Germany*
[c] *Royal College of Surgeons in Ireland, Department of Anatomy, Dublin, Ireland*
[d] *University Medical Center Groningen, University of Groningen, Dept of Biomedical Engineering, Groningen, the Netherlands*
[e] *University of Twente, Dept of Biomechanical Engineering, Enschede, the Netherlands*

Abstract. For simple constructions a mechanical analysis to determine internal stresses and deformation is possible using theoretical formulas. However, for complex constructions, like joint prostheses, this is not possible. Numerical simulation of internal stresses and deformations offers a solution for these constructions. The so-called Finite Element Analysis divides the complex structure in simple ones (elements), applies the mechanical formulas and adds the effect on each element to predict the behaviour of the complex contruction.

Keywords. Finite Element Analysis, mechanics, differential equation, node, stiffness matrix

Introduction

Human life expectancy is increasing and so too, therefore, is the need for joint replacements to maintain mobility. However, the functional life expectancy of the implants is limited to 10-20 years, thus in many patients the original implant must be removed and replaced. One of the reasons for the limited lifetime of these devices is the fact that bone adapts its shape to a change in load distribution as the geometry of an artificial joint differs from the original, particularly in its stem.. As the bone adapst its shape, the fixation of implants is affected resulting in movement of the artificial joint, malfunction and pain, making revision surgery necessary. Such revision surgery is difficult and dangerous, so an increase in the lifetime of joint prostheses is preferable.

To improve those implants, an analysis of the loading situation and response of the implant is required However, the geometry of such an implant and the surrounding bone is complicated, and bone is a complex material, with non-homogenous properties. The Finite Element Analysis (FEA), which originates from the aerospace and nuclear industries, offers a solution, Finite element packages avail of advanced imaging

techniques and are capable of solving the most sophisticated problems, not just in structural analysis, but for a wide range of phenomena such as static and dynamic loading of orthopaedic implants and fluid flow problems in cardiovascular applications. Its popularity is due to a number of factors, which include:

- A variety of loading conditions and environmental parameters can be analysed
- A variety of materials can be analysed and tested under *in vivo* circumstances
- Geometrical modifications are easily made
- It reduces the number of animal and pre-clinical tests required
- It is cost and time efficient

1. How does FEA work in practice?

1.1. Geometry

It typically starts with the generation of a geometrical model. Currently a variety of techniques is available which include µCT and MRI scanning. These machines are able to provide data files, which contain three-dimensional data points of the geometry, which can be a femur, mandible or artery. These data point are than linked together, resulting in a wire frame which represents the geometry. Using these lines, areas can be created from which a solid 3D model is generated. In Figure 1A, a wire frame model is shown of half a mandible, in which a plane fracture (blue arrow) and a mini plate for reconstruction (red arrow) are modelled. The angle of the fracture can be modified in the solid model (Figure 1B).

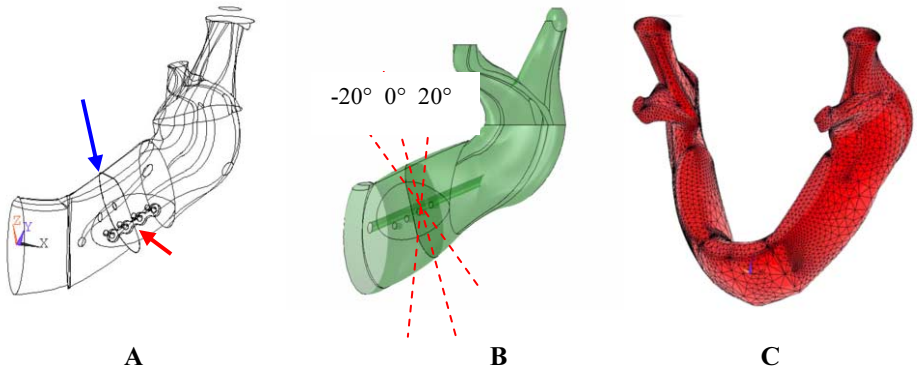


Figure 1. From the MRI or µCT scan a model is created of a mandible (A). The data points are connected through lines and a solid model is created (B). Finally the solid model is meshed (C).

To carry out a finite element analysis, the model must be divided into a number of small pieces known as finite elements. In simple terms, a mathematical net or "mesh" is required to carry out a finite element analysis. If the system is one dimensional, it can be represented by line elements, if it is two dimensional, then a 2D mesh is required.

Correspondingly, if the problem is complex and a 3D representation of the continuum is required, then we use a 3D mesh. Figure 1C shows the 3D-meshed model of a mandible.

1.2. Material Properties

The next step is to give the finite elements mechanical properties such as Young's modulus, which determines its elasticity. The final step is to place a load (or different loads) at one or more nodes, or intersection points of the mesh, and to fix other nodes to the environment.

Now the load on all elements will be computed and can be compared to the maximal allowable load before failure. If this load is exceeded in some element, we know that the structure will collapse.

If the maximal load is not exceeded, the resulting deformation on every element can be calculated and, by adding all deformations, the total deformation of the complete structure is found..

In more detail, the finite element is a mathematical method for solving ordinary and partial differential equations. As these types of equations occur naturally in virtually all fields of the physical sciences, the applications of the finite element method are limitless as regards the solution of practical design problems. Within each element, the variation of displacement is assumed to be determined by simple olynomial shape functions and nodal displacements. Equations for the strains and stresses are developed in terms of the unknown nodal displacements. From this, the equations of equilibrium are assembled in a matrix form, which can easily be programmed and solved on a computer. After applying the appropriate boundary conditions, the nodal displacements are found by solving the matrix stiffness equations. Once the nodal displacements are known, element stresses and strains can be calculated. From the elements and their nodes the 'stiffness' matrix is formulated. This square matrix contains details of the material properties, the model geometry and any assumptions of the stress-strain field.

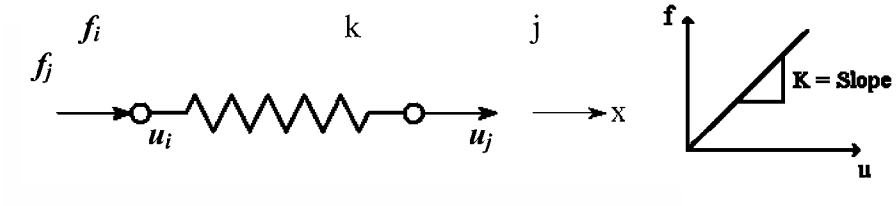


Figure 2. Single line element.

A mesh is essentially an assembly of springs. For a spring in general we can say that the force F is proportional to the spring constant k times the displacement u, or $F = k \cdot u$. Let us consider a single line element, consisting of two nodes, 'i' and 'j',

Figure 2. Two these nodes we can apply forces, f_i and f_j respectively, resulting in a displacement u_i and u_j. The displacement of these two nodes will be dependant on the stiffness of the spring 'k'. In any case where linear elastic material properties can be assumed, the displacement u increases proportionally with the force f (if k is constant).

In order for this spring to be in a state of equilibrium, the forces f_i and f_j have to be of equal magnitude in opposite directions. Mathematically we can say that:

$$f_i = -F = -k(u_j - u_i) = ku_i - ku_j \text{, for node } i, \tag{1}$$

$$f_j = -F = -k(u_j - u_i) = -ku_i + ku_j \text{ at node } j. \tag{2}$$

This can also be expressed in matrix form, as is usual in FEA, as follows:

$$\begin{Bmatrix} f_i \\ f_j \end{Bmatrix} = \begin{bmatrix} k & -k \\ -k & k \end{bmatrix} \bullet \begin{Bmatrix} u_i \\ u_j \end{Bmatrix} \tag{3}$$

This is similar to $F = K \cdot u$ in this case. Here F is the matrix of the load vectors, derived from the loads of every cases, u is the matrix of the displacement of nodes and K is the coefficient matrix of the system, the so called stiffness matrix. Since FEA does not use one single element but several, we can do this in the same way for a spring assembly.

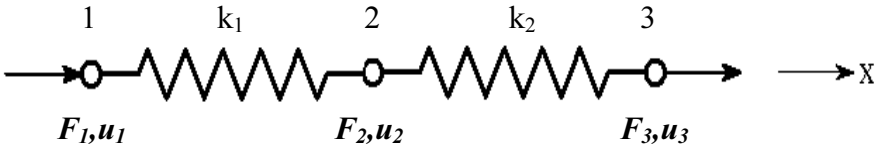


Figure 3. Two line elements.

In Figure 3 we have two elements; one is connected by nodes 1 and 2, while the second element is connected through the elements 2 and 3. This can be expressed in matrix form as follows:

$$\begin{Bmatrix} f_1^1 \\ f_2^1 \end{Bmatrix} = \begin{bmatrix} k_1 & -k_1 \\ -k_1 & k_1 \end{bmatrix} \bullet \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} \text{ for element 1 and} \tag{4}$$

$$\begin{Bmatrix} f_1^2 \\ f_2^2 \end{Bmatrix} = \begin{bmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} \bullet \begin{Bmatrix} u_2 \\ u_3 \end{Bmatrix} \text{ for element 2.} \tag{5}$$

The equilibrium forces at node 1 are therefore $F_1 = f_1^1$, at node 2, $F_2 = f_2^1 + f_1^2$ and at node 3, $F_3 = f_2^2$. These matrices can now be combined to form the stiffness matrix for these to elements, which look like:

$$\begin{Bmatrix} F_1 \\ F_2 \\ F_3 \end{Bmatrix} = \begin{bmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{bmatrix} \bullet \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix} \tag{6}$$

In this two-dimensional example we have relatively little to calculate. Due to increased computational power, more sophisticated three-dimensional models can be produced, which can consist up to 100,000 elements and require several hours to solve a particular problem.

Once the stiffness matrix is created, external loads are applied to evaluate the displacements of the structure (hence the term displacement analysis). These loads can be static, dynamic or represent impact loading, depending on the location and function. Initially, the construction of a 'free-body-diagram' in which all muscle forces, joint load and reaction forces are drawn is useful. In cases where dynamic loads are to be investigated a 'kinetic diagram', which is similar to the 'free-body-diagram', can be made in which the expected accelerations can be mapped. This information can than be inserted into our meshed model (see Figure 4).

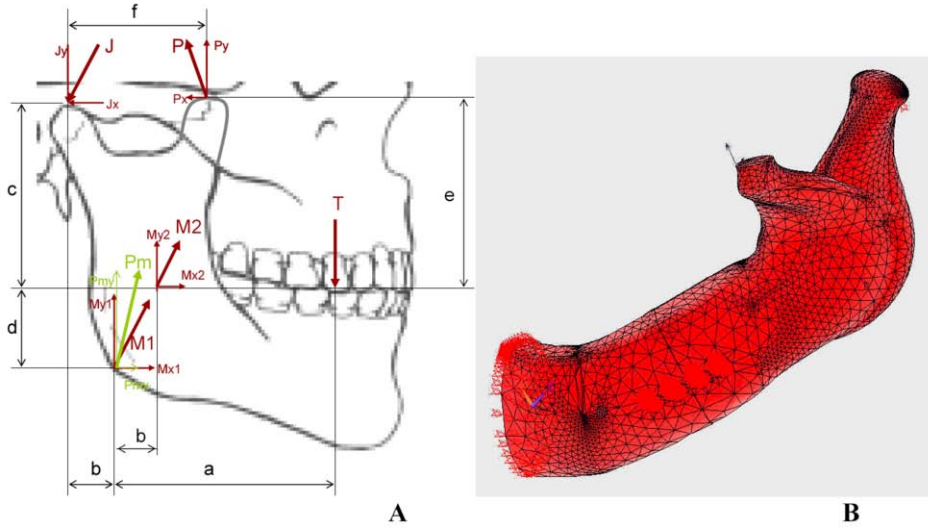


Figure 4. On the left-hand side (A) an image of a free-body-diagram in which P is the posterior part of temporalis, M1 is the superficial part of masseter, M2 is the deep part of masseter, Pm is the medial pterygoid and J is the lateral pterygoid muscle. On the right-hand (B) side is an image of the forces applied to the meshed model of the mandible.

On evaluation of the displacements, they are differentiated to give six strain distributions, 3 mutually perpendicular direct strains and 3 corresponding shear strains. Finally, six stress distributions are determined via the stress/strain relationships of the material. Figure 5 shows the stress and shear distributions in a 2D case. A point to note is that at least one of the displacements must be known before the rest can be determined (before the system of equations can be solved). These known displacements are referred to as *boundary conditions* and are oftentimes a zero value. An alternative solution may be obtained via the force matrix method. In the previous description, the displacements were the unknown, and solution is obtained via the stiffness method. In the force method, the forces are the nodal unknowns, while the displacements are known. The solution is obtained for the unknown forces via the flexibility matrix and the known displacements. The stiffness method is more powerful and applicable than the flexibility approach.

Typically in implant design, account has to be taken of the non-linear behaviour of biological tissues. In order to explain non-linearity in stress analyses, let us focus on the nature of linear solutions. The primary assumption in linear stress analysis is that the stress/strain relationship is directly related to the deformation. In general, there are four causes for nonlinear behaviour. It could be that the *material* deforms in a non-linear way. This is where the material stress-strain relationship is actively nonlinear. In this case, material behaviour depends on the current deformation state and possibly on the past history of the deformation.

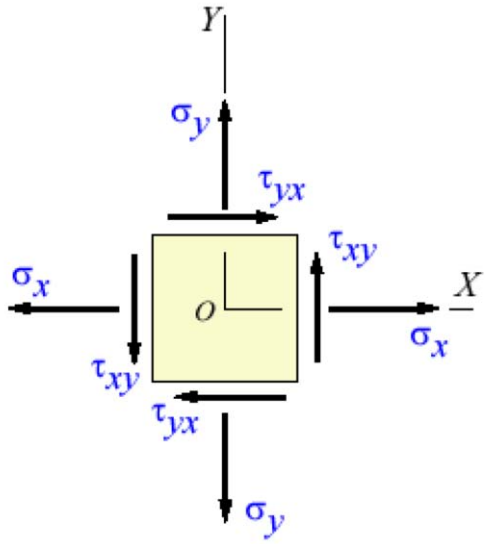


Figure 5. Stress and shear distributions in 2D

Material nonlinearity can be observed in structures undergoing nonlinear elasticity, plasticity, viscoelasticity, creep or other inelastic effects. There could be *geometrical* causes, which may be due to large strains (in impact trauma analysis) or small strains

but with large displacements (in the analysis of muscles). Muscles contractions are also a good example where nonlinear *forces* are applicable. This is where the magnitude or direction of the applied forces change with application to the structure. Displacement *boundary condition* nonlinearities, such as occurs in modelling of friction in knee joints. This is where the displacement boundary conditions depend on the deformation of the structure. The most important and obvious application is in contact problems, the displacement is highly dependent on the relationship between two contact surfaces (e.g. a fracture site in bones). It is important to note that the bodies in contact could be in a state of linear elastic stress; the nonlinearities all come from the contact definition in this case.

All non-linearities are solved by applying the load slowly (dividing it into a number of small load increments). The model is assumed to behave linearly for each load increment, and the change in model shape is calculated at each increment. Stresses are updated from increment to increment, until the full applied load is reached. In a nonlinear analysis, an initial condition at the start of each increment is the state of the model at the end of the previous one. This dependency provides a convenient method for following complex loading histories. At each increment, the solver iterates for equilibrium. Due to the iterative nature of the calculations, non-linear FEA is computationally expensive, but reflects the real life conditions more accurately than linear analyses.

2. The Power of FEA

The previous section has provided some of the background regarding FEA. But why is it commonly used in the analysis of implant biomechanics and implant design? These days it is very easy to obtain accurate models of various body parts of interest using MRI and CT-scanning. This means that irregular shaped objects are easy to analyse. Anatomical studies can be used as a great source of information regarding muscle attachments, insertions and directions. Analysis can be done for static as well as for dynamic loading conditions. All these features contribute to more sophisticated models that mimic *in vivo* situations in a more realistic way. But the true power of FEA lies in the ability to change parameters relatively easily once the model has been created.

In the case of the analysis of a fractured mandible, we can choose a specific material for the plate that stabilises the fracture site. The analysis will indicate what the local stress or strains are in the tissue. Perhaps one wants to investigate what happens if a softer and more flexible material is used. This would only require altering the material properties and rerunning the analysis. The same applies if the location and size of the plate or the fracture angle is to be investigated. This can provide valuable information, even at tissue level. The phenomenon of stress shielding is a common problem in implant design, which can result in tissue loss or altered structure. Typically a tissue is stimulated by the daily loads to which it is subjected. However, following the insertion of an implant, this load is partly reduced since it is shared by the implant. Factors that are of importance here are the geometry and the fit of the implant with the natural structure. Interface conditions between the implant and its surrounding tissue might be designed to promote bounding between them and therefore realise a better stress distribution from the implant to the surrounding tissue and so increase the lifetime of the implant.

References

[1] T.M. Wright and S.B. Goodman, Implant Wear in Total Joint Replacement: Clinical and Biological Issues, Material and Design Considerations, American Academy of Orthopaedic Surgeons, 2002

[2] P.J. Prendergast, Finite element models in tissue mechanics and orthopaedic implant design,. *Clin Biomech (Bristol, Avon)* **12** (1997),.343-366.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-67

Introduction to Chapter II. Bioelectronics

Richard B. REILLY and T. Clive LEE (eds.)

Bioelectronics can be defined as the interface area between electronics, photonics, computer science, and mathematics, on the one hand, and the nervous system at the molecular, cellular, organ, and systemic levels, on the other. Therefore it is the aim of this chapter to explore this definition and introduce bioelectronics from the fundamental properties of electricity, sensors for measuring biological activity and the analysis of electrical signals acquired from the body.

There have been numerous successes in bioengineering due to bioelectronics including the ability to measure brain activity using the electroencephalogram to provide information on the functioning of the brain, implantable bladder stimulators which can dramatically improve quality of life for paraplegics, cochlear implants which can restore partial hearing to the profoundly deaf, cardiac pacemakers, cardiac defibrillators and deep brain stimulation which improves the quality of life for those suffering from Parkinson's Disease.

Despite these successes which have had a profound effect on quality of care and quality of life for patients, bioelectronics is at a relatively primitive stage of development, but is recognised as being a strategically important area at the beginning of the 21st century.

As our understanding of human physiology advances, along with new developments in biomaterials and mathematical analysis, the impact of bioelectronics will increase. Therefore, an appreciation of the fundamental properties of bioelectricity is essential to apply bioelectronic approaches and devices in the clinic, while appreciating the trends in this area into the future.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-69

II.1. Elementary Electrodynamics

Jacques JOSSINET
Inserm, U556, Lyon, F-69003
University of Lyon, Lyon, F-69003, France

Abstract. The paper describes the physical phenomena involved in the conduction of electricity, with particular reference to living tissue. The conduction of electricity depends on the flow of charge carriers in the material, while the dielectric properties are due to the rotation of dipoles that can align along an applied electric field. The relation between the electric variables in a conducting medium and their physical meanings are explained. The phenomena responsible for the electric and dielectric properties of living tissues are described. The presence of cells limits the flow of charge carriers, in particular at low frequency, and the membranes are responsible for dielectric relaxation. The passive response of cell membranes to weak applied signals enables bioelectrical tissue characterisation. Practical tools for electrical impedance spectroscopy are given with an overview of the most recent applications.

Keywords. Electric conduction, dielectric permittivity, bioimpedance, tissue characterisation

1. Definitions

1.1. Electric Charge (q)

Electric charge (coulomb, C) is a property of matter and is transported at the atomic level by ions: elementary particles and charged molecules. There are two kinds of electric charge, positive and negative. The term "*charge*" may also represent a dimensionless point bearing a given electric charge. Electricity is quantized, meaning that charge always comes in integer multiples of the elementary charge (the electron charge, $e = 1.6\times10^{-19}$ C). Due to the great number of atoms in any piece of matter, electricity is handled at the macroscopic level as a continuous variable. The net charge of an object is the sum of the positive and negative charges its carries. The normal state of matter is electric equilibrium, with equal number of positive and negative charges. The conservation of electricity is verified in any process or reaction.

1.2. Electric Field $\vec{E}$

The electric field is a property of space in the presence of an electric charge. A point charge q_0 produces an *electric field* which is a vector, denoted $\vec{E}$, passing by the point charge and verifying the following equation

$$\left|\vec{E}\right| = q_0 / \left(4\pi\varepsilon_0 d^2\right) \quad (1)$$

The quantity ε_0 is the *dielectric permittivity* of free space and d is the distance to the source. The major property of the electric field is the force, $\vec{f} = q\vec{E}$, it exerts on a charge. Two charges exert opposite forces on each other. The electric force is attractive when the charges have opposite signs and repulsive when the charges have identical signs.

1.3. Electric Potential or Voltage (v)

Voltage or electric potential, *u*, at a given point is the electric potential energy per unit charge (in joules per coulomb = volts). The definition of potential is generally presented under the form: $\vec{E} = -\overrightarrow{grad(v)}$, where *grad* is the vector operator *gradient*. This vector represents the magnitude and direction of the greatest rate of change of the potential. This definition indicates that the existence of electric field in a medium implies a non uniform potential distribution. Conversely, if the potential is uniform, the field is equal to zero at any point (e.g. in a conductor at electric equilibrium).

1.4. Electric Current (i)

Electric current is the ensemble movement of charged particles, termed *charge carriers*. The electrons in a metal or the ions in an electrolyte are examples of charge carriers. The intensity of the current (ampere, A) is the quantity of electricity that is transported per unit time: current is the time derivative of electric charge:

$$i = \partial q / \partial t \quad (2)$$

1.5. Current density $\vec{j}$

It is often useful to consider the flow of electric charge per unit surface. This quantity is a vector termed current density, $\vec{j}$ (A/m^2). The current density through a surface element of area *dS*, is given by:

$$\vec{j} = \vec{n}.\overrightarrow{di} / dS \quad (3)$$

where $\vec{n}$ is the unity vector perpendicular to the surface. The dot product shows that only the orthogonal component of the flow gives rise to a current density through the surface.

The current density determines the effects of electricity in a medium. A practical example is the heating of a wire by the *Joule effect*. For a given current, the heating is greater in a conductor of small cross section than in a conductor of larger cross section. In body tissues, the desired or adverse effects of electrical current depend on how the current concentrates in certain pathways or spreads in a larger section, resulting in different current densities.

1.6. Electric Conductivity

Electric *conductivit*y (siemens, S) characterises the ability of a material to conduct electricity. The electric conductivity of an insulating material is equal to zero. The conduction of electricity through a medium implies the presence of charge carriers able to move within the medium. Charges than cannot move (bound charges) do not contribute to the conduction of electricity. If the charge carriers can move freely (e.g. electrons in metals) the medium is highly conductive. The ability of a charged particle to move within a given medium is characterised by its "*mobility*". Mobility is the limit speed reached by this particle in a given medium in the presence of a unit uniform and constant electric field. This speed is reached when the viscous force (increasing with the speed of the particle) becomes equal to the electric force $q\vec{E}$.

1.7. Static Conductivity of Solutions

In solution of dissolved electrolytes, the quantity of electricity that can be carried by an ionic species is proportional to its number of charge per ion, the concentration of dissolved ions and their mobility. The overall conduction σ_0, of a solution containing several species is the sum of the conduction of each species α_i. This is written as

$$\sigma_0 = F \sum_i \gamma_i \, \alpha_i \, C_i \, n_i \, \mu_i \qquad (4)$$

The activity coefficient, γ_i, ($\gamma \leq 1$) accounts for the deviation from linear behaviour for increasing ionic concentrations. For a given species, C_i is the concentration of this species in the solution, n_i the number of charges of the ion, and μ_i is the mobility and Σ denotes summation. The above equation deals with the static conductivity for direct current. However, the viscous forces delay the movement of charges carriers and the establishment of the steady state is not instantaneous. This effect however becomes appreciable for rapidly varying fields. Hence, in biomedical applications the conductivity of ionic solutions can be considered independent of the applied field frequency.

1.8. Ohm's Law

Ohm's law relates to the proportionality between the DC current, denoted *I*, and the DC voltage difference, denoted *V*, across an ideal conductor, circuit or current pathway. The general expression of Ohm's law is

$$V = R \times I \qquad (5)$$

The constant of proportionality, *R*, is termed the "*resistance*", (ohm, Ω) of the circuit. If the resistance of a conductor is constant over a wide (ideally infinite) range of voltages, the conductor is said to be "*ohmic*".

The reciprocal of the resistance is termed the "*conductance*", G (siemens, S). Ohm's law also applies at the macroscopic scale and then is written differently.

Considering the flow of charge in an elementary volume element, it can be demonstrated that the "local" expression of Ohm's law is:

$$\vec{E} = \vec{J} / \sigma = \rho \vec{J} \tag{6}$$

The quantity ρ, reciprocal of the conductivity σ, is the resistivity of the medium (ohm.metre). This equation shows that for the current density, $\vec{J}$ (A/m^2), i.e. the quantity of electricity passing though a unit surface per unit time is proportional to the electric field and the electric conductivity of the medium.

The heat production per unit time (energy) due to the current in a circuit is the electrical "*power*" W, measured in watts. In a circuit of resistance *R* passed by a current *I* the dissipated energy is RI^2. Using Ohm's law this also writes V^2/R, with *V* and *R* defined above. Instead of a circuit's resistance, one may consider the energy dissipated within a medium. The energy dissipated in a volume element of volume *dt*, within a medium. In such an element, the energy dissipated per unit volume is given by

$$dW = J^2 / \sigma \tag{7}$$

(where σ is the electric conductivity and *J* the current density). This equation shows that in a heterogeneous medium, the local Joule effect depends on the conductivity and current density at each point.

1.9. Impedance, Admittance,

In a circuit passed by a sinusoidal signal, the voltage and the current vary in synchrony, but are shifted in time, according to the properties of the circuit. The voltage-to-current ratio is termed *impedance*.

A sinusoidal signal (voltage or current) is characterised by its magnitude and phase angle and is represented by a complex number formed by a *real part* and an *imaginary part*.

It is expressed as a complex number denoted *Z*. The *real part* and the *imaginary part* of *Z* are denoted *R* and *X* respectively.

The *magnitude* $|Z|$ of this complex number is the ratio of the magnitudes (*V/I, V* in volt and *I* in ampere).The *argument*, θ, is the phase difference (in radians) between voltage and current. Simple mathematical considerations lead to the following classic equations.

$$|Z| = \left(R^2 + X^2 \right), \; R = R|Z| \cos\theta \quad \text{and} \quad X = |Z| \sin\theta \tag{8a}$$

The real part, *R*, of the impedance is termed *resistance* and its imaginary part, *X*, is termed its *reactance*. Impedance is the generalisation to sinusoidal signals for the notion of resistance for DC current. The reciprocal of impedance is also a complex number denoted Y ($|Y|, \theta'$), termed admittance, and verifying:

$$Y = 1/Z, \; |Y| = 1/|Z| \text{ and } \theta' = -\theta \tag{8b}$$

The impedance of a capacitor is given by $X = 1/(j2\pi fC)$, where C is the capacity (farad, F) and f the applied signal frequency (hertz, Hz).

1.10. Dielectric Permittivity

Dielectric properties are observed in a medium containing bounded *electric dipoles*. An electric dipole is the rigid system of two equal and opposite point charges, located at points Q-, Q+, and bearing opposite charges of equal magnitude, $-q$ and $+q$, respectively. The dipole is characterised by its *moment* which is the vector defined by

$$\overrightarrow{M} = q \times \overrightarrow{Q_- Q_+} \tag{9}$$

The dipoles in a dielectric medium cannot move but can rotate in the presence of an electric field. The two charges experience opposite forces. It can be shown that these forces create a torque tending to align the dipole along the direction of the electric field.

The rotation of dipoles creates the *dielectric polarisation* of the material. Near the boundary surface, the orientation of dipoles makes charges of a given polarity come closer to the surface and the charges of the opposite polarity move away from it. This results in a distribution of charges at the outer surface of the material. This distribution depends on the concentration of dipoles and their ability to rotate. A dielectric material is characterised either by its dielectric permittivity ε (farad/metre) or by its "*dielectric constant*" $\varepsilon_r = \varepsilon/\varepsilon_0$ (dimensionless).

1.11. Capacitive Coupling C

The polarisation of a dielectric slab in the presence of an orthogonal field results in surface charges of opposite signs on each side of the slab. These surface charges can attract or repel charges from the outer circuit connected to the slab. Similarly, the external circuit can bring or remove charges at the surface and produce polarisation of the dielectric slab. In both cases, the corresponding flows of charges is termed "displacement current". It can be shown that the charges on both sides are of opposite signs and equal magnitudes and vary in synchrony. This effect is termed the "capacitive coupling", as the charges seem to pass from one side to the other one. In practice, capacitive coupling appears when a membrane separates two conductors, forming a capacitor. The "*capacity*" (farad, F) is the ratio of the magnitude of the charge accumulated on either side to the voltage across the dielectric. The capacity of a capacitor increases with the area of the plates and decreases with the separation between the plates according to the relation

$$C = \varepsilon_0 \varepsilon \frac{S}{d} \tag{10}$$

S is the area of charged surfaces, d is the distance between them and ε is the dielectric constant of the material. The main feature of a capacitor is to block DC signals and be permeable to varying signals.

1.12 Dielectric Polarisation

The rotation of dipoles in a dielectric is hindered by viscous forces exerted by the medium. Dielectric permittivity has therefore a dependence on frequency. This dependence is called the dielectric dispersion. The associated time constant is called the "relaxation time" of the medium. The term dielectric relaxation denotes the delayed polarisation in response to the application of an electric field.

The time lag between electrical field and dipole rotation implies irreversible energy losses (dielectric losses). For ideal dielectrics, containing dipoles with a unique time constant, τ, the relaxation is described in terms of permittivity as a function of frequency, which can be described by Debye's equation:

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{1 + j\omega\tau} \tag{11}$$

$\Delta\varepsilon$ is the dielectric increment, ε_∞ the high frequency limit permittivity (both characterising the material), ω is the angular frequency ($\omega = 2\pi f$) and j the base of complex numbers ($j^2 = -1$).

2. Electric Properties of Cells and Tissues

The conduction of electricity in the human body presents a particular set of features that is not met is other materials, except in animal tissues and, to some extent in plants. The main feature is the simultaneous existence of electric conduction and dielectric behaviour. The presence of cells is responsible for the remarkable electric and dielectric properties of tissue.

2.1. The Cell

The cell is the basic building block of the human body. The living cell is delimited by the *plasma membrane* providing rigidity and protection. The cell membrane also binds different cells together. The membrane consists of an asymmetric molecular lipid bi-layer which is basically impermeable to ions. However, protein molecules inserted in the bi-layer ensure the transport of materials, including ions, across the membrane. The interior of the cell contains the nucleus and the cytoplasm which is a jelly-like material consisting of the cytosol and the organelles. The cytosol is made up of water, salts, and organic molecules and is therefore conductive. The content and structure of the interior of a cell depend of each cell type.

2.2. Membrane potential

Under normal conditions of activity, there is a difference in ion concentrations on both sides of the cell membrane. This charge separation results in an electrostatic field and a voltage difference across the membrane. The equilibrium voltage (or *resting potential*) corresponds to the balance between the flow of ions due to the electrostatic field and

the flow produced by the ionic concentration gradient. This potential difference is called the *Nernst potential*. The resting value for nerve cells is about 70 mV.

2.3. Biological Tissue

A tissue is formed by cells performing a similar function. An organ is a structure containing several types tissues working together to carry out specific functions. The archetypal representation of a tissue is a population of cells separated by the interstitial space. However, its structure may be complex, particularly in mammalian tissues. The extracellular matrix is a structural entity supporting cells; it is composed of 3 major classes of biomolecules including structural proteins (collagen and elastin), specialized proteins and proteoglycans (composed of a protein core attached to glycosaminoglycans, long chains of repeating disaccharide units). There is a great variability in cell shape, concentration and arrangement, depending on each tissue's structure, function and physiological state. The interstitial space between cells contains a conductive medium the conductivity of which depends on the particular tissue. For these reasons, the dielectric properties of body tissue are highly variable from one tissue to another.

2.4. Cell Membrane Passive Properties

The overall conduction of electric signals through a tissue depends on the conductivity of the interstitial space, that of the intracellular medium and the capacitive behaviour of the cell membranes. The membrane is practically impermeable to DC current, except for the active transport of materials through it by the trans-membrane proteins. For applied signals, the membrane acts like a passive insulator separating two conducting media and causes capacitive coupling between the interstitial space and the interior of the cell. The capacitance of the cell membrane is about 1 $\mu F/cm^2$ [7, 12]. Finally, DC and low frequency signals do not penetrate into cells. For increasing applied signal frequency, the interior of cells is progressively involved in the overall conduction due to capacitive coupling through the membrane.

2.5. Dielectric Relaxation in Tissue

Observation shows that body tissues are dispersive media. In a dispersive dielectric medium, the permittivity and dielectric constant vary with the applied field frequency. In general, the dielectric relaxation in body tissue does not follow in the Debye equation. Experiments show that, in the frequency range from a few hertz to many gigahertz ($1\ GHz = 10^9\ Hz$), the dielectric permittivity of tissues is a decreasing function of frequency and shows several dispersion domains. Several types of phenomena, preponderant in different frequency domains, are responsible for the observed dispersion [6, 14, 15].

2.5.1. α -Relaxation

This relaxation occurs at low frequency roughly below 1 Hz. This effect is due to the counter-ion atmosphere surrounding the membrane surface. In the presence of an electric field, the ions in the cell vicinity accumulate on diametrically opposed locations and then form a cell-size dipole. Such a dipole cannot follow rapidly varying

signals. This low frequency relaxation does not give information on either cell membrane or cellular content.

2.5.2. β *-Relaxation*

This structural relaxation is also termed Maxwell-Wagner relaxation. It may occur up to the radio-frequency range, depending on the properties of each particular tissue. The basic mechanism is the accumulation of charge at interfaces such as cell membranes. In body tissue, the time constant of this relaxation depends on the conduction in the media and the capacitive coupling through the interfaces.

2.5.3. γ *-Relaxation*

This relaxation is due to the orientation of dipoles. Several types of dipoles contribute to the polarisation of the medium. It takes place typically in the microwave frequency domain ($f < 1$ GHz). The permanent dipoles include dipolar molecules (mainly H_2O), hydrogen ionic or covalent bonding. Induced dipoles result from the temporary separation from the barycentre (the point between two objects where they balance each other) of positive and negative charges in non-polar molecules by either neighbouring molecules or the applied field. Transient dipoles (also called London forces or van der Waals forces) result from the random mutual perturbation of electron clouds of two atoms or molecules. When an uneven distribution occurs, a temporary dipole is created. In polar molecules, transient dipoles are usually smaller than the permanent dipolar moment.

The above relaxation phenomena give tissues high dielectric constants compared to usual materials (Table 1).

2.6. Equivalent circuit model

An equivalent circuit model (ECM) is a circuit yielding the same impedance as a tissue in a given frequency domain. There are *lumped circuits* and *discrete element* circuit models. The most commonly used discrete element ECM are composed of two resistors and one capacitive element. The response of such circuits is characterised by a single central frequency; denoted f_0 according to the general equation [10, 11]:

$$Z = R_\infty + \frac{R_0 - R_\infty}{1 + j\left(f / f_0\right)^\beta} \tag{12}$$

Z denotes the overall impedance of the circuit, R_0 the low frequency limit resistance, R_∞ the high frequency limit resistance and j the base of the complex numbers ($j^2 = -1$). Exponent β is termed the "*fractional power*" ($\beta \leq 1$). At frequency f_0 the magnitude of the imaginary part passes by a maximum. The plot in the complex plane of the imaginary part against the real part (termed an Argand diagram) is an arc of circle the centre of which is shifted from the horizontal axis (Fig. 1).

Table 1. Example values of electric conductivity ($S.m^{-1}$) and dielectric constant in tissues and some usual materials at 50 Hz or "below 100 Hz" [2]. The above figures are supplied for illustration and different values can be found according to measurement frequencies and experimental conditions [12].

	Conductivity Siemens/metre	Dielectric constant
Vacuum	0	1
Dry air	3×10^{-15}	1
Copper	56×10^{6}	1
PolyVinyl Chloride	10^{-6}	3.4
Water at 25 °C.	5.5×10^{-6}	78
NaCl saline 0.9 g/l	1.68	77
Kidney	0.089	1.0×10^{7}
Fat	0.02	1.5×10^{6}
Nerve	0.027	1.6×10^{6}
Liver	0.037	1.8×10^{6}
Grey matter	0.075	1.2×10^{7}
Muscle	0.23	1.8×10^{7}
Tendon	0.27	1.7×10^{7}
Whole blood	0.7	5.3×10^{3}
Cerebro-spinal fluid	2	109

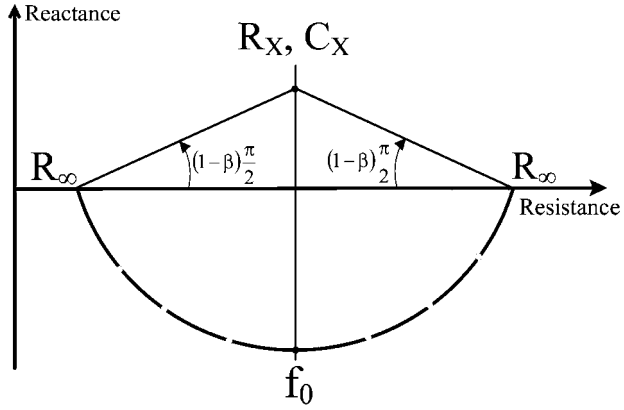


Figure 1. Sketch of Eq. (13) in the complex plane: Horizontal axis is the resistance and vertical axis is the reactance. The reactance of body tissues is negative. The white dots represent particular measurement frequency points. The abscissa of the intercepts with the horizontal axis are the so-called "*limit resistances*" at low frequency (R_0) and at high frequency (R_∞). The bottom of the arc corresponds to f_0 in Eq. (13). The shift of the centre from the horizontal axis is a function of the fractional power β .

There are two types of 3-elements circuits giving rise to the above response in Fig. 2. In these circuits, Z_{cpa} is a "pseudo-capacitance" (subscript "cpa" stands for "constant phase angle"). Its frequency response is given by $Z_{cpa} = K(j\omega)^{-\beta}$. The exponent β is the fractional power in (12). A constant phase element is a hypothetical element that does not exist in practice, but is useful to describe the frequency response of body tissues. If the particular case where β is equal to unity, Z_{cpa} is equivalent to a capacitor of capacity $C=1/K$. The relation between ECMs' elements and the variables in (12) are given in Table 2.

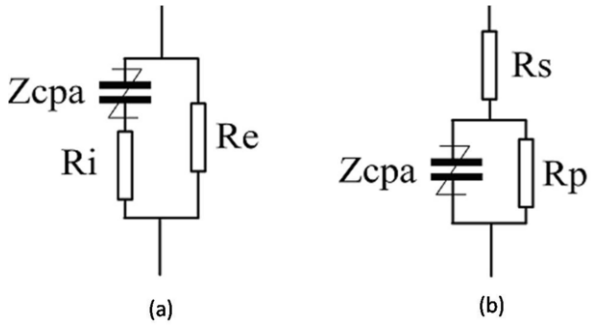


Figure 2. Equivalent circuit models comprising 3-element and yielding identical frequency responses. Type (a) circuit is generally used for the deep tissue, with Re representing the resistance of the interstitial space and Ri the intracellular resistance. Type (b) circuit is generally used to represent the impedance of layered media such as skin.

In practice, fitting an arc of circle to the experimental data points enables the calculation of R_0, R_∞, and β (Table 2). The calculation of these parameters makes it possible to compare data sets obtained at different measurement frequencies. Table 3 gives the equations enabling the calculation of the elements of ECM using the four parameters resulting from the fit of a circular arc to the experimental data. The table confirms that, with appropriate values of their elements, both circuits can model a given response (same R_0, R_∞, K and β).

Table 2: Equations for R_0, R_∞ in equation (12) for the equivalent circuit models of Fig. 2.

ECM	R_0	R_∞	F_T
(a)	$R_0 = Re$	$R_\infty = Re\,Ri/(Re+Ri)$	$F_T = \frac{1}{2\pi}\left(K\frac{R_0 - R_\infty}{R_0^2}\right)^{1/\beta}$
(b)	$R_0 = Rp + Rs$	$R_\infty = Rs$	$F_T = \frac{1}{2\pi}\left(\frac{K}{R_0 - R_\infty}\right)^{1/\beta}$

Table 3: Equations to calculate the elements of the equivalent circuit models of Fig. 2a and Fig. 2b from the variable in Equ. (12) Quantity X_c denotes the ordinate of the centre of the circular arc. Centre's abscissa is equal to $(R_0+R_\infty)/2$.

ECM			β	K
(a)	$Re = R_0$	$Ri = R_0 R_\infty / (R_0 - R_\infty)$	$\frac{2}{\pi}\text{atan}\left(\frac{R_0 - R_\infty}{2X_C}\right)$	$(R_0 - R_\infty)(2\pi F_T)^\beta$
(b)	$Rp = R_0 - R_\infty$	$Rs = R_\infty$	$\frac{2}{\pi}\text{atan}\left(\frac{R_0 - R_\infty}{2X_C}\right)$	$(R_0 - R_\infty)(2\pi F_T)^\beta$

3. Applications

The measurement of the electric and dielectric properties of tissues has given rise to range of biomedical applications [1, 8]. The most widespread are EIS, EIT and BIA. *Electrical Impedance spectroscopy* (EIS) is the study of the frequency response of a tissue or an organ to characterise its physiological and/or pathological state (e.g. detection of oedema, ischaemia, inflammation, and tumour growth) [7]. *Electrical Impedance Tomography* is an imaging method that appeared in the 1980s. The purpose is the reconstruction of the distribution of conductivity inside the body from multiple surface measurements. The main features of this method are the relatively small number of measurements, limited spatial resolution, high time resolution, real-time imaging capability and true3-D image reconstruction. This method can be associated with EIS [3, 4, 5, 9].

Impedance measurements at the body level or across limb segments are termed *Bio-Impedance Analysis* (BIA) [13]. The purpose is to monitor the changes in fat free mass and/or total body water. This method has given rise to devices for various applications including fitness, sports and nutrition. One popular implementation is the incorporation of BIA system in personal weighing scales to monitor the effects of exercise and/or nutritional regimes.

The above applications involve measurements using low intensity currents (of the order of milliamps or less) avoiding any significant interaction with the explored tissue at the used measurement frequencies. Therapeutic devices are based on the interaction of relatively strong signals with the body. Typical applications include electrosurgery, transcutaneous electric stimulation of muscles or nerve (TENS), cardiac pacing and defibrillation.

References

[1] J. Ackman and M. Seitz, Methods of complex impedance measurements in biologic tissue, *CRC Crit. Rev. in Biomed. Eng.* 11(1984), 281-311.

[2] D. Andreuccetti, R. Fossi and C. Petrucci, *Internet resource for the calculation of Dielectric Properties of Body Tissues*, Italian National Research Council, Institute for Applied Physics "Nello Carrara", Florence.

[3] K. Boone, D. Barber and B.H. Brown, Imaging with electricity: report of the European concerted action on impedance tomography, *J. Med .Eng. Technol*. 21(1997), 201-232.

[4] S.G. Dawids, Evaluation of applied potential tomography: a clinician's view, *Clin. Phys. Physiol. Meas*., 8A(1987), 175-180.

[5] A.M. Dijkstra, B.H. Brown *et al.*, Clinical applications of electrical impedance tomography, *J. Med. Eng. Technol.* 17(1993), 89-98.

[6] K.R. Foster and H.P. Schwan1, Dielectric properties of tissues and biological materials: a critical review, *CRC Critical Reviews in Biomedical Engineering*, 17(1989), 25-104.

[7] J.P. Grant and N.M. Spyrou, Complex permittivity differences between normal and pathological tissues: mechanisms and medical significance, *J. Bioelectricity*, 4(1985), 419-458.

[8] P. Heroux and M. Bourdages Monitoring living tissues by electrical impedance spectroscopy, *Ann. Biomed. Eng*., 22(1994), 328-337.

[9] H.C.N. Jongschapp, R. Wytch, J.M.S. Hutchinson andV. Kulkarni, Electrical impedance tomography: A review of current literature, *Europ. J. of Radiology,* 18(1994):165-174.

[10] J.R. MacDonald, Impedance spectroscopy, *Ann. Biomed. Eng*., 20(1992), 289-305.

[11] J.P Morucci, M.E. Valentinuzzi, B. Rigaud, C.J. Felice, N. Chauveau and P.M. Marsili, Bioelectrical Impedance techniques in Medicine, *Critical Reviews in Biomedical Engineering*, J.R. Bourne Ed., 24(1996), ISSN 0278-940X.

[12] R. Pethig, Dielectric properties of biological materials, *Clin. Phys. Physiol. Meas*, 8A(1984), 5-12.

[13] P.J. Riu, J. Rosell, R. Bragosand and O. Casas, Electrical Bioimpedance Methods, *Ann. NY Acd. Sci.*, **873**(1999), ISBN 1-57331-191-X.

[14] H.P. Schwan, Electrical properties of tissue and cell suspensions, *Advan. Biol. Med. Phys.*, 5(1957), 147-208.

[15] R.D. Stoy, K.R. Foster and H.P. Schwan, Dielectric properties of mammalian tissues from 0.1 to 100 MHz:A summary of recent data, *Phys. Med. Biol*., 27(1982), 501-513.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-81

II.2. Electrical Safety

Jacques JOSSINET
Inserm, U556, Lyon, F-69003,
University of Lyon, Lyon, F-69003, France

Abstract. Correct use of medical equipment within the clinical environment is of prime importance. This includes awareness of the safety issues regarding equipment, particular when it is an electrically powered device. Incidents can occur in the clinic in which a medical device is suspected of contributing to patient or staff injury. It is important that one can identify in advance any potential hazards which may arise with electrical equipment due to technical or environmental factors. This paper gives an overview of electrical safety.

Keywords. Electrical Safety, electrical injury, shock, prevention, first aid

Introduction

The contact with either a man-made or natural source of electrical energy can create injury. Such a contact is termed "*electrification*". The term "*electrocution*" should be reserved for cases resulting in a victim's death. The detrimental effects of electrification depend on several parameters including the source of electrical energy and current conduction through the body.

For electrification to occur, the victim must complete the circuit: at least two *contact points* are necessary with sufficient voltage difference to drive the current through the circuit. Electrification can have direct and indirect effects. For instance, the explosive force produced by lightning can throw the victim a distance and the sudden muscular spasm caused by the current can cause the person to fall from a height.

However, most lesions and trauma due to electrification are the direct effects of current. It must be pointed out that it is the current that produces lesions. This is confirmed, for instance, by the harmlessness of the static discharge, where the voltage is high, but the current is limited in intensity and duration. This feature is reminded by the familiar slogan:

"*The volts jolt, the mills kill",* mills signifying milliamps, the unit of current

For a given applied voltage, the current is controlled by the series combination of the impedance of the sources and the impedance of the body through the contact points. In lightning and electrostatic discharge (ESD), the resistance of the body is small compared to that of the total circuit and has very little influence on current intensity. In contrast, body impedance is the most important variable when considering electrocution by power lines.

Note: In the literature on electrical safety, tissues are often characterised by their *resistances*. This is incorrect use of the terminology since for AC current a tissue is characterised by its impedance. However, at the frequencies used for power distribution

(e.g.: 50 Hz in Europe or 60 Hz in the U.S.A), the imaginary part (the reactance) of tissue impedance is small and is generally ignored.

There is a wide variation in body resistance between people; therefore the same voltage level may result in different effects. Skin impedance is normally high and acts as an insulating barrier. This feature, however, disappears if skin is wet or damaged. For given tissue resistances, the nature and importance of lesions created in the body by electric current depends on the electrical parameters of the source of electrification [11].

1. Sources of Electrical Hazard

1.1. Electrostatic Discharge

Electrostatic electricity is that which accumulates at the surface of persons or objects (e.g.: carpets, pieces of furniture and clothing). It is mainly produced by friction *triboelectricity*. Contact with charged objects results in electric shocks due to the discharge current through the body. The voltage can reach several kilovolt or more, but the duration of the current is so short that there is no dangerous effect on a person. Electrostatic discharge, however, can damage electronic components and produce sparks powerful enough to cause a fire or explosion in the presence of an explosive atmosphere. Specific equipment for ESD protective workstations includes a dissipative floor or table mat, dissipative work surface, ground connected wrist strap and a common point ground system.

1.2. Lightning

Lightning is the natural and violent discharge between an electrically loaded cloud (cumulonimbus) and the ground. Once the "*stepper*" has established the path, the current flows along the "*arc*", i.e. a tube of plasma (high temperature, highly conductive ionised matter). Lightning voltages are typically greater than 20 kV and can reach millions of volts. The average current during the discharge is about 20-30 kA. The lightning generally consists of several strokes and lasts from about 0.2 to 1 sec. Thunder is due to the shock wave resulting from the rapid temperature change of the air in the close vicinity of the arc.

Direct stroke is not the only danger of lightning. The lightning current spreads through the ground and gives rise to gradients in ground potential. The "*step voltage*" is defined as the voltage difference in electrical potential between a person's feet. As a body is a far better conductor of electricity than the earth, a current can flow through the legs and cause lesions, trauma and even death, especially if excessive current passes through the heart [8].

1.3. Power Lines

By construction, power lines (and the mains) maintain a constant voltage even if large current is current is drawn either in the circuit connected to it or through the body of a person. The contact with power lines can therefore result in severe electrical injury, as a large current can continue to pass through the body as long as the contact is

maintained. This is particularly critical when the victim is not able to release the contact, above the so-called *"let go threshold*".

High-voltage power lines can also give rise to arc-flash between the line and conducting objects connected to the ground. In air, this may occur for distances shorter than 1 cm for voltage levels of approximately 4000 volts. The arc is made of plasma and gases at high temperature produce second and third degree burns by direct contact or by ignition clothing. Arcs can also produce pressure wave caused by the rapid thermal expansion of gases, giving rise to the explosion of circuit components.

2. Electrical injury

The detrimental effects of electricity through the body can range from a barely perceptible tingle to immediate cardiac arrest [2, 12]. The undesirable effects of electric current consist of *burns* and *electric shocks*. Burns are due to the excessive electrical energy dissipated in the tissue. An electrical current flowing through the body can be hazardous or fatal if it causes local current densities in vital organs that are sufficient to interfere with their functions.

2.1. Burns

Burns are produced by the heating of body tissues (Joule effect). The appreciable conductivity differences between organs result in an inhomogeneous distribution of current density. Consequently, the thermal effects on electrification (burns) depend on the pathway of the current through the body. Electrical burns are usually most severe at the contact points, where current concentration results in high current density. This effect is exploited in electrocautery devices for the cauterization of tissue, where the tissue is burned by an electrical current. Electrosurgery uses alternating current to directly heat the tissue itself. This is known as diathermy.

2.2. Electric shock

Electric shock is the physiological and pathological effect of the passage of an electrical current through the body. The electric current may result in action potentials and produce nerve stimulation and muscle contraction (also termed *tetany*). According to the intensity and path, electrification can produce spasm of the muscles, lung paralysis, loss of consciousness, irregular heartbeat, heart fibrillation and/or cardiac arrest [4, 5, 10].

3. Factors Influencing Electrical Injury

The severity of electrical injury depends on the *type of source*, the *intensity* of the current, the *pathway* through the body and the *duration of the contact*. Other factors are the applied current frequency, the phase of the heart cycle when the shock occurs and the general health status of the person.

The effect of electric shock decreases with applied signal frequency [7]. High frequencies currents do not excite muscles and do not cause cardiac arrhythmias.

Unless otherwise noted, the values quoted here are implicitly for the frequency of power distribution.

The current and pathway of the current through the body depend on the location and impedance of contact points and the resistance of the body between these points. This latter issue depends on each subject's morphology and body composition. Finally, it must be kept in mind that there are no absolute thresholds or known values quantifying the exact injury from a given current or voltage.

3.1. Voltage

Sources producing electrical injury are generally classified into *"low* voltage" (less than 500 volts) and "*high voltage*" (greater than 1000 volts). Both high and low voltage can be fatal. Low voltage injuries occur at home or in a residential environment. Electrocutions in bathtubs and by electric dryers are the most common causes of low-voltage deaths.

Low voltage does not imply low hazard. In most cases, electrification by the mains does not cause skin damage unless the contact point is small or the victim has delicate skin. A low-voltage source can, however, produce major cardiopulmonary complications and death if a sufficient current passes across the chest (depending on the resistance) to induce ventricular fibrillation (depending on the pathway).

High-voltage injuries generally occur in an outdoor environment near power sources and power lines, either by direct contact or arc flashing. High-voltage has a greater potential for tissue destruction and can be responsible for severe injuries leading to amputations and tissue loss. However, remembering that the adverse effect of electrification is due to current a shock of 100 volts is not less dangerous than a shock of 10000 volts. Individuals have been electrocuted by appliances connected to the mains.

3.2. Amperage

The term "*amperage*" refers to the intensity of the current passing through the body. The "*let-go threshold*" is the limit at which a person becomes unable to let go off the current source because of muscular tetany. Typical values of let-go currents for men, women and children are about 9, 7 and 4 mA, respectively. Electricians familiar with this effect often refer to an immobilized victim of electric shock as being "frozen on the circuit." Typical effects of electrification of the human body by 50 or 60 Hz AC currents are summarized in Table 1.

3.3. Source Type

The type of circuit involved, either direct current (DC) or alternating current (AC) is one of the factors affecting the nature and severity of electrical injury. Direct current (ESD, lightning, batteries) tends to throw the victim from the source after. Alternating current (power lines) is more dangerous for it causes muscle contraction that can maintain the contact with the power source. High-voltage AC injuries are more severe than DC ones mainly because of increased risk of muscular tetany and prolonged contact. AC exposures are said to be three times more dangerous than DC exposure at the same voltage.

Table 1. Archetypal effects of electric current on the human body against the applied currents for frequency of 50/60 Hz

Current (mA)	Effect
0.1 - 1	Perception threshold
1 - 4	Faint tingling sensation
1 - 10	Muscular contraction, slight shock, not painful but disturbing, “no let go” danger
6 - 30	Painful shock, muscular contraction, muscular control lost
20 - 75	Breathing becomes difficult due to thoracic muscle tetany
50 - 150	Extreme pain, severe muscular contraction, respiratory arrest, ventricular fibrillation above 100 mA, death is possible
100 - 200	Currents between 100 mA and 200 mA are lethal
200 - 1000	Severe burns, unconsciousness, strong muscular contraction, heart clamped during the shock (no ventricular fibrillation), good chances for survival if the victim is given immediate attention
1000 - 4300	Muscular contraction, nerve damage, ventricular fibrillation, death is most likely
10000	Cardiac arrest, severe burns and probable death

3.4. Duration of Exposure

The severity of injury is increases with the time of current flow through the body. The effects of short duration current are more shock than burns. For instance, the electric ray produce discharges of several hundred volts. This fish, however, does not produce lethal shocks since the discharge duration, of the order of tens of microseconds, is too short.
However, even brief exposures to very high amperage can produce important tissue damage. Lightning current is seldom long enough to cause severe burns, but it produces electric shock than can be fatal. Thoracic tetany can occur at levels just above the let-go current and result in respiratory arrest.

3.5. Current Pathway

The pathway of the current between the *contact points* determines the tissues at risk, the type of injury regardless of whether high, low, or lightning voltages are being

considered.

The hand is the most common site of *source contact point*. Bone, tendon, and fat have a very high resistance and tend to heat up and coagulate rather than transmit current. Similarly high resistance skin resistance dissipates much of the energy, producing surface burns and limiting internal damage. If the current were delivered through low resistance contact points, it can produce cardiac arrest but no surface burns, such as in a bathtub injury.

3.6. Contact Impedance

The impedance at the contact point depends on the piece of material in contact with the skin and the properties of the skin. In particular, the contact impedance depends on the contact surface area: the wider the area, the lower the contact resistance. Hence, the contact impedance can be different if the victim is grasping a handle or coming accidentally into contact with an electrified object. In practice, the contribution of skin resistance to the contact impedance is preponderant.

For usual contacts, practical values of skin resistance vary from 1000 Ω for wet skin to over 500 KΩ for dry skin. When a victim comes into contact with a high-voltage power source, the epidermis may be destroyed by heat within milliseconds. Moisture, sweating and immersion in water can dramatically reduce skin resistance. Pure water is a poor conductor, but tap water is almost as conductive as body tissues. In practice, water always increases the risk of electrical injury by decreasing the resistance of contact points.

3.7. Tissue Resistance

Body tissues other than skin are relatively good conductors; large currents can then pass through them and produce tissue damage [1,6]. The resistances of the organs determine the distribution of the current through the body for the current tends to pass through low resistance regions. Nerves, muscles, and blood are better electrical conductors than bone, tendon, and fat. Other body tissues have intermediate resistances.

The actual resistance of the body varies depending upon the points of contact and the skin condition (moist or dry). For instance, skin resistance excluded, the internal resistance between the ears is about 100 ohms, while the resistance from hand to foot is about 500 ohms.

4. Symptoms of Electrical Injury and First Aid to Victims

The outcome of an electric shock depends on the speed and adequacy of the treatment [3,9]. The current's path through the body determines for prognosis and therapy. It is impossible to know the exact amperage because of the variability of the resistance. However, an estimate can be calculated knowing the voltage of the source. The pathway can be estimated knowing the contact points.

Pain is the least significant result of electric shock. The outside of the victim’s body may appear to have only minor lesions, but the internal injuries may still be significant. More generally, the symptoms include:

- ✓ Skin burns,
- ✓ tingling, weakness, unconsciousness,
- ✓ Muscle contraction, muscular pain,
- ✓ Bone fractures
- ✓ Headache, hearing impairment, seizures
- ✓ Respiratory failure, arrhythmias, cardiac arrest.

4.1. Cares to Victims

The classical *dos and don'ts* can be summarized as follows:

1. Switch off the electrical current at the control box, if safely possible.
2. Call for emergency medical help.
3. If the current can't be switched off, use an insulator (dry rope, cloth, broom handle…) to drag the victim away from the contact with the source.
4. Once the victim is free from the source of electricity, check the victim's breathing and pulse. If either has stopped or seems dangerously slow or shallow, initiate first aid..
5. Give first aid for burns
6. If the victim is faint, pale, or shows other signs of shock, lay the victim down, and cover the person with a blanket or a coat.
7. Stay with the victim until medical help arrives.
8. Avoid moving the victim's head or neck if a spinal injury is suspected. Give first aid needed for other wounds or fractures.

DO NOT	get close to a victim in contact with a high-voltage source
DO NOT	touch the victim in contact with the source of electricity with bare hands
DO NOT	attempt to rescue a victim near active high-voltage lines
DO NOT	move a victim of electrical injury unless there is immediate danger.
DO NOT	remove dead skin or break blisters if the victim has acquired burns.
DO NOT	apply ice, butter, medications, cotton dressings or adhesive bandages to a burn.

5. Protection and prevention measures

Simple rules improve electrical safety at home and at the laboratory. The use of electricity at the working place must be in agreement with the general safety regulations and the specific ones applying to every particular type of activity (e.g. use of chemical, flammable gases, protection against ESD). However, the application of simple basic rules minimizes electrical hazards at home and in the laboratory. Some of them are given below.

5.1. General Electric Safety Tips

- ✓ Unplug electrical appliances before cleaning
- ✓ Always unplug an appliance that overheats and have it checked by a qualified repair person before using it again,

- ✓ Check appliances and extension cords for fraying or cracking,
- ✓ Always pull the plug and not the cord; never carry an appliance by the cord

5,2, Tips for the laboratory

- ✓ Maintain equipment properly
- ✓ Learn the location of electrical panels and shut-off switches for a quick power disconnection in the event of an emergency
- ✓ Minimal clearance of a 1 metre around electrical panels for ready access
- ✓ Don't overload circuits by using power strips or multiple outlets on regular sockets.
- ✓ Carefully place power cords so they don't come in contact with water or chemicals
- ✓ Do not allow cords to dangle from counters or hoods
- ✓ Do not allow cords to contact hot surfaces, water or chemicals
- ✓ Do not place electrical cables where can be walk on or tripped over
- ✓ Never attach an exposed connector such as an alligator clip to a power supply
- ✓ No jewellery or other metal objects around electricity
- ✓ Don't work with electricity with wet hands or feet, or if the floor is wet.

5.3. Problems Needing Inspection orRepair

The device must be inspected by a qualified repair person in case of:

- ✓ Recurring problems with blowing fuses or circuit breakers,
- ✓ Discoloration of wall outlets,
- ✓ Burning smell or unusual odour coming form an appliance or wiring,
- ✓ Feeling a tingle or a shock when you touch an electrical appliance,
- ✓ Sizzling sound at wall switches or outlets,

Bibliography

[1] C. Biegelmeier, New knowledge of the impedance of the human body, in *Electric shock safety criteria*, J. Bridges, L. Ford, L. Sherman, M. Vainberg, eds, Pergamon Press, 1985.
[2] C.F. Dalziel, The threshold of perception currents, in: *IEEE Trans. Power Apparatus and Systems* . **73**(1954), 990-996.
[3] A.R. Dimick, Electrical Injuries, in *Harrison's Principles of Internal Medicine*, Anthony S. Fauci *et al*. Eds, McGraw-Hill, New York, 1997.
[4] R. Fish, Electric shock- Part I: Physics and pathophysiology, *J Emerg Med*., **11**(1993), 309-312.
[5] R. Fish, Electric shock, Part II: Nature and mechanisms of injury, *J Emerg Med*., 11(1993), 457-462.
[6] IEC Technical Specification, Electrical impedance of the human body, effects of sinusoidal alternating current in the range of 15 Hz to 100 Hz, effects of direct current, IEC publication IEC/TS 60479, Part 1, (1994-09).
[7] IEC Technical Specification, Effects of alternating current with frequencies above 100 Hz, Effects of special waveforms of current, Effects of unidirectional single impulse currents of short duration", IEC publication IEC/TR 60479, Part 2 (1987-03).
[8] IEC Technical Specification, Effects of lightning strokes on human beings and livestock, IEC publication IEC/TR 60479, Part 4 (2004-07).
[9] A.C. Koumbourlis, Electrical injuries, *Crit Care Med*., **30**(2002), 424-430.

[10] R.C. Lee, D. Zhang and J. Hannig, Biophysical injury mechanisms in electrical shock trauma, *Ann. Rev Biomed Eng*., **2**(2000), 477-509.
[11] D. Leibovici, J. Shemer, S.C. Shapira, Electrical injuries: current concepts, *Injury*, 26(1995), 623-627.
[12] J.P.Reilly, Scales of reaction to electric shock, in *Electrical Injury*, Annals of N.Y. Acad. Sci., 720(1994), 21-37.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-90

II.3. Electrograms (ECG, EEG, EMG, EOG)

Richard B. REILLY and T. Clive LEE
Trinity Centre for BioEngineering, Trinity College, Dublin 2, Ireland
Department of Anatomy, Royal College of Surgeons in Ireland, Dublin 2, Ireland

Abstract. There is a constant need in medicine to obtain objective measurements of physical and cognitive function as the basis for diagnosis and monitoring of health. The body can be considered as a chemical and electrical system supported by a mechanical structure. Measuring and quantifying such electrical activity provides a means for objective examination of heath status. The term electrogram, from the Greek *electro* meaning electricity and *gram* meaning write or record, is the broad definition given to the recording of electrical signal from the body. In order that comparisons of electrical activity can be made against normative data, certain methods and procedures have been defined for different electrograms. This paper reviews these methods and procedures for the more typical electrograms associated with some of the major organs in the body, providing a first point of reference for the reader.

Keywords. Biopotential, membrane potentials, electrograms, cardiogram, encephalogram, myogram, oculogram

1. Biopotentials

The term biopotential refers to the electric potential that is measured between points in living cells, tissues, and organisms and which accompanies all biochemical processes. Biopotentials from the organs of the body provide rich physiological and clinical information, playing often significant role in diagnosis.

Bioelectric potentials are produced as a result of electrochemical activity in a certain class of excitable cells. These cells are components of nervous, muscular, and glandular tissues. Electrically, these cells exhibit a resting potential and, when appropriately stimulated, an action potential. Their activity is necessary for information transfer (e.g. sensory information in the nervous system or coordination of blood pumping in the heart).

1.1. Mechanism Behind Biopotentials

Neurons are designed to respond to stimulation, which they do by generating electrical impulses. These impulses are expressed as changes in the electrical potentials conducted along the plasma membranes of the dendrites, cell body, and axon of each neuron. The difference in potential across the plasma membrane of the neuron results from differences in the concentration of certain ions on either side of the membrane. Cell potential is a function of membrane permeability and concentration gradient to

various molecules (i.e. K^+, Na^+, Cl^-, and Ca^{2+}). The cell membrane is a very thin (7-15 nm) lipoprotein complex that is essentially impermeable to intracellular protein and other organic anions (A^-). The membrane in the resting state is only slightly permeable to sodium Na^+ and rather freely permeable to potassium K^+ and chlorine Cl^-. The permeability of the resting membrane to potassium ions is approximately 50~100 times larger than its permeability to sodium. The diffusion and electrical forces acting across the membrane are opposed to each other, and a steady state is ultimately achieved. This steady membrane potential is called the equilibrium potential for potassium on account of the main ion potassium K^+ involved in the resting state. As a result an approximation of steady membrane potential can be calculated using the Nernst equation , Eq. (1). The Nernst equation provides a quantitative value for potential generated, E_K given the intra- and extracellular concentrations of potassium.

$$E_K = \frac{RT}{nF} ln \frac{[K]_o}{[K]_i} = 0.0615 log_{10} \frac{[K]_o}{[K]_i} \tag{1}$$

where n is the valence of potassium, $[K]_i$ and $[K]_o$ are the intra- and extracellular concentrations, R is the universal gas constant, T is the absolute temperature in Kelvin, F is the Faraday constant [1].

A more accurate expression for the membrane equilibrium potential can be calculated from the Goldman- Hodgkin-Katz equation, Equations 2 and 3, which provide a more accurate estimate of the potential E, by taking into consideration the intra- and extracellular concentrations of sodium and chlorine ions as well as potassium.

$$E = \frac{RT}{F} ln \left[\frac{P_K [K]_o + P_{NA} [Na]_o + P_{Cl} [Cl]_i}{P_K [K]_i + P_{NA} [Na]_i + P_{Cl} [Cl]_o} \right] \tag{2}$$

$$E = 0.0581 log_{10} \left[\frac{P_K [K]_o + P_{NA} [Na]_o + P_{Cl} [Cl]_i}{P_K [K]_i + P_{NA} [Na]_i + P_{Cl} [Cl]_o} \right] \tag{3}$$

Here P is the permeability coefficient of the given ion.

When membrane stimulation exceeds a threshold level of about 20 mV, an action potential occurs. Sodium and potassium ionic permeabilities of the membrane change. Sodium ion permeability increases very rapidly at first, allowing sodium ions to flow from outside to inside, making the inside more positive. The more slowly increasing potassium ion permeability allows potassium ions to flow from inside to outside, thus returning membrane potential to its resting value. At rest, the sodium and potassium (Na-K) pump restores the ion concentrations to their original values. The number of ions flowing through an open channel is greater than 106/sec.

The change in the electrical potential difference across a plasma membrane is the key factor in the initiation and subsequent conduction of a nerve impulse. A stimulus that is strong enough to initiate an impulse in a neuron is called a threshold stimulus. When such a stimulus is applied to a polarized resting membrane of an axon, sodium

ion channels into the cell open, and sodium ions rush in, reversing the electrical charge at the point of stimulus (In the l millisecond that a channel is open, about 20,000 sodium ions flow through). Thus, at the point of stimulus, the inside of the membrane becomes positively charged relative to the outside, a condition known as depolarization. When a stimulus is strong enough to cause depolarization, the neuron is said to fire.

Once a small area on the neuron is depolarized, it stimulates the adjacent area, and an action potential, or nerve impulse, is initiated and conducted along the plasma membrane.

Shortly after depolarization, tbe original balance of sodium and potassium ions is restored by the action of the membrane pumps; the relative electrical charges inside and outside the membrane are also restored. The membrane is then said to be repolarized. The transmission of a nerve impulse along the plasma membrane may be visualized as a wave of depolarization and repolarization.

After each firing, there is an interval of from 0.5 to 1 millisecond before it is possible for an adequate stimulus to generate another action potential. This period is called the *refractory period.* Most nerve fibers are capable of generating about 300 impulses per second

In the resting state there exists a steady electrical potential difference between internal and external environments—typically between -70 to -90mV, relative to the external medium.

In the active state an electrical response is instigated by adequate stimulation. It consists of an “all-or-none” action potential after the cell threshold potential has been reached.

The body is an inhomogeneous volume conductor and these ion fluxes create measurable potentials on the body surface. By recording these biopotentials, a process known as electrography, an analysis of the underlying physiological can be carried out providing information for clinical diagnosis.

Biopotentials have extremely low voltages values, typically in the order of micro or millivolts. Amplification of these biopotentials is required prior to their acquisition by computer and subsequent analysis.

2. Bioamplifers

A bioamplifier is an electronic instrument used to acquire and increase the signal amplitude of biopotential electrical activity for output to various sources. The design of biopotential amplifiers requires great technical expertise, as besides a linear amplification of biopotential amplitude to improve the signal-to-noise ratio, the device must maintain signal fidelity in terms of time resolution. The design parameters of biopotential amplifier are specific to the biopotential under investigation. A more detailed review of bioamplifers in provided in the accompanying related paper on Biosensors.

There are many electrograms, however the most commonly used in the clinic are the electrocardiogram (ECG), electroencephalogram (EEG), electrooculogram (E0G), electromyogram (EMG), and electroretinogram (ERG).

3. Electrocardiogram (ECG)

An electrocardiogram is the most familiar electrogram and is the graphical output of an electrocardiograph, which is the temporal variation of electrical activity of the heart. The term cardio derives from the Greek for heart and the word electrocardiogram is often abbreviated to ECG or EKG. The ECG possesses great information content, due to the ease at which can it can measured in different circumstances and it is pivotal in providing clinical diagnoses. By analyzing the perturbations in normal electrical activity, the ECG can provide critical information on an evolving myocardial infarction, different cardiac arrhythmias, the effects of hypertension, as well as important information for cardiac rehabilitation and exercise.

The first practical electrocardiogram was recorded by the Dutch doctor and physiologist Willem Einthoven in 1903. He was awarded the Nobel Prize in Physiology or Medicine in 1924 “for his discovery of the mechanism of the electrocardiogram”.

3.1. Basis of the ECG

The heart consists of four chambers: the left and right atria and also the left and right ventricles. Blood enters the right atrium and passes through the right ventricle, which in turn pumps the blood to the lungs where it becomes oxygenated. The oxygenated blood is brought back to the heart by the pulmonary veins which enter the left atrium, flowing then into the left ventricle. The left ventricle pumps the blood to the aorta which will distribute the oxygenated blood to all parts of the body.

There are three groups of cells in the heart. Pacemaker cells, Electrical Conducting Cells and Myocardial Cells [2]. In the resting state these cells are electrically polarised, but it is their depolarisation that is the fundamental electrical event of the heart. Depolarisation is propagated from cell to cell, producing a wave of depolarisation that is transmitted across the heart. This wave sets up a biopotenital across the heart, resulting in the flow of electrical current [3]. This depolarisation can be detected by a number of electrodes attached at specific points on the body. Once the depolarisation has occured, repolarisation commences to restore the cardiac cells to their resting state. This repolarisation phase can also be recorded with the same electrodes. The ECG is a recording of the depolarisation and repolarisation of the myocardial cells of heart, which makeup the majority of heart cells and represent the heart’s contractile machinery. The ECG indicates the overall rhythm of the heart and weaknesses in different parts of the heart muscle [4].

Since the human body can be regarded as a volume conductor, changes in potential are transmitted throughout the body. Therefore to be able to record myocardial activity, one needs to be able to detect small changes in potential on the body surface. These waves can be measured at electrodes attached to the skin. An ECG displays the voltage between pairs of these electrodes and the muscle activity in different locations.

In order to record the ECG, each lead pair is connected to a bioamplifier. Bipolar leads measure the difference in electrical potential between any two points on the body, unipolar leads measure the potential at a single point on the body and a virtual reference point, with zero electrical potential, located in the centre of the heart.

A lead records the electrical signals of the heart from a particular combination of recording electrodes that are placed at specific points on the patient's body. In various cases, the detected signals have the following polarities (see Figure 1):

- When a depolarization front propagates toward a positive electrode, it creates a positive deflection on the ECG in the corresponding lead.
- When a depolarization wavefront moves away from a positive electrode, it creates a negative deflection on the ECG in the corresponding lead.
- When a depolarization wavefront moves perpendicular to a positive electrode, it creates an biphasic complex on the ECG. This will be positive as the depolarization wavefront approaches (A), and then become negative as it passes by (B).

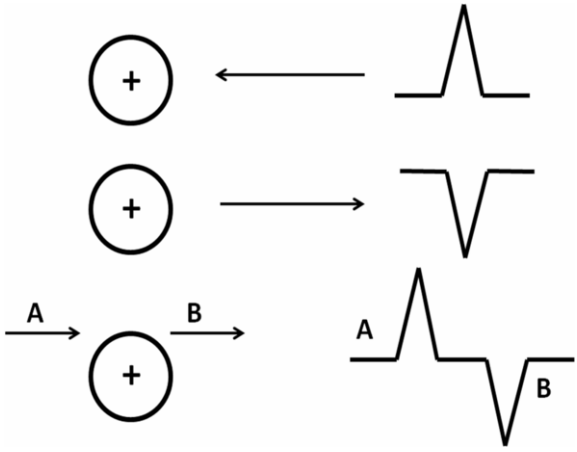


Figure 1. Graphic showing the relationship between positive electrodes, depolarization wavefronts (or mean electrical vectors), and complexes displayed on the ECG.

3.2. Measuring ECG

3.2.1. ECG Leads

In 1908 Willem Einthoven published a description of the first clinically important ECG measuring system. The Einthoven triangle is an imaginary equilateral triangle having the heart at the centre and formed by lines that represent the three standard limb leads of the electrocardiogram (Figure 2).

3.2.2. Standard Limb Leads

The most basic arrangement of ECG leads is the three limb lead configuration, which are bipolar leads measuring the potential difference between the right arm, left arm, and left leg. They are positioned as follows:

Lead I – Difference between the left arm and right arm, the left arm being positive

Lead II – Difference between the left leg and right arm, the left leg being positive.

Lead III – Difference between the left leg and left arm, the left leg again being positive.

A fourth electrode is placed on the right leg to use as reference or “electric ground”.

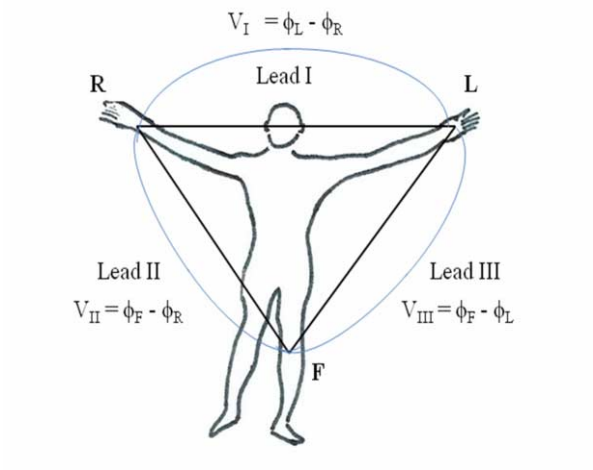


Figure 2. Einthoven lead system

The Einthoven limb or Standard leads are defined in the following way:

Lead I: $V_I = \Phi_L - \Phi_R$

Lead II: $V_{II} = \Phi_F - \Phi_R$

Lead III: $V_{III} = \Phi_F - \Phi_L$

where

V_I = the voltage of Lead I, V_{II} = the voltage of Lead II,

V_{III} = the voltage of Lead III

Φ_L = potential at the left arm (LA)

Φ_R = potential at the right arm (RA)

Φ_F = potential at the left leg (LL)

According to Kirchhoff's law, these lead voltages have the following relationship:

$$V_I + V_{III} = V_{II} \tag{4}$$

Hence only two of these three leads are independent.

In clinical practice, ECG is recorded using extra leads in addition to the standard three. This is necessary due to the three dimenstional nature of the heart. The addtional leads provide important information on the heart's electrical activity in three orthogonal directions.

3.2.3. Augmented Limb Leads

The same three leads that form the standard leads also form the three unipolar leads known as augmented leads. These are referred to as "augmented" because they are unipolar. These three leads are referred to as aVR (right arm), aVL (left arm) and aVF (left leg) and also record a change in electric potential in the frontal surface of the chest.

3.2.4. Precordial Leads

For measuring the potentials close to the heart, Wilson introduced the *precordial leads* (chest leads) in 1944. These leads (V_1 to V_6) are located over the left chest as shown in Figure 3.

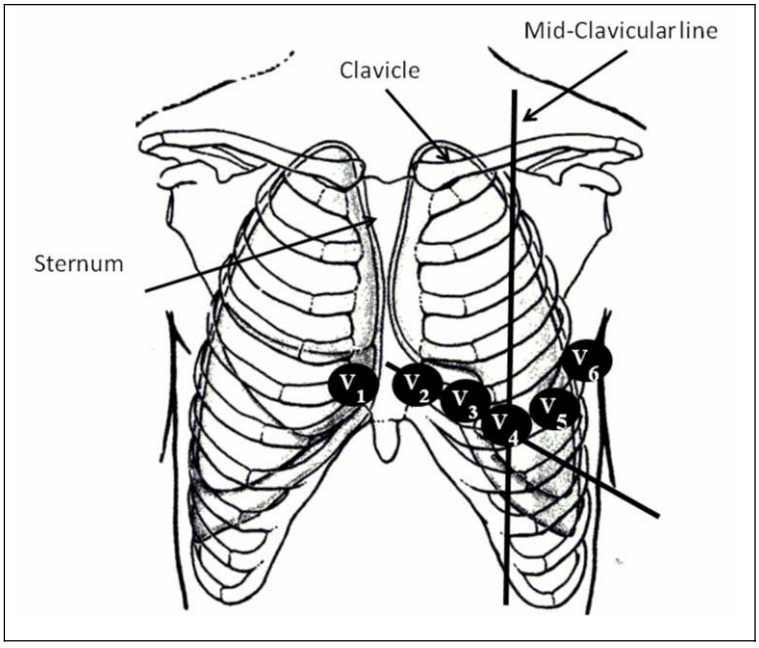


Figure 3. Precordial leads.
Placement of Precordial Leads
V_1 – 4th intercostal space, just to the right of the sternum
V_2 – 4th intercostal space, just to the left of the sternum
V_3 – Halfway between V_2 and V_4
V_4 – 5th intercostal space in the mid-clavicular line
V_5 – Halfway between V_4 and V_6
V_6 – 5th intercostal space in the mid-axillary line

3.2.5. The 12-Lead System

The most commonly used clinical ECG-system, the 12-lead ECG system, consists of the following 12 leads:

- I, II, III
- aV_R, aV_L, aV_F
- V_1, V_2, V_3, V_4, V_5, V_6

Of these 12 leads, the first six are derived from the same three measurement points. Therefore, any two of these six leads include exactly the same information as the other four.

3.3. Formation of the ECG Waveform

In order to interpret the ECG, one needs to understand the electrical events that underpin the cardiac cycle of contraction (systole) and relaxation (diastole) [2]. The dominant pacemaker cells in the heart, are known as the Sinoatrial Node. These are located at the top right hand side of the heart, in the right atrium. These cells fire spontaneously at a rate of 60 to 100 times per minute, and a wave of depolarization begins to spread outward into the atrial myocardium cells. It is the depolarization of atrial myocardium cells that results in atrial contraction. Electrodes on the surface of the body will record this contraction as a burst of electrical activity lasting a fraction of a second. This is called the P wave. The atria begin contracting around 100 ms after the start of the P wave. Once the atrial depolarisation is complete the ECG returns to baseline.

Following this the ventriclar myocardial cells depolarise, causing a ventriclar contraction. Electrodes on the surface of the body will record this contraction as a large increase in amplitude. This is known as the QRS complex. The amplitude of the QRS complex is larger than the P wave as the mass of the ventricular muscle is much larger than that of the atria. The ventricles begin contracting shortly after the R wave. The smaller T wave indicates ventricular repolarization. The atria repolarize during the QRS complex and therefore this repolarization cannot be observed separately.

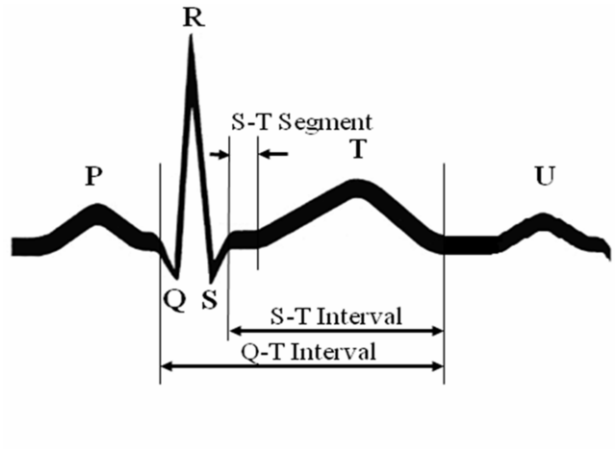


Figure 4. Heart excitation related to electrocardiogram (ECG)

Therefore a typical ECG tracing of a normal heart cycle consists of a P wave, a QRS complex and a T wave. A small *U wave*, caused by repolaristaion of Purkinje fibers the nerve cells of the heart, is normally visible in 50 to 75% of recorded ECGs. The baseline voltage of the electrocardiogram is known as the isoelectric line. Typically the isoelectric line is measured as the portion of the tracing following the T wave and preceding the next P wave. Figure 4 shows the important features of a sample ECG.

3.3.1. Electrical heart axis

When the heart depolarizes and repolarizes, it is convenient to represent the electrical activity as an electric dipole or mathematically as a vector. A vector has both magnitude and an angle of orientation. The magnitude of the heart vector represents a maximum amplitude of the recorded electrical activity, while the angle of orientation represents the average direction of the current flow in the heart. The magnitude and direction of this heart vector can be beneficial in diagnosis.

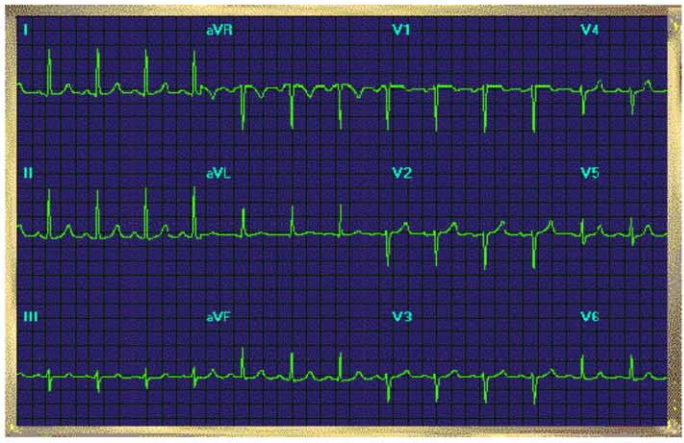
The augmented and standard leads allow analysis of the heart on what is known as the frontal plane. Therefore, the heart vector refers to the direction of the electrical depolarization obtained from the sum of all different vectors in the frontal plane. It is usually oriented in a right shoulder to left leg direction. The normal angle of orientation of the QRS vector generally lies between 0 and 90 degrees. A change in the angle of orientation provides diagnostic information. For example, a left deviation (-30^{o} to -90^{o}) may indicate left anterior fascicular block or Q waves from inferior myocardial infarction (MI). Right axis deviation ($+90^{o}$ to $+180^{o}$) may indicate left posterior fascicular block, Q waves from high lateral myocardial infarction, or a right ventricular strain pattern.

3.3.2. ECG – Clinical significance

ECG is the standard method for the diagnosis of cardiac arrhythmias (Figure 5). Pathology that can be identified and studied using ECGs includes rhythm disturbances, ischaemia and infarction, chamber enlargements, electrolyte abnormalities, and drug toxicities.

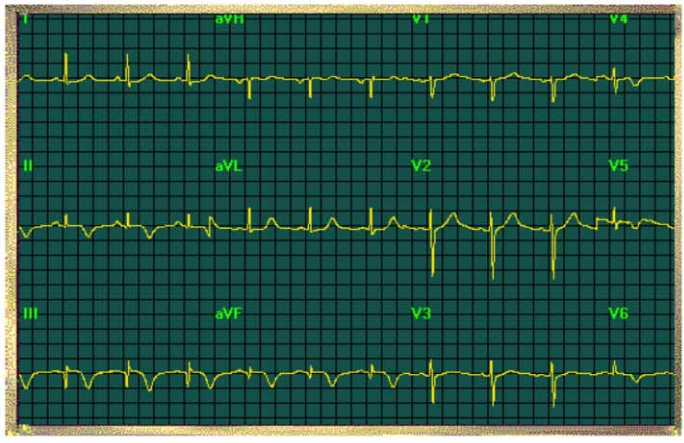
Some abnormalities that may indicate illness:

- An extended P-R interval may be diagnosed as AV node block
- Widening of the QRS complex conduction problems in the bundle of His
- Elevated ST segment may indicate occurrence of MI
- Negative polarity T wave may be due to coronary insufficiency. QRS amplitude, polarity, time domain, PR interval (indicator of heat beat per min. & T-wave amplitude are some very important distinctive features.



Normal ECG with horizontal axis

* Upright P and T waves in lead I
* R > 0.20 sec
* Duration of Q > 0.02 sec in leads I and aVL
* Frontal plane QRS vector 0-90 degrees
* Precordial progression from R in V1 to R in V5&V6
* QTc > 0.44 sec



Inferior myocardial infarction

* Q waves longer than 0.04 sec in duration in leads II, III, and aVF
* ST segment elevation in leads II, III, and aVF in an acute infarction
* T wave inversion in leads II, III, and aVF in an old or evolving infarction
* Common presence of a lateral wall component in inferior wall myocardial infarction (T wave changes in leads V4-V6)

Figure 5. Clinical significance of ECG (Source: www.bioscience.org)

4. Electroencephalography

Hans Berger was the first to standardize recordings of the electrical activity in the brain. In a series of experiments in 1920s, he pioneered electroencephalography (EEG), by placing electrodes on the scalp and measuring electrical activity during different tasks. The name is derived from the Greek *encephalo* meaning brain.

The conventional EEG is recorded from scalp electrodes, and shows cortical electrical activity. This includes cortical manifestations of the sub cortical regions (projection pathways, thalamus, reticular formation, mesencephalon or midbrain) [5].

As the EEG reflects the functional status of the brain, it can be used to monitor its functional integrity and thus assist clinicians in the diagnosis of a variety of neurological problems. It is very important to relate scalp potentials to the underlying neurophysiology. The pathological states most commonly diagnosed using EEG include, common headaches and dizziness, seizure disorders, stroke, brain tumours, epilepsy, multiple sclerosis, sleep disorders and movement disorders [5].

A method similar to the EEG is intracranial EEG (icEEG), also referred to as subdural EEG (sdEEG) and electrocorticography (ECoG). The electrodes are placed directly on the exposed surface of the brain to record electrical activity from the cerebral cortex. ECoG is an invasive procedure in which electrodes are placed over a specific brain area. In a very small selected group of patients being considered for epilepsy surgery, ECoG may be recorded over several days to map the distribution and spread of seizure activity.

EEG records the spontaneous activity in the brain. The amplitude the signal record on the scalp is typically 100 μV, while that of ECoG from the surface of the cortex it is in the order of 1-2 mV.

4.1. Electrode Montages

There is a standard system for electrode placement, known as the 10-20 international system which includes 64 electrodes. In this system the electrodes are located on the surface of the scalp, as shown in Figure 6. The electrode positions are determined as follows: Reference points are *nasion*, which is the delve at the top of the nose, level with the eyes; and *inion*, which is the bony lump at the base of the skull on the midline at the back of the head, the external occipital protrusion. The first mark is placed 10% of the distance along this line and others are arranged at 20% intervals. With this electrode placement system, activity at specific brain locations can be recorded with specific electrodes and thus can easily be compared across individuals [5].

When carrying out EEG experimentation, it is typical to record specific sensory processing areas of the cortex: visual, auditory or somatosensory. For example, visual processing is carried out in the occipital regions at the back of the brain. Activity at electrodes O_1 and O_2 would be more active during visual processing and these electrodes would be termed active electrodes while the brain is engaged in processing visual information.

64-channel EEG systems are typical, but some high-density systems exist, with 128 and 256 electrodes. Placement is based again on the 10-20 system, with the new electrodes assigned to locations in between the 32 electrodes of the 10-20 system.

In order to obtain good signal quality, impedance between the scalp and each electrode should ideally be kept below 5kOhms. If the electrode impedance increases, background noise and movement artifacts may obscure the EEG signal. As a

consequence, the signal-to-noise ratio decreases and the ability to extract diagnostic information may be limited.

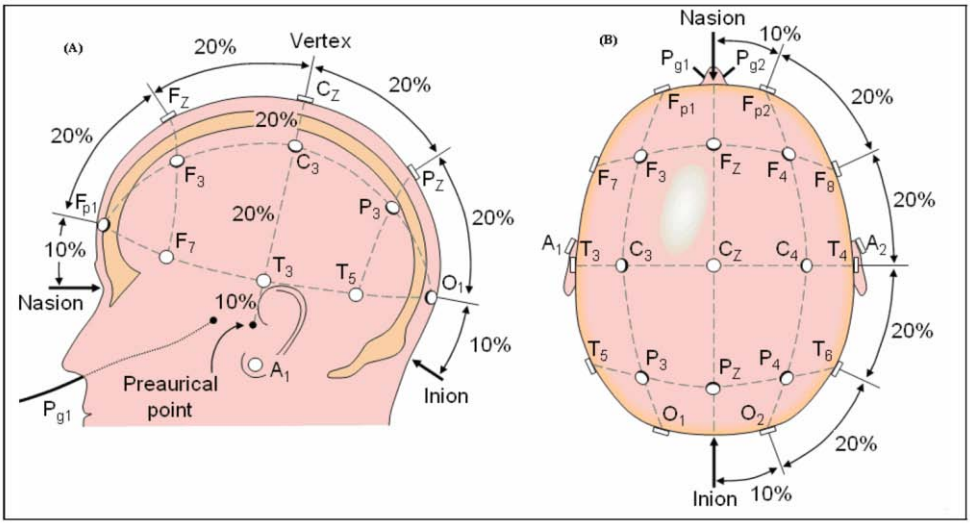


Figure 6. The international 10-20 system seen from (A) left and (B) above the head. A = Ear lobe, C = central, Pg = nasopharyngeal, P = parietal, F = frontal, Fp = frontal polar, O = occipital.

Electrode placements and the different patterns of combining electrode pairs to measure potential differences on the scalp constitute the electrode montage. There are two basic types of EEG montage: Referential and Bipolar. In the referential montage the potential difference is measured between an active electrode and an inactive reference electrode. While, with the bipolar montage, the potential difference is measured between two active electrodes.

4.2. Basis of the EEG

The EEG signal recorded on the scalp or on the cortex is generated by the summation of the synchronous activity of thousands of neurons that have similar spatial orientation and placement radial to the scalp. An action potential in a pre-synaptic axon causes the release of neurotransmitter into the synapse. The neurotransmitter diffuses across the synaptic cleft and binds to receptors in a post-synaptic dendrite. The activity of many types of receptors results in a flow of ions into or out of the dendrite. This results in currents in the extracellular space. It is these extracellular currents which are responsible for the generation of EEG potentials.

4.3. Processing of EEG

The character of the EEG signal is highly dependent on the degree of the activity of the cerebral cortex. EEG observed during states of wakefulness and sleep are remarkably

different. EEG is typically described in terms of patterns of rhythmic activity and transients. The rhythmic activity is divided into bands by frequency. Most of the cerebral signal observed in the scalp EEG falls in the range of 1-40 Hz. EEG frequency ranges are categorized as Delta (σ), Theta (τ), Alpha (α) and Beta (β) and Gamma (γ) [6].

Alpha rhythm is usually observed between 8 and 12Hz. Alpha is usually best observed in the posterior regions of the head on each side. It tends to dominate during eyes closed and by relaxation, and tend to diminish with eye opening or cognition processing (thinking, calculating). It is the major rhythm seen in normal relaxed adults. It is present during most of life especially beyond the age of thirteen.

Beta activity is a higher frequency rhythm than alpha, with a frequency between 14-20 Hz and is classed a fast activity. It is usually most evident frontally. It is accentuated by sedative-hypnotic drugs especially the benzodiazepines and the barbiturates. It may be absent or reduced in areas of cortical damage. It is generally regarded as a normal rhythm. It is the dominant rhythm in patients who are alert or anxious or who have their eyes open.

Theta activity has a frequency between 3.5 to 7.5 Hz and is classed as slow activity. It is abnormal in awake adults but is perfectly normal in children up to 13 years of age and also in sleep.

Delta activity is 3 Hz or below. It tends to be the greatest in amplitude of all EEG rhythms. It is the dominant rhythm in infants up to one year and in stages 3 and 4 of sleep. It is usually most prominent frontally in adults and posteriorly in children.

Gamma activity is associated with 40 Hz and up to 70Hz. Activity in this band can often be from 24Hz. Gamma activity is thought to be associated with perception and consciousness. Many encephalographers do not distinguish gamma waves as a distinct class but include them in beta brain waves.

Since the interpretation of EEG is time consuming and requires much experience, different processing methods have been developed to facilitate its interpretation. Typically, Fourier Analysis is applied to recorded EEG. Fourier analysis provides an estimate of the frequency content of a signal and results in the power spectrum of that signal. In the case of EEG, the spectrum of EEG can be divided into the specific spectral bands: delta, theta, alpha, beta and gamma. In terms of diagnostic information, of great interest is the ratio of power of one frequency band to another during a specific brain state, for example the ratio of alpha power to beta power. Another quantity of diagnostic interest is the change over time in power of a frequency band during a specific brain state.

4.4. Evoked Potential

The EEG reflects the functional status of the brain. The strict definition is continuous EEG, as it represents ongoing brain activity. Another useful electrical signal that can be recorded is the evoked or event related potential. Evoked potentials arise in response to a stimulus (auditory, visual or somatosensory, etc.). They provide great diagnostic information on sensory processing. They are much smaller in amplitude than continuous EEG and are not readily distinguished. Therefore, a train of stimuli is often applied to the individual and signal averaging method used to improve the signal-to-noise ratio of the recorded signals to allow interpretation [5],[6].

4.4.1. Visual Evoked Potentials

These are tests of the pathway between the eye and the back of the brain. The individual being examined fixates on the centre of a computer monitor that has a moving black and white checkerboard pattern. Each eye is tested separately, with a patch covering the eye not being investigated. The recorded visual evoked potential (VEP) is achieved by averaging over a number of responses to checkerboard pattern and are of the form shown in Figure 7.

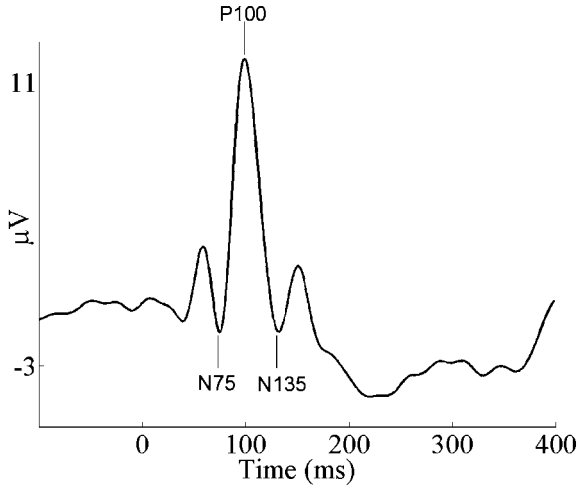


Figure 7. A normal pattern reversal VEP

Measurements are made of the latencies and amplitudes of particular peaks. In the normal VEP, a very distinguishable peak is found approximately 100ms after stimulation, known as the P100. There are also two clear troughs called the N75 and N135, around 75ms and 135ms after stimulation. The latencies and amplitudes of these peaks provide objective diagnostic information on the brain's ability to process visual information. Individual data can be with compared to normative data.

4.4.2. Auditory Evoked Potentials

Audio evoked potentials (AEP) reflect the function of the auditory pathway. AEP can be evoked by repeated clicks of short duration (100-500 ms) given into an ear piece. Trigger-synchronized averaging of a defined number of EEG-segments is used to extract the AEP by reduction of the underlying EEG signal (background noise). The reduction in background noise is proportional to the square root of the number of averaged segments; the more averaged segments, the better the quality of the AEP. The extracted AEP signal consists of a number of peaks and troughs. As with VEP analysis measurements are made of latencies and amplitudes of particular peaks. Three main groups of peaks can be distinguished and they can be correlated to the anatomical structures:

- *Brainstem AEP* (BAEP) with latencies shorter than 10 milliseconds. Anatomical structures: cochlea, acoustic nerve (BAEP wave I, II), brainstem (BAEP wave III-V)
- *Middle latency AEP* (MLAEP) with latencies of 10-50 milliseconds. Anatomical structures: medial geniculate body and primary auditory cortex (temporal lobe).
- *Late cortical waves* with latencies over 50 milliseconds. Anatomical structures: frontal cortex, association fields.

4.5. Interpretation of EEG

Changes in EEG are not necessarily specific to underlying mechanisms [5]. For example, slowing of EEG may reflect either changes in anesthetic concentrations or cerebral ischemia. EEG has poor spatial resolution, meaning that it not possible to know exactly the source of electrical activity from the recordings made on the scalp. Mathematically one can consider the electrical activity being generated by an electric dipole. Considerable research attention is focused on providing estimates of the location of electric dipoles that would generate the potential similar to those recorded. The magnitude and orientation of these dipoles can be used as a basis for diagnosis in neurological conditions.

Despite these disadvantages, EEG has considerable advantages over neuroimaging. Its poor spatial resolution is offset by excellent temporal resolution, which is in milliseconds. The ease of recording EEG makes it of great clinical importance.

5. Electromyography (EMG)

Electromyography (EMG) is the study of the mechanical properties of muscles at rest and in contraction. The electrical activity associated with muscle is recorded using two groups of electrodes over the muscle site. These can be indwelling (intramuscular) electrodes or can be surface electrodes for non-invasive recordings [7].

EMG is used to diagnose two general categories of disease: neuropathies and myopathies. EMG may aid with the diagnosis of nerve compression or injury (such as carpal tunnel syndrome), nerve root injury (such as sciatica), and with other problems of the muscles or nerves. Less common medical conditions include amyotrophic lateral sclerosis, myasthenia gravis, and muscular dystrophy.

5.1. Basis of the EMG

Muscle consists of motor units, which is areas consisting of a motor neuron and the muscle fibres it innervates. Each motor unit is controlled by the motor neurons. The motor end plate is the synaptic junction between the motor neuron and the controlled motor unit. Depolarization of the post synaptic membrane arises in case of activation of the motor unit. End plate potential (EPP) is the potential that is recorded. The depolarization wave moves along the direction of the muscle fibers**.** The signal between the EMG electrodes corresponds to the depolarization wave front and to the subsequent repolarization wave.

The electrical source is the muscle membrane potential of about -70mV. Due to movement of the muscle, the measured potentials range between less than 50 μV and 20 to 30 mV. EMG signals are made up of superimposed motor unit action potentials (MUAPs) from several motor units. For a thorough analysis, the measured EMG signals can be decomposed into their constituent MUAPs.

5.2. Location and orientation of the electrode

The electrode should be placed between a motor point and the tendon insertion or between two motor points, and along the longitudinal midline of the muscle. The longitudinal axis of the electrode should be aligned parallel to the length of the muscle fibers. The reference electrode also called the ground electrode should be placed as far away as possible and on electrically neutral tissue [7].

The voltage waveform that is recorded is the difference in potential between the two electrodes. The electrical signal generated in the muscle fibers is called a muscle action potential (MAP). As mentioned above, the surface electrodes or indwelling electrodes record the algebraic sum of all MAPs that are being transmitted along the muscle fibers.

6. Electrooculography (EOG)

Electrooculography (EOG) is the study of retina function by recording changes in steady, resting electric potentials of the eye. The eye can be considered to be a fixed electric dipole with a positive pole at the cornea and a negative pole at the retina. The electric potential is not generated by excitable tissue but is attributed to the higher metabolic rate in the retina. The magnitude of the corneoretinal potential is in the range of 0.4-1.0 mV.

The electric potential is independent to light stimulus. The field may be detected with the eye in total darkness and/or with the eyes closed. This potential difference and the rotation of the eye form the basis for a signal measured at a pair of periorbital surface electrodes. Electrodes are placed near the canthi, the corner of the eye where the upper and lower eyelids meet The resulting signal is called the electrooculogram. If the eye travels from the center position in the direction of one electrode, this electrode perceives the positive side of the retina while the opposite electrode views the negative side of the retina. As a result, a potential difference occurs between the electrodes. If the resting potential is constant, the recorded potential is a measure for the eye position (Figure 8).

The EOG has clinical importance in the study of eye movement. The small movements of the eye, termed nystagmus, are measured with the help of EOG. The resulting signal is called an electronystagmogram.

The retina is the site of cells that are sensitive to the incident light energy; as with other peripheral nerve cells, they generate receptor potentials. The collective behavior of the entire retina is a bioelectric generator, which sets up a field in the surrounding volume conductor. This potential field is normally measured between an electrode on the cornea (contact-lens type) and a reference electrode on the forehead. The recorded signal is known as the electroretinogram (ERG).

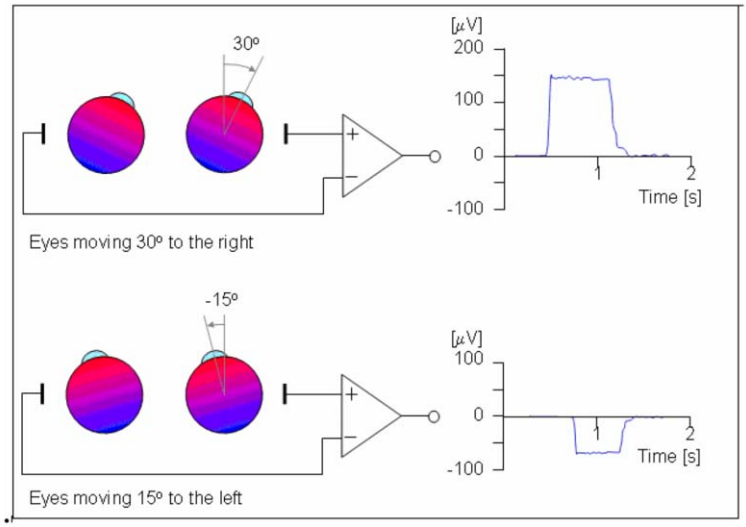


Figure 8. An illustration of the electrooculogram (EOG) signal generated by horizontal movement of the eyes. Two electrodes are employed and each electrode connected to the input of a bioamplider. The polarity of the signal is positive at the electrode to which the eye is moving

EOG is a sensitive electrical test for detection of retinal pigment epithelium dysfunction. The main applications are in ophthalmological diagnosis and in recording eye movements.

7. Conclusion

Electrical activity associated with the brain, heart, muscle and eye can be recorded and analysed. Comparisons against normative data provide important clinical information. It is important to relate the electrical activity to the underlying physiology so that this information can be used for diagnosis.

References

[1] Deutsch S. and Deutsch A. (1993), *Understanding the Nervous System-An Engineering Perspective,* IEEE Press, New York.
[2] Katz A.M. (1986), *Physiology of the Heart*, Raven Press, New York
[3] DiFrancesco D. and Noble D. (1985) "A model of cardiac electrical activity incorporating ionic pumps and concentration changes", *Phil Trans R. Soc, London* [B] 307, 307-353
[4] Geselowitz D.B. (1989) "On the theory of the electrocardiogram", *Proc IEEE* 77, 857.
[5] Nunez P.L., Srinivasan R. (2005) *Electric Fields of the Brain: The Neurophysics of EEG*, Oxford University Press, USA.
[6] Buzsaki G. (2006), *Rhythms of the Brain*, Oxford University Press, USA.
[7] Katirji B. (2007) *Electromyography in Clinical Practice: A Case Study Approach*, 2nd Edition, Mosby.

Glossary of Terms

Action potential: An action potential is initiated by a stimulus above a certain intensity or threshold.
Atrium: There are four chambers in the human heart, two atria and two ventricles. The right atrium receives de-oxygenated blood from the superior vena cava, inferior vena cava and coronary sinus. The left atrium receives oxygenated blood from the left and right pulmonary veins.
Artery: A vessel that carries blood away from the heart to the farthest reaches of the body
Atrioventricular node: An electrical relay station between the atria (the upper) and the ventricles (the lower chambers of the heart).
Aorta: The large artery that receives blood from the left ventricle of the heart and distributes it to the body.
Biopotentials: electric potential that is measured between points in living cells, tissues, and organisms as a result of electrochemical activity of excitable cells.
Bioamplifier: Device used to gather and increase the signal integrity of human neurophysiologic electrical activity for output to various sources.
Bipolar lead: Registration of the potential difference between two electrodes.
Diastole: Phase of cardiac relaxation
Einthoven triangle: An imaginary equilateral triangle having the heart at its center and formed by lines that represent the three standard limb leads of the electrocardiogram
Electrical heart axis: The direction of the electrical depolarization obtained from the sum of all different vectors in the frontal plane.
Electrocardigraphy (ECG): Electrical recording of the heart and is used in the investigation of heart disease.
Electrocorticography (ECoG): Graphical recording of electrical activity in the brain by placing electrodes in direct contact with the cerebral cortex
Electroencephalograph (EEG): Graphical recording of electrical activity in the brain recorded by scalp electrodes.
Electromyography: Graphical recording of the electrical activity of muscle
Electrooculography: Graphical recording technique for measuring the resting potential of the retina.
Electroretinography: Graphical recording electrical responses of various cell types in the retina.
Elevation: The rise of the ST-segment above the iso-electrical line.
Electrode: A metallic part which is used for transmitting the electrical activity from the body to the input circuit of a bioamplifier.
Electrolyte: It is a conducting solution (gel or paste) or may be fluid of living tissue as when electrode inserted below skin.
Excitable cells*:* Those that can be stimulated to create a tiny electric current.
Frequency: The number of beats per minute
Frontal plane leads: The frontal plane leads are the leads I, II, III, aVR, aVL and aVF.
His, bundle of: Fast conducting bundle which runs from the AV-node to the cells of the ventricles.
Infarction: The formation of an infarct, an area of tissue death due to a local lack of oxygen.
Interval: A specific distance on the ECG.

Ischemia: Shortage of oxygen in a tissue, caused by insufficient blood flow towards that tissue.
Isoelectric line: The baseline voltage of the electrocardiogram
Lead: Electrode for registering the electrical potentials.
Membrane potential: Potential difference across the cell membrane
Myocardial infarction (MI): Medical condition that occurs when the blood supply to a part of the heart is interrupted.
Precordial leads: Leads placed on the chest to record the ECG.
P-wave: The depolarization of both atria.
Physiology: The study of how living organisms function including such processes as nutrition, movement, and reproduction.
Repolarization: Recovery of the resting potential.
Rhythm strip or an ECG strip: The printed record of the electrical activity of the heart is called a rhythm strip or an ECG strip
Sinus node: Primary pacemaker of the heart located in the right atrium - sinoatrial (SA) node.
Systolic: Phase of cardiac contraction
Transmembrane potential: Resting potential across the cell membrane.
T-wave: The repolarization of both ventricles.
Unipolar lead: Lead in which the potential differences are registered in one point compared to the central terminal.
U-wave: Repolarization of the Purkinje fibbers
Vector: Electrical force in a specific direction and with a specific magnitude.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-109

II.4. Biosensors

Richard B. REILLY and T. Clive LEE*
Trinity Centre for BioEngineering, Trinity College, Dublin 2, Ireland
**Department of Anatomy, Royal College of Surgeons in Ireland, Dublin 2, Ireland*

Introduction

The ability to accurately measure physiological and chemical changes in the body is fundamental to the development of new diagnostic and therapeutic methods. In order to carry out such measurements, specific tools are necessary to detect physiological and chemical changes and to transform these changes into a form that can easily be analyzed and processed. A biosensor is one such tool and can be defined as a*n analytical device which converts a biological response into an electrical signal*. The name signifies that the device is a combination of two components, a biological element and a sensor or transducer element.

The biological element may be tissue, living cells, an enzyme, antibody or antigen. The sensor/transducer element includes electric current, electric potential, intensity and phase of electromagnetic radiation, mass, conductance, impedance, temperature and viscosity.

This paper will outline the components and characteristics of a biosensor, describe some of the typical transducers found in biomedical engineering and biomedical instrumentation amplifiers. A focus is placed on sensing within electrophysiology. How transducers are used in medicine to measure key biophysical characteristics in the clinic will form the concluding section.

1. Components of a Biosensor

The basic components of a biosensor include a biological element and the physiochemical transducer, Figure 1. The output of the transducer is passed to the detector for further processing and analysis.

The biological element is the target system under investigation, such as tissue, microorganisms, cell receptors, enzymes, antibodies, nucleic acids or biologically derived material. The transducer transforms the signal from the sensor into a more easily measurable and quantifiable signal. Typically, a computer or microprocessor device acquires (more formally termed samples), amplifies and processes the signal from the transducer. Following signal processing, the data are often converted to other units and transferred to a display or/and data storage device.

The electrical signal from the transducer is often low and may be superimposed upon a relatively high and noisy baseline. One of the more fundamental signal

processing steps involves subtracting a 'reference' baseline signal, derived from a similar transducer without being coupled to the sensor, from the sampled signal.

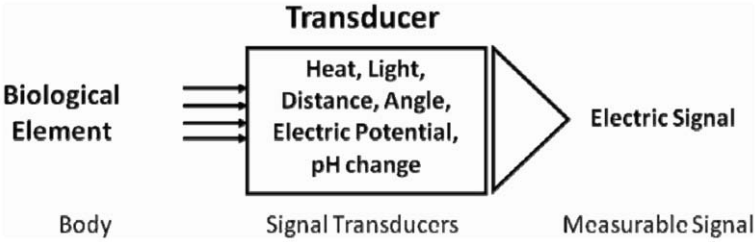


Figure1. Components of a biosensor

1.1. General Biosensor Characteristics

An ideal biosensor must fulfill a number of criteria to measure a physical quantity [1]. The biosensor must be able to provide a signal that is directly proportional to the quantity under investigation, providing all the useful information about the physical event. The biosensor must also meet requirements connected with measurement such as repeatability, sensitivity, reproducibility, selectivity, specificity, good response times and linearity of response. These are defined as follow:

Specificity/selectivity: Ability to recognize a single event/condition among other events/conditions in the same sample/signal. The selectivity of the biosensor is determined by both the sensor and transduction method.

Sensitivity: The sensitivity of a sensor is measured by the change in its response as a function of the analogous change in the quantity examined. The sensitivity s of a biosensor is defined by the ratio

$$s = \frac{\Delta output}{\Delta input} \tag{1}$$

where
$\Delta output$ = change in the amplitude of output and
$\Delta input$ = change in the amplitude of input.
The sensitivity determines the aptness of the sensor for a particular application.

Repeatability: The results are deemed to have high repeatability if two sets of results, obtained using the same sensor and the same sample, are numerically close to each other.

Reproducibility: The method is reproducible if similar results are obtained by other scientists at different laboratories.

Linear response: The linear region of a biosensor is obtained from a calibration curve of its response to different inputs. A good calibration curve is also indicative of stability of the biosensor's response, which should neither drift nor oscillate with time.

A biosensor having a linear response can be more easily described mathematically, than one which has a nonlinear response.

Response time: Response time is defined as the time taken to reach a steady state output from the instant of variation in the input value to the sensor.

1.2. Transducer Principles

The transducer is responsible for converting a biological signal to a measurable electrical signal. The transducer is a device, usually electrical, electronic, electro-mechanical, electromagnetic, photonic, or photovoltaic that converts one form of energy to another. The transducer can take many forms depending upon the parameters being measured - electrochemical, optical, mass and thermal changes are the most common. The signal produced by a transducer after a change is measured, may or may not need to be converted to an electrical signal, depending on the type and application of the biosensor.

1.3. Types of Transducers

There are several types of transducers and they can be classified based on their mechanism of action. Some of the types are listed below with examples.

Displacement Transducers: A displacement transducer measures distance or angle traversed by the sensor. A device that produces movement in a straight line is a linear displacement transducer and if it measures movement through an angle it is angular displacement transducer. Displacement transducers are widely used in kinesiology, the study of human movement.

Resistive Transducers: A resistive transducer contains a translational or angular displacement sensing shaft. This shaft drives a wiper in the transduction element that slides along the resistive layer. The resistance measured between the wiper and one of the ends of the layer is a function of the position, motion, or displacement. Resistive transducers are often used in kinesiological studies.

Inductive Transducers: Inductive transducers are used to sense small displacements of movements and are based on the inductance L of a coil. This type of transducer measures movement by changing the value of L, in proportion to the movement of a ferrite or iron core in the bore of the coil assembly. L is given by the following equation $L = n^2 \, x \, G \, \mu$, where n is the number of turns in the coil, x the distance of the ferrite core within the coil assembly, G is the Geometric form constant, and μ is the permeability of magnetically susceptible medium inside the coil. Inductive transducers are often used to measure respiration, where the sensor is connected to a flexible woven belt worn around the chest.

Capacitive Transducers: The sensing shaft in a capacitive transducer changes the position of the dielectric between the capacitor's plates in the transduction element, or it changes the distance and area between the plates. A change in these three parameters leads to a change in capacitance, which is then measured.

Piezoelectric Transducers: Piezoelectric crystals (e.g. quartz) vibrate under the influence of an electric field. The frequency of this oscillation depends on their thickness and cut, each crystal having a characteristic resonance frequency. Piezoelectric crystals can generate an electric field when subjected to physical pressure. As a result, piezoelectric crystals can sense light pressures, such as touch. They are the

basis of tactile sensors. They can also be used to provide vibrotactile stimulation in the study of movement disorders and in force plates to measure ground reaction forces.

Transducers in temperature Measurement: The most popular method of temperature measurement is by using a mercury thermometer, but this is slow, difficult to read, not reliable and prone to contamination. In many cases, continuous monitoring of temperature is required, as in an operating theatre or intensive care unit. Electronic thermometers are convenient, reliable and hence used for continuous monitoring. They use probes incorporating a thermistor or thermocouple sensor, which is an electronic component that has rapid response characteristics of voltage with temperature. Thermocouples measure the temperature difference between two points and not absolute temperature.

Electrochemical Transducers: The underlying principle for this class of biosensors is that many chemical reactions produce or consume ions or electrons which in turn cause some change in the electrical properties of the solution. These electrical properties can be measured.

2. Sensing and Instrumentation in Electrophysiology

Sensing electrical activity within the body is an important task in clinical diagnosis. This section focuses on the physics underlying the acquisition of this electrical activity.

2.1. Electrodes

Sensing is typically achieved with a metal plate electrode, consisting of a metallic conductor, often silver, in contact with the skin with an electrolyte gel in between. These electrodes are often made using a metal foil for flexibility and sometimes in the form of a suction electrode. Silver-silver-chloride (Ag-AgCl) electrodes have been shown to have electrically superior characteristics to silver electrodes, especially when recording low level AC and DC potentials. Chlorided silver electrodes present less low frequency "noise" than silver electrodes. Such electrodes are employed in electrocardiography. A metal disk electrode with a gold surface of conical shape is used in electroencephalography. The gel is used to provide a good electrical contact between the skin and the metal.

A hydrogel electrode is a film saturated with electrolytic solution and made up of sticky materials placed on the electrode surface. The opposite side of the gel layer can be connected to the skin. In this case, the electrolyte directly sticks to the body and reduces motion artifact. This type of electrode is more typically used when the subject is required to move or perform exercise. Hydrogel electrodes are typical in electrocardiography and electromyography.

2.2. Basic Biosensor Amplifier:

The amplifier that is often used in conjunction with transducers is the operational amplifier or Op-Amp [2]. An op-amp is a differential amplifier that amplifies the difference between two inputs. One input has a positive effect on the output signal; the other input has a negative effect on the output. The op-amp is powered by a dual polarity power supply in the range of +/- 5 volts to +/- 15 volts. The theoretically

perfect op-Amp has an infinite voltage gain, theoretically being able to amplify a signal to any level. It also theoretically has infinite bandwidth, being able to amplify any frequency without distortion. It also theoretically has infinite input impedance, being able to sense an input voltage level without distorting that voltage in any way. The perfect Op-Amp also has zero-Ohm output impedance, allowing it to be connected to any other electronic device without distortion. In reality such characteristics are not met, but real op-amps do achieve great performance when correctly configured.

2.3. Inverting Amplifier

The inverting amplifier is one such configuration [2]. The op-amp is connected using two resistors R_A and R_B such that the input signal is applied in series with R_A and the output is connected back to the inverting input through R_B. The non-inverting input is connected to a reference voltage, typically electrical ground or zero volts. This can be achieved by connecting to the centre tap of the dual polarity power supply (Fig. 2a).

During operation, as the input signal becomes positive, the output will become negative. The opposite is true as the input signal becomes negative. The amount of voltage change at the output relative to the input depends on the ratio of the two resistors R_A and R_B. As the input changes either positively or negatively, the output will change in the opposite direction, so that the voltage at the inverting input remains constant or zero volts in this case. If R_A is 1K and R_B is 10K and the input is +1 volt then there will be 1 mA of current flowing through R_A and the output will have to move to -10 volts to supply the same current through R_B and keep the voltage at the inverting input at zero. The voltage gain in this case would be R_B/R_A or 10K/1K = 10. Since the voltage at the inverting input is always zero, the input signal will see input impedance equal to R_A, or 1K in this case. For higher input impedances, both resistor values can be increased.

2.4. Non-inverting Amplifier

Another configuration is the non-inverting amplifier, where the input signal is connected to the non-inverting input and the input resistor R_A is at electrical ground (Fig. 2b). As the input signal changes either positively or negatively, the output will follow in phase to maintain the inverting input at the same voltage as the input [2]. The voltage gain is always greater than 1 and can be calculated as $Vgain = 1 + R_B/R_A$.

2.5. Voltage Follower

The voltage follower, also called an electrical buffer, provides high input impedance, a low output impedance, and unity gain. As the input voltage changes, the output and inverting input will change by an equal amount (Fig. 2c). The voltage follower is often used to electrical isolate the transducer from a computer, to allow subsequent analysis and signal processing to take place.

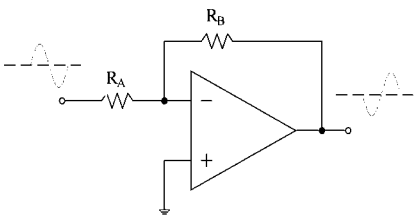


Figure 2. (a) Inverting Amplifier

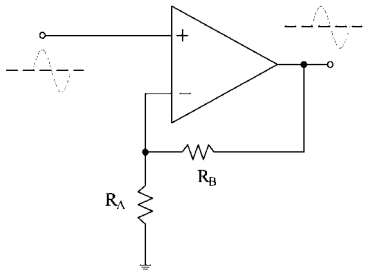


Figure 2. (b) Noninverting Amplifier

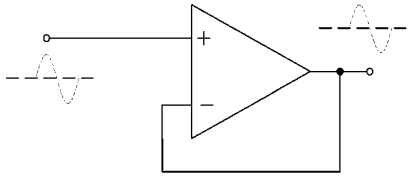


Figure 2. (c) Voltage Follower

2.6. Input Impedance

Impedance (symbol Z) is a measure of the overall opposition of a circuit to current. It is like resistance, but it also takes into account the effects of capacitance and inductance. Impedance is measured in Ohms. Impedance is more complex than resistance because the effects of capacitance and inductance vary with the frequency of the current passing through the circuit and thus implies that impedance varies with frequency. Input impedance (Z_{IN}) is the impedance resulting from the connection of any input to a circuit or device (such as an amplifier). It is the combined effect of all the resistance, capacitance and inductance connected to the input inside the circuit or device. The effects of capacitance and inductance are generally most significant at high frequencies. For good signal amplification, the input impedances should be high; at least ten times the output impedance of the transducer supplying a signal to the input. This ensures that the amplifier input will not overload the transducer producing the signal and thus reduce the voltage of the signal by a substantial amount.

Human skin presents large impedance if connected via an electrode to an amplifier. The impedance is greatly reduced if the surface layer of the skin is removed. This layer is known as the stratum corneum. In practice, for electrocardiography or electromyography, the stratum corneum is removed by lightly abrading the skin. The skin contact with the electrolyte gel and silver-silver chloride electrode now presents significantly lower impedance. For electroencephalography, input impedance at each electrode site is kept below 5kOhms.

2.7. Instrumentation Amplifiers

An instrumentation amplifier is a type of differential amplifier that has been specifically designed to have characteristics suitable for use in measurement and test equipment. Its characteristics include very low DC offset, low electrical drift, low noise, very high gain, very high common-mode rejection ratio, and very high input impedances [1][2]. Instrumentation amplifiers are used where great accuracy and stability of the circuit, both short- and long-term are required. The ideal common-mode gain of an instrumentation amplifier is zero. Instrumentation amplifiers can be built with individual op-amps and precision resistors, but are also available in integrated circuit form.

2.8. Filtering

Filtering is a necessary processing step with any biosensor, in order to reduce noise to a minimum and accentuate frequencies of interest. Filtering may be carried out in software using digital signal processing, but it can also be accomplished in hardware [3].

When designing a filter, the range of frequencies we wish to pass through the filter without attenuation is known as the passband. The range of frequencies we wish to attenuate is known as the stopband. The steepness of the filter's transition from the passband to the stopband, is related to the filter order. The higher the filter order, the sharper the transition and more precise is the filter in pass and attenuating frequencies.

Three 2^{nd} order filters are described here. A low pass, high pass, and bandpass filter. Each of these filters will attenuate frequencies outside their passband at a rate of 12dB per octave equivalent to 25% of the voltage amplitude for each octave of frequency increase or decrease outside the passband (Fig. 3(a) and Fig. 3(b)).

First order low or high pass cutoff frequency is given by the choice of a resistor and capacitance value.

$$Frequency_{Cutoff} = \frac{1}{2\pi RC} \tag{2}$$

2nd order low or high pass cutoff frequency (-3dB point) is also chosen by resistor and capacitance values.

$$Frequency_{Cutoff} = \frac{1}{2\pi\sqrt{R_1 R_2 C_1 C_2}} \tag{3}$$

A bandpass filter passes a range of frequencies while rejecting frequencies outside the upper and lower limits of the passband Fig 3(c). The range of frequencies to be passed is called the passband and extends from a point below the centre frequency to a point above the centre frequency where the output voltage falls to approximately 70% of the output voltage at the centre frequency. These two points are not equally spaced above and below the centre frequency but will look equally spaced if plotted on a log graph. The percentage change from the lower point to the centre will be the same as

from the centre to the upper, but not the absolute amount. The filter bandwidth (BW) is the difference between the upper and lower passband frequencies. A formula relating the upper, lower, and center frequencies of the passband is

$$Centre\ Frequncy = \sqrt{\text{LowerFrequency} * \text{Upper Frequency}}$$

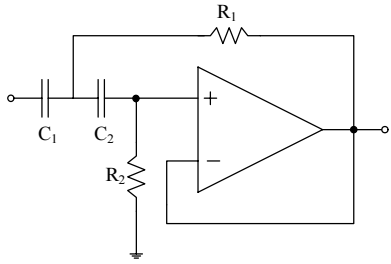


Figure 3. (a) High Pass Filter

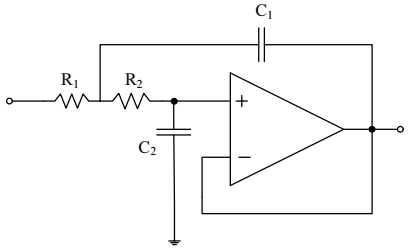


Figure 3. (b) Low Pass Filter

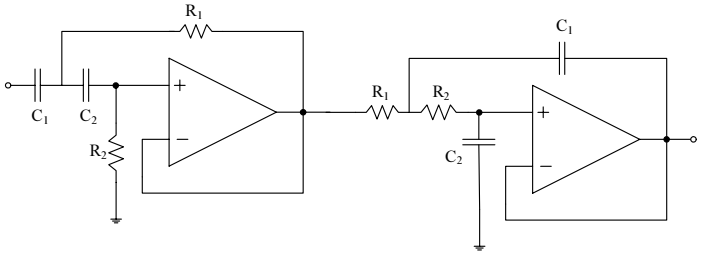


Figure 3. (c) Band Pass Filter

2.9. Electrical Safety with respect to electrophysiology:

Electrical safety is very important in hospitals as patients may be undergoing a diagnostic or treatment procedure where the protective effect of dry skin is reduced. Burn injuries can occur when using skin-electrode interfaces. This is specifically relevant to electromyography or functional electrical stimulation (FES). In FES, electrical current is injected into a muscle to initiate a muscle contraction. Burn injuries can occur when there is small contact between the skin and the electrode. The injury received from electric current depends on the magnitude of current, the pathway that it takes through the body and the time for which it flows. As a result, physically large electrodes, which provide a wide area for the current to pass are necessary in FES to form a large area of the current to pass and reduce the risk of injury.

3. Examples of Biosensors in Use

This section provides an example of the use biosensors in clinical engineering, a branch of biomedical engineering which is concerned with equipment routinely used in hospitals and clinics [4].

3.1. Direct BP Measurements:

Direct blood pressure measurements are made by introducing a cannula (needle) into an artery and then coupling this to a pressure transducer. The advantage of this approach is that pressure is constantly monitored beat-by-beat and a clinically important waveform (a graph of pressure against time) can be displayed. Patients with invasive arterial monitoring require very close supervision, as there is a danger of severe bleeding if the cannula line becomes disconnected. It is generally reserved for critically ill patients where rapid variations in blood pressure are anticipated.

3.2. Indirect BP Measurements:

Indirect measurement involves use of an inflatable cuff coupled to a pressure gauge (sphygmomanometer). In the standard sphygmomanometric approach, the cuff is wrapped round the arm (placed at about heart level) and a pressure, higher than the expected blood pressure, applied to the artery. A stethoscope placed over the brachial artery is used to listen to the changes in sounds as the cuff is slowly deflated. The first Korotkoff sounds occur when the systolic pressure, the highest pressure reached when the ventricles contract and restrict blood flow, first exceeds the pressure in the cuff so that blood once again flows through the artery beneath the stethoscope. At systolic pressure a clear tapping sound is heard in time with the heart beat. The Korotkoff sounds become muffled and disappear when the pressure in the cuff drops below the diastolic pressure, the minimum pressure that occurs at the end of ventricular relaxation.

Another indirect approach of measuring blood pressure is the oscillometric method. Here a microprocessor periodically inflates and deflates the cuff. When the blood breaks through the occlusion caused by the cuff, the walls of the artery begin to vibrate slightly due to the turbulent nature of the blood flow. The onset of these oscillations in pressure correlates with systolic pressure. Diastolic pressure corresponds to the pressure when the amplitude of the oscillations drops to zero. It is the oscillometric method that is employed in Holter monitors, which record blood pressure over a 24-hour period while the patient-under-observation goes about their daily activity.

3.3. Respiratory System Measurements (Spirometry)

Spirometry is the measurement of lung function and is the most common pulmonary function test. It is an important tool used in assessing conditions such as asthma, pulmonary fibrosis, cystic fibrosis, and chronic obstructive pulmonary disease (COPD). There are many designs of spirometers, but most modern devices are electronic based. They all involve the measurement of the volume and flow of breath during inhalation and exhalation. The most common electronic based devices consist of a small handheld

tube into which the patient exhales as hard as possible and for as long as possible, having been just requested to take the deepest breath possible. A flow sensor in the tube measures the direction and volume of the breath flow. Following exhalation, the patient rapidly inhales (inspiration). Inspiration is important when assessing possible upper airway obstruction and is sensed in the tube by the directional sensor. Sometimes, the test will be preceded by a period of quiet breathing in and out from the tube, so that sensor can measure tidal volume. During the test soft nose clips may be used to prevent air escaping through the nose.

4. Conclusions

As can be appreciated, the majority of biosensors are essentially the same as sensors used in other commercial or industrial applications. The distinctive aspect about biosensors is their application. However, there are problems that are encountered by biosensors that are unique to them. These problems are associated with the interface between the biosensor and the biological system being measured. The presence of foreign materials, especially implanted materials, can affect the biological environment surrounding the sensor. Therefore one must be aware of potential physical or chemical reactions as a result of the sensing process. The presence of the sensor may also hinder the subject's physical and cognitive ability. Therefore the placement of the sensor on the body must be given due care and attention.

Despite these issues, biosensors have an enormous clinical impact both in diagnosis and therapy, as well for being critical for data collection for biomedical research.

References

[1] Geddes L.A. and Baker L.E., (1989), *Principles of Applied Biomedical Instrumentation*, Wiley-Interscience, New York.

[2] Horowitz P. and Hill W., (1989), *The Art of Electronics,* Cambridge University Press.

[3] Wise D.L. (1991), *Bioinstrumentation and Biosensor*, Marcel Dekker, New York

[4] Harsanyi G. (2000), *Sensors in Biomedical Applications: Fundamental Technology and Applications*, CRC, Florida.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-119

Introduction to Chapter III: Medical Informatics for Biomedical Engineering

Paul McCULLAGH and T. Clive LEE (eds.)

Medical Informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and healthcare.

Chapter III.1 will introduce the student to the range of computing devices, their operating systems and programming environment. Most devices no longer work in isolation so the option for networking will be explored. When data is moved from one system to another across a network security issues such as privacy, authentication, and data integrity are important. A Case study will illustrate the role of ICT and the Internet in the rehabilitation of stroke patients.

Data should be structured for storage and retrieval, as indicated in Chapter III.2. The database is the fundamental storage repository. Data may be stored as free text, and in image format, but coding is important to aid analysis and audit. The chapter will discuss the most important formats for coding to assist classification. Unambiguous description and standards are important. To date interoperability has been most advanced in ECG equipment. A Case study on a markup language to describe the electrocardiogram (ecgML) will be presented.

As the amount of data becomes ever larger, mining techniques are required to extract relevant information from the database. A goal of medical informatics is that 'knowledge' can be deduced which will aid in decision support. Chapter III.3 explores these topics and discusses a Case Study which uses a machine learning approach to discover the optimal placing of electrodes for the recording of the ECG.

Advances in Information and Communication technology (ICT) have resulted in the use of the Internet to support medical applications (eHealth) along with the use of high speed communication links to support Telemedicine and Telecare. This is particularly important when large distances separate the healthcare professional and the patient along with the provision of home based assistive services. Chapter III.4 illustrates these issues by using a Case Study to support the home based management and dispensing of medication.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-121

III.1. Medical Informatics and eHealth

Paul J. McCULLAGH[a], Huiru ZHENG[a], Norman D. BLACK[a], Richard DAVIES[a], Sue MAWSON[b] and Kieran McGLADE[c]
[a]*Faculty of Engineering, University of Ulster at Jordanstown,Co. Antrim, N Ireland*
[b]*Centre for Health and Social Care Research, Sheffield Hallam University, UK*
[c]*Department of General Practice, Queens University of Belfast, N Ireland*

Introduction

Medical Informatics combines the disciplines of medicine and computing, and knowledge of this area is a fundamental requirement for a student of biomedical engineering, providing support for software development, information gathering and critical assessment, and evidence based medicine [1]. This review assumes the following definition:

Medical Informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and healthcare [2].

One often hears the term "eHealth". This is a related area, usually understood to apply to direct applications of medical informatics to health care. It often involves use of the Internet and includes the activities of patients and clients as much as health care professionals. We will use the following definition:

"eHealth" is concerned with the application of electronic information systems in the organization of and access to health care including such examples as booking and referral systems; it covers integrated information tools which allow secure access to personal health data for all those who need it to deliver optimal healthcare; and it includes complex clinical applications which can support the clinician in diagnosis and treatments and ultimately supports the citizen in their own environment [3].

Essentially eHealth places the citizen at the centre of their own well being and care. In this review we will introduce computing devices, their architecture, operating systems and programming environments. As most devices no longer work in isolation, the options for connectivity and networking will be explored. When data are moved from one system to another across a network, security issues such as confidentiality, incorporating preservation of privacy, authentication and data integrity are important. An eHealth case study will illustrate the role of Information and Communication Technology (ICT) and the Internet, utilizing a biomedical engineering solution to assist the rehabilitation of stroke patients.

1. Role of the Computer in Medical Informatics and eHealth

The computer is ubiquitous in everyday-life and is an essential tool for the biomedical engineer. It is fundamental to the acquisition, storage, processing and analysis of data,

and can be used to summarize and display information via the human computer interface (HCI) so that it is comprehensible to users. This interface is normally graphical nowadays and can be very sophisticated, e.g. 3-dimensional image construction of the fetus in the womb, by processing reflections of externally applied ultrasound, providing the means for clinical assessment. Indeed processing capacity is such that images can be constructed and displayed in real-time to provide a non-invasive 'movie' of the fetus [4]. Alternatively, a more straightforward example is the use of automated speech recognition to allow a specialist to record clinical observations directly into the computer. An example of this is the reporting of diagnostic imaging results in radiology [5].

The ability to be connected to private networks and the Internet is essential for information sharing and for information retrieval from large authoritative information repositories. The Electronic Patient Record (EPR), which is now commonplace in Primary Care, is an example of the former, with many additional, often incompatible systems existing in Secondary Care departments of large hospitals. In the United Kingdom, a key aim of the government's Connecting for Health program [6] was to establish a central database of 50 million patients' medical records, accessible over NHSnet[1]. This is the world's largest Information Technology development program, with an estimated cost from six to twenty billion pounds. The United States National Library of Medicine's publicly accessible database PUBMED [7, 8] with over 16 million citations from MEDLINE is an example of an Internet based resource. Others include large data banks such as the human genome project.

Van Bemmel has provided an important conceptual model for structuring computer applications in health [9]. It comprises six levels, with the lowest level being largely automated and the higher levels requiring progressively more human interaction. This model provides a useful context in which to discuss the interaction of computers, networks and eHealth. We will use the recording, storage and interpretation of the electrocardiogram (ECG), an electrical recording used to assess the physiology of the heart, as an example.

Level 1: *Communication and telematics* deals with data acquisition and transfer, e.g. the recording of the ECG potential waveform with electrodes and physiological amplifier and its subsequent transfer, in a standardized format, to a computer screen for human interpretation. The computer can be thousands of miles distant, a capability which has promoted the emergence of telemedicine as a discipline.

Level 2: *Storage and retrieval* deals with the storage of information in a data repository, e.g. multimedia EPR which capitalizes upon advances in memory storage capacity and database technology to form digital libraries comprising labeled text, physiological recordings and images.

Level 3: *Processing* requires a computer program to extract information from the captured data. For an ECG, this could be important diagnostic parameters such as heart rate, or pattern abnormalities such as ST elevation[2], which may have some diagnostic significance. Similarly for an image (e.g. an echocardiogram), the computer can be used to segment boundaries and highlight potential abnormalities, for human verification.

Level 4: *Diagnosis* combines features extracted from the processing of acquired data with existing medical knowledge to assist with decision making, e.g. the '*ECG is*

[1] NHSnet is a private network connecting hospitals and general practitioner sites in the UK
[2] The well know ECG complex is designated by QRST markers

abnormal' or '*ECG indicates an arrhythmia*' and some therapy is required. A Decision Support System (DSS) includes an knowledge-base (KB) containing information such as statistical norms for sex and age. It can provide alerts, critiques and advice. The DSS is a useful tool to support a diagnosis. However, the treating physician may be aware of many additional case-specific factors, such as previous illness and medication, which may have significant influence far beyond the current knowledge in the KB. The physician always makes the final diagnosis. It is possible to conclude that a physician's diagnosis may be as much an art as a science.

Level 5: *Treatment* utilizes the information derived from the diagnosis to formulate an intervention, e.g. fit an on-demand pacemaker, which will stimulate the patient's heart when the cardiac beat becomes irregular. A computer algorithm will form part of the embedded control system.

Level 6: *Research* allows users to interact with data to formulate hypotheses and build computer models, e.g. to simulate the electric depolarization of the heart, which can then be used to promote understanding of normal and abnormal function. The possibilities for research are only limited by prevailing knowledge, tools and the ingenuity of the human mind.

1.1. Architecture of the Computer

Logically, the computer comprises three elements: (i) a processor (control unit and arithmetic and logic unit (ALU)), (ii) random access memory, some form of persistent storage, currently a hard disk (memory) and (iii) input/output devices. This is the classical von Neumann architecture [10] or "stored-program computer", which has been in use for decades and forms the design of most of today's computers. Parallel architectures are also in use, particularly for high performance applications. At the heart of a computer is an algorithm, which is a set of programmed instructions to carry out some task. The design uses a single store to hold both instructions and data, and the control unit instructs the ALU to execute instructions sequentially.

Any hardware specification quickly becomes outdated due to the pace of technological progress. This can be summed up by Moore's Law (proposed by Gordon E. Moore a co-founder of Intel, the major silicon chip manufacturer) which observes that the transistor density of integrated circuits, with respect to minimum component cost, doubles every 24 months [11]. A similar law has held for hard disk storage cost per unit of information, and indeed random access memory (RAM) storage capacity has increased at the same rate as processing power. Kurzweil's expansion of Moore's Law [12] shows that the underlying trend has held true from integrated circuits to earlier transistors, vacuum tubes, relays and electromechanical computers. He projects continued progress until 2019 with transistors just a few atoms in width, and predicts that it is possible that Moore's Law may be extended by a paradigm shift into new technology, e.g. quantum computers.

The computer has many different incarnations: supercomputer, grid computer, mainframe, server, desktop, laptop, personal data assistant, smart phone and embedded device. A summary of characteristics is provided in Table 1.

Table 1: Classifying computers by scale and application

Computer type	Description	Example Application	Possible Operating System (OS)
Supercomputer	High power number cruncher using an array of powerful processors	Modelling (computationally intensive), simulation, research e.g. identifying genetic disposition for various disease from the human genome	Proprietary, Unix, or Linux
Grid	Virtual cluster of desktop computers for sharing resources	Modelling (computationally intensive) or access to information or computational resources	Heterogeneous operating systems, linked by 'open' middleware using standards
Mainframe	Large central computer, accessed by 'terminals'	Administration and storage of data, Hospital Information System, EPR, Patient Administration System, Picture archive and retrieval	UNIX, VMS or other proprietary systems
Server	High performance computer	Web, Database, email, etc	UNIX, Window NT
Desktop, Workstation	General computer for personal use and storage of information; workstation may have enhanced graphics	Multi-purpose, office, web access, email, program development. Workstation suited to image processing applications e.g. Radiology department	Windows, Linux, Apple
Laptop, Tablet PC	Similar specification to desktop but with portability the main requirement	Multi-purpose, as above, normally with built in connectivity	Windows, Linux, Apple
Personal data assistant	Computer in your hand	Cut down versions of office programs, specialized 'calculator' programs, normally with built in connectivity.	Windows CE, Palm OS
Smart phone	Telephony, with storage and limited processing (e.g. games)	Organizer, small programs, with connectivity to the cellular phone systems (GSM, GPRS, 3G)	Symbian OS
Embedded device	Dedicated control of hardware	ECG monitor / defibrillator. Bespoke software development	Windows CE or Linux

The supercomputer and grid perform specialized tasks, typically in the research world, where enormous processing is required. For example, *supercomputers* have been used to model avian flu progression by scientists from Los Alamos National Laboratory [13], and to model HIV's vulnerability to a class of drugs known as protease inhibitors [14] by researchers at Stony Brook University. The former application was implemented on a platform developed for the nuclear weapons programme comprising 2048 processors, the largest cluster in the world. The latter simulation took 20,000 hours of processing (about three months) on the USA National Center for

Supercomputing Applications. However, it would have taken more than a year to complete the work on a conventional computing system.

Grid computing [15] involves sharing heterogeneous resources (based on different platforms, hardware/software architectures, and computer languages), located in different places belonging to different administrative domains over a network using open standards. This offers a model for solving huge computational problems by making use of the resources of large numbers of computers, treated as a virtual cluster embedded in a distributed telecommunications infrastructure. It has the goal of solving problems too big for any single supercomputer, whilst retaining the flexibility to work on multiple smaller problems. Secure authorization techniques allow remote users to control computing resources. The availability of large amounts of data (clinical, genomic) in heterogeneous sources and formats, and the progress in fields such as computer based drug design, medical imaging and medical simulations have lead to a growing demand for large computational power and easy accessibility to data sources. Functional imaging research in schizophrenia [16] is an example application, integrating of large numbers of medical images, with patient's medical and biological data, to assess genetic pre-disposition to the condition.

In order to help combat malaria, which kills more than 1 million people annually, CERN (Conseil Européen pour la Recherche Nucléaire, the European Organization for Nuclear Research) has launched a grid computing effort to run a simulation program called MalariaControl.net [17], developed by researchers at the Swiss Tropical Institute. The program simulates how malaria spreads through Africa, enabling researchers to better understand the impact of introducing new treatments. On a test phase of a few months with 500 volunteers, the grid was able to run simulations equivalent to 150 years of processing time on a single computer.

Mainframes are enterprise computers with high specification and associated high cost. In the context of health they are used for large Hospital Information Systems to store patient administration systems and more recently the EPR, with technical support from an Information Systems department.

By comparison to the previous systems, a *server* is a smaller computer system with a dedicated function, such as a database, email or web servicing. The hardware and software specification tends to be more demanding than the conventional desktop, as it may need to serve multiple simultaneous connections, and secure data backup is often a requirement. The *workstation* can be a dedicated high specification desktop with for example high specification graphics, as required for imaging applications. The *desktop* provides cost effective processing power with significant storage and easy connectivity for stationary users. The system unit typically interfaces to a keyboard, mouse and screen display. Printers and scanners are common peripherals and high volume backup may be achieved by compact disk or digital versatile disk rewriter. The floppy disk drive has been replaced by a memory stick (or card) with capacity in Giga (10^9) bytes for file transfer, ad-hoc storage and backup[3]. The interface options to external devices are typically Universal Serial Bus (USB2.0), with firewire (IEEE 1394 High Speed Serial Bus) and memory card readers as alternatives. Local area network (LAN) connectivity to a fixed wired Ethernet typically provides a data rate of between 10-

[3] Terms for storage: Terabytes (10^{12} bytes), Gigabytes (10^9 bytes), Megabytes (10^6 bytes), Kilobytes (10^3 bytes),

1000 Mbps (10^6 or million *bits per second*). While this provides adequate quality of service for medium sized networks (up to say 30 users), the growing demand for multimedia applications (including the transfer of sound, images and movies, all of which are required in the health domain) continues to challenge the performance of the network.

Mini desktop PCs are beginning to emerge. These have smaller footprints, and significantly lower energy requirements than the conventional tower or desktop. Average Power Consumption of only 29Watts, is approximately 30% less energy than a standard PC. This is important in the reduction of the carbon footprint of computer technology. *Virtualisation* lets one larger computer do the job of multiple computers, by sharing the resources of a single computer across a network. A small energy efficient 'thin client' (no hard disk and minimal peripherals) is required on the desktop. In addition to energy savings and lower cost due to more efficient use of hardware, virtualisation provides better desktop management, increased mobility and security, and improved disaster recovery.

The *home computer* sacrifices little functionality by comparison to the desktop. The network connection was initially via a telephone dial-up (Public Switched Telephone Network) to an Internet Service Provider (ISP) at 56 kbps (10^3 or thousand bits per second) but increasing is via broadband (Asymmetric Digital Subscriber Line or cable television) at speeds typically of 1-10 Mbps for download, and 512 kbps for upload. The speeds have increased as sophisticated signal processing yields more bits per second from the available bandwidth.

For mobility, a number of options are available. The *laptop* comprises practically all the functionality of the desktop, but with a slightly inferior specification-price ratio. Wireless Bluetooth using radio waves or infrared light (The Infrared Data Association, IrDA) provide unfettered desktop connectivity. Both wired Ethernet and wireless Ethernet (known as wireless fidelity or WiFi) are normally supported. Wireless offers nominal speeds of the order of 11-54 Mbps, depending on the characteristics of the network access point, and the wireless interface built into the laptop. This is inferior to the wired network speed of typically 100 Mbps. Reliability of connection is probably a more important issue for the user, with the wired network providing better service. The *tablet* PC is a variant of the laptop, with automatic handwriting, a standard input feature. A major issue with portable computing is the capacity of the battery, typically supporting a couple of hours of unfettered computing and hence a major design issue is the conservation of power. This has spawned research into harnessing power from the environment, e.g. from kinetic energy or solar power.

The *personal data assistant* (PDA) provides computing power which can be stored in the pocket. This is very appealing in the area of medicine, e.g. a ward round in a hospital where even a lightweight laptop can be obtrusive. The PDA comprises similar connectivity interfaces to the laptop and supports similar applications to desktop such as word processing, spreadsheet, database, calendar, Internet browsing, and email, normally with limited functionality. However, due to the small screen size and small keypad, usability is the major difficulty. Some models have virtual keypads, others have small keyboards or incorporate speech recognition. Many medical applications have been developed for the PDA, including dictionaries, calculators, guidelines and databases, see [18].

In many respects, from technology and user perspectives, the PDA and the *smart phone* have converged. The smart phone primarily provides telephony within a Global System for Mobile communication (GSM) cell (IS-95 in USA). However the use of

data services such as Short Message Service (SMS), and Internet access using Wireless Access Protocol (WAP) or I-Mode[4] using General Packet Radio Service (GPRS) are in widespread use. The telephone companies have introduced third generation (3G) cellular systems based on Universal Mobile Telecommunications System (UMTS), with data rates of 384 kbps and beyond, which can support video conferencing and hence basic telemedicine. Further technological enhancements offering performance approaching LAN speeds, e.g. High Speed Packet Download Access (HSPDA) are under development (download speed of 3.6 Mbps with 14.4 Mbps planned for the future). The smart phone also contains a processor and can be programmed for bespoke applications. By enhancing a PDA with GPRS connectivity, the PDA essentially becomes a telephone. Similar convergence from ICT means that a computer, laptop or PDA connected to a 'hot-spot' can utilize voice over internet protocol (VOIP) to provide telephony or 'chat' services.

Of course, embedded devices with microprocessor controllers are also ubiquitous in biomedical engineering. These can utilize advances in miniaturization to perform dedicated functions, under the control of a program. An example is an ECG monitor, which uses purposely designed hardware and software to acquire and display the electrocardiogram. If a new clinical measure is required, then an updated program can be downloaded from a web site to the device firmware. Semi-automated ECG defibrillation systems rely more heavily on embedded software for decision support.

1.2. Operating Systems

An operating system (OS), is a computer program that manages the hardware and software resources of a computer. The OS performs basic tasks such as controlling and allocating memory, prioritizing system requests, controlling input and output devices, facilitating networking, and managing files. It may provide a graphical user interface. Microsoft Windows is the dominant OS on the PC and Apple Macintosh platforms, with 90% market share. The main competition comes from Open Source[5] protagonists who support Unix-variants such as Linux. On larger mainframe computers Unix and proprietary systems such as OpenVMS dominate. The latest Windows OS, Windows 7, includes added functionality in security and network administration.

The most common operating systems used by smart phones are Symbian, Linux, Windows Mobile, and Palm OS. Windows Compact Edition (CE) is an OS for minimal computers, such as PDAs. It is optimized for devices that have minimal storage and may run in under a megabyte of memory. Embedded systems normally use dedicated operating systems and versions of Linux.

Programming environments for the smart phone and PDA are Microsoft.NET Compact Framework and Java[6]. For the former, applications may be authored in languages such as C# or Visual Basic.NET. For the latter, Java programs (know as MIDlets) may be authored for embedded devices.

---

[4] NTT DoCoMo's I-Mode is a wireless Internet service made popular in Japan, providing access to a number of consumer sites

[5] Open principles and practices are applied to the development of source code that is made available for public collaboration, and released as open-source software

[6] Java is an open source programming language from Sun Microsystems

1.3. Client-Server Computing

Client/server is a network-based architecture, see Figure 1, which separates the client computer (often an application that uses a browser) from the server computer. It uses the HyperText Transfer Protocol (HTTP), using methods to publish and retrieve HTML pages. The destination server, which stores or creates resources such as HTML files and images, is called the origin server. An HTTP client initiates a request by establishing a Transmission Control Protocol (TCP) connection to a particular address (called a port) on a remote host (the address is port 80 by default). An HTTP server listening on that port waits for the client to send a request message. Upon receiving the request, the server sends back a confirmation, and a message of its own, the body of which is perhaps the requested file, an error message, or some other information. Resources to be accessed by HTTP are identified using Uniform Resource Identifiers (URIs).

Each instance of the client software can send requests to a server, e.g. a file server, terminal server, or mail server. Examples of servers programs are Apache and Windows Internet Information Services (IIS). The Apache Server Project [19] maintains an open-source HTTP server for modern OSs, including UNIX and Windows NT. IIS provides similar services, but restricted to the Microsoft platform.

The three-tier architecture separates the user interface, functional process logic, and data storage/access. The programs are developed and maintained as independent modules (possibly on separate platforms), so that any of the tiers may be upgraded or replaced independently as requirements or technology change. Typically, the user interface runs on a desktop PC or workstation and uses a standard graphical user interface. Functional logic may consist of one or more separate modules running on an application server, and a Relational Database Management System (RDBMS) on a database server contains the data storage logic. Structured Query Language (SQL) is used to create, modify, retrieve and manipulate data from relational database management systems.

Interoperability is important when interacting with heterogeneous platforms. A Web service is a software system designed to support interoperable machine-to-machine interaction over a network, using a distributed function (or method) call interface. Simple Object Access Protocol (SOAP) is a protocol for exchanging (extensible markup language) XML-based messages over the computer network, normally using the HTTP protocol.

Whilst client-server is the dominant architecture, peer to peer systems have become more popular for file sharing and 'chat' applications on the internet. In this case, files may be transferred without the involvement of a server. Often a server-based database holds the location of information, e.g. music files, but not the actual files, permitting a subsequent peer to peer connection for transfer.

1.4. Computer Networks

Networks may be categorized by scale as: personal area networks, local area networks and wide area networks.

A *personal area network* (PAN) is a computer network used for communication among computer devices (including telephones and personal digital assistants) a few meters (normally 10m but up to 100m is possible) apart. PANs can be used for communication among the personal devices themselves, or for connecting to a higher

level network and the Internet. Personal area networks may be wired (USB and FireWire) or wireless (WPAN) using technologies such as IrDA and Bluetooth. ZigBee is the name of a specification for a suite of high level communication protocols using small, low-power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks. ZigBee operates in the Industrial, Scientific and Medical (ISM) radio bands; 868 MHz in Europe, 915 MHz in the USA and 2.4 GHz worldwide. The technology is simpler and cheaper than Bluetooth.

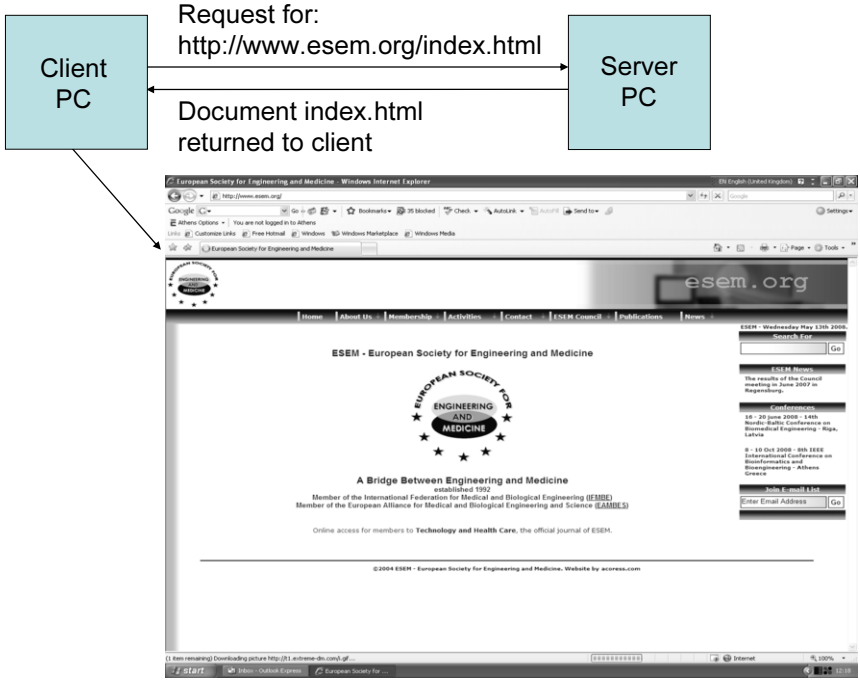


Figure 1: Example of Client Server Interaction

A *local area network* (LAN) is a computer network normally covering an office, but it is increasingly used in the home. Current LANs are most likely to be based on switched IEEE 802.3 Ethernet running at 100 or 1,000 Mbit/s or on WiFi technology. The defining characteristics of LANs in contrast to WANs (wide area networks) are: their much higher data rates; smaller geographic range; and that they do not require leased telecommunication lines.

A *wide area network* (WAN) covers a broad geographical area, e.g a country. WANs are used to connect local area networks (LANs) together, so that users and computers in one location can communicate with users and computers in other locations. Many WANs are built for one particular organization and are private. Others, built by Internet service providers, provide connections from an organization's LAN to the Internet. WANs are most often built using telephone lines. At each end of the

leased line, a router connects to the LAN on one side and a hub within the WAN on the other. Network protocols including TCP/IP deliver transport and addressing functions.

The *Internet* is the worldwide, publicly accessible network of interconnected computer networks that transmit data by packet switching using the standard Internet Protocol (IP). It is a "network of networks" that consists of home, academic, business, and government networks, which together carry various information and services, such as electronic mail, online chat, file transfer, and the interlinked Web pages.

1.5. Security

In eHealth, sensitive confidential information about a person's health status, may be stored in computer systems, or transferred across computer networks. Security deals with the unauthorized access to computers, networks or data. Basic security requires that access to a computer system is controlled by username and password, preventing unauthorized access to programs and data. Enhanced access control measures include the use of a smart card and card reader and the use of biometric data such as fingerprint scanning.

Encryption is the process of obscuring information to make it unreadable without special knowledge of the algorithm used and the key to unlock it. Encryption is now employed in protecting widely-used systems, such as Hospital records, Internet e-commerce, mobile telephone networks and bank automatic teller machines. Encryption can use either secret key algorithms such as Data Encryption Standard (DES) and Advanced Encryption Standard (AES) or public/private key algorithms such as RSA, named after its inventors Rivest, Sahmir and Adelman. Encryption can be used to ensure secrecy, but other techniques are still needed to make communications secure, particularly to verify the integrity and authenticity of a message; for example, a message authentication code (MAC) or digital signature. In computer security, authentication is the process of attempting to verify the digital identity of the sender of a communication such as a request to log in. The sender being authenticated may be a person using a computer, a computer itself or a computer program. Authentication verifies a person's identity, while authorization is the process of verifying that a known person has the authority to perform a certain operation. Authentication, therefore, must precede authorization.

Security measures include the use of firewall technology to restrict access to computer systems, except via an 'electronic drawbridge'. This permits control of accepted programs, ports and IP addresses. Anti-virus software is also important to prevent malicious infiltration of and attack on computer systems. Good security also dictates that users adhere to strict guidelines. For example wireless networks are normally considered to be less secure than wired equivalents and may permit the opportunity to by-pass the secure firewall. Hence Wired-Equivalent Privacy (WEP) should be utilized on wireless devices. This security discussion is probably equally relevant to the home and office environments.

1.6. eHealth

eHealth is health care practice which is supported by electronic processes and communication. eHealth can use advances in ICT to empower citizens and patients, enabling them to play an active role in decisions relating to their health. The European Union has launched a portal [20] to provide citizens with access to comprehensive

information on Public Health initiatives, to positively influence behaviour and promote the steady improvement of public health in Europe. Possible services are summarized in Table 2. eHealth may be particularly appropriate to the management of chronic disease in an era when the population is ageing and the costs of providing support are becoming prohibitive.

Table 2: eHealth examples

eHealth Component	Explanation
Electronic Medical Records	Communication of data between patient and healthcare professionals e.g. GP, specialists, care team, pharmacist
Telemedicine	Physical and psychological measurements that do not require a patient to travel to a specialist
Evidence Based Medicine	Healthcare professional can look up whether his/her diagnosis and treatment are in line with current scientific research.
Citizen-oriented Information	Provides both healthy individuals and patients with information on medical topics
Specialist-oriented Information Provision	Overview of latest medical journals, best practice guidelines or epidemiological tracking
Virtual healthcare	Healthcare professionals who collaborate and share information on patients through digital equipment

2. CASE Study: SMART Rehabilitation: Tele-Rehabilitation and eHealth

We will use the following case study to emphasize the importance of computers and networking to the eHealth paradigm and relate this to the van Bemmel conceptual layer model [2], introduced in section 2.

2.1. Stroke and Rehabilitation

In the United Kingdom, stroke is the most significant cause of severe disability affecting a quarter of a million people [21]. At six months post-stroke, around half of patients need help with bathing, a third need help with dressing and a third need help with feeding [22]. Rehabilitation is primarily concerned with maximising the functional and cognitive ability of the patient combating social exclusion and increasing patient independence [22]. The National Framework for Older People recommends that rehabilitation should continue until maximum recovery has been achieved [24]. Research suggests that intensive, task specific and repetitive training may be necessary to modify neural organisation [25]. However, due to cost factors, in-patient length of stay is decreasing and outpatient rehabilitation facilities are limited. Furthermore, current rehabilitation techniques used in hospitals and rehabilitation centres require access to specialised equipment and a dedicated laboratory set-up [26]. Therefore there is a strong argument for the need to develop a low-cost, accessible system that can

augment existing rehabilitation services for post-stroke patients. Currently, two main approaches have been explored, namely tele-rehabilitation and home-based rehabilitation.

Tele-rehabilitation, was firstly proposed in 1997 by the National Institute on Disability and Rehabilitation Research (United States Department of Education), it is defined as:

the use of communication and information technologies to deliver rehabilitation services and exchange information over geographical distances [27].

The advantages of rehabilitation in a home environment are [28]:

- Enhanced evaluation of daily living activities, which improves outcome;
- More frequent and timely treatment that would otherwise burden health care systems;
- Patients can stay at home for extended periods, reducing costs;
- Feedback can enhance user motivation;
- Empowerment of the user by enabling them to make decisions about the time and frequency of their intervention;
- Chronic patients can benefit from regular and frequent intervention.

SMART rehabilitation: technological applications for use in the home with stroke patients, examined the scope, effectiveness and appropriateness of eHealth to support home-based rehabilitation for older people and their carers [29]. It utilized Information and Communication Technology and motion sensors to provide tele-rehabilitation alongside home-based rehabilitation.

The design and development phases were guided by an expert panel of national and international academics. In addition an expert group of stroke patients and their carers contributed to developments by way of focus groups and consultations [30]. Throughout the development, neurological physiotherapists engaged with decisions about the interface and the feedback given to the user. A biomedical engineering design team was involved in the development of the sensor attachment to clothing ensuring that the clothing was acceptable and usable by the stroke patients.

2.2. SMART Rehabilitation System Overview

The SMART rehabilitation system employs an activity monitoring system linked to a decision support platform that provides therapeutic instruction, supports the rehabilitation process and monitors the effectiveness of rehabilitation interventions on patient function (*Level 5*). Information relating to this process is fed back to patients, their carers and/or health care professionals.

The system consists of three primary components; a motion tracking unit, a computer base station and a web-server shown in Figure 2. The motion tracking unit provides the ability to monitor a patient's upper limb during exercise such as reaching and drinking. The motion information, which includes both position and rotation, is transmitted via Bluetooth to a base station (*Level 1*) for visualisation and analysis.

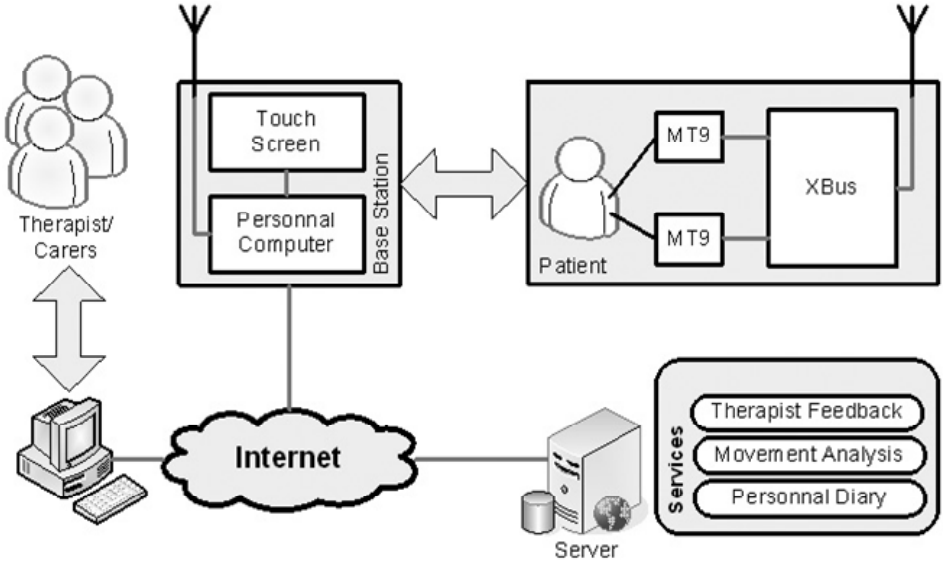


Figure 2: SMART System Architecture

(*Level 3*). The decision support platform (*Level 4*) operating at the base station provides various types of feedback to user, such as movement display, exercise history and variable measurement results. A personal computer and touch screen monitor make up the base station which is connected to a central server providing web services to the user. Patients and their carers can view their activities in a three dimensional (3D) environment on the ICT platform, and can communicate with their therapists using the web services (*Level 1*).

2.3. Motion Tracking Wnit

Various technologies have been applied in the area of motion tracking and rehabilitation:

- Switches: pedometer and functional electrical stimulation (FES);
- Video systems;
- Gyroscopes (angle / rate of turn);
- Accelerometers (velocity);
- Robot arms.

In the SMART project, inertial sensors (Xsens Dynamics Technologies, Netherlands, known as MT9) were used to track patients' upper limb movements. The MT9 sensor consists of a three axis accelerometer, a three axis gyroscope, and a three axis magnetic field sensor, which can be used to measure rotation rates and linear

accelerations. For a more comprehensive review of accelerometers and general motion sensing techniques see [31].

As illustrated in Figure 3, two MT9 sensors may be attached to a patient's upper limb via wearable harnesses to record movement during daily activities. The MT9s record movement information, such as positional and angular data relating to upper limb joints - the wrist, elbow and shoulder. The data are then transmitted wirelessly to the base station via a digital box called the XBus which is worn on a belt around the patient's waist. The tracking unit utilises sensor fusion and optimisation techniques and is implemented using Visual C++, based on a PC with a 1.2GHz CPU (*Level 2*). Motion data are generated continuously with a sampling rate of 25Hz.

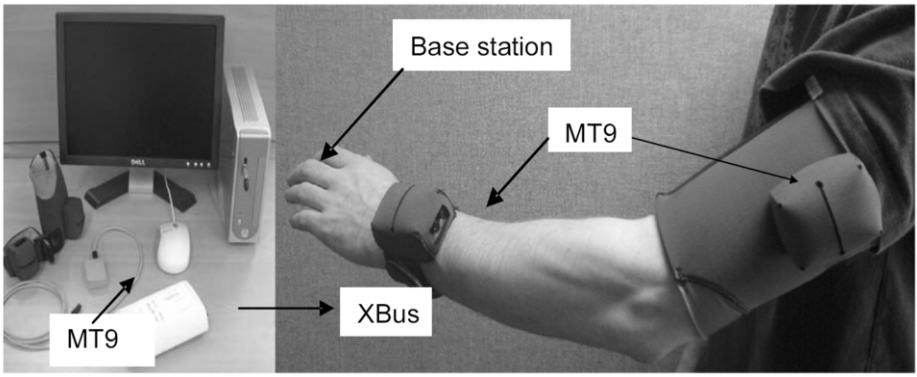


Figure 3: Accelerometer based motion tracking unit in Smart rehabilitation system with embedded processing

2.4. ICT Decision Support Platform

The ICT platform consists primarily of five modules: database, interface, decision support, communication and a feedback module.

The database module (*Level 2*) maintains information such as system information, personal details, rehabilitation movement parameters; comments and instructions from healthcare professionals and movement analysis. The system information stores a series of questions designed to ascertain the patient's ability to safely complete the daily exercises. A calendar service stores a patient's entire history of rehabilitation. The database module also provides access to relevant personal information on the patient.

The design of the interface module was largely informed by the fact that users would interact with the software via a touch screen display. Its aim is to provide easily accessible and understandable menus. One such menu is illustrated in Figure 4 allowing the user to start an exercise, view their calendar and diary and obtain some feedback on their rehabilitation progress.

The decision support module (*Level 4*) analyses the movement data and provides key outcome variables relating to physical performance such as length of reach, elbow angle. The communication module manages the transfer of information between the base station and the central server. The feedback module provides different types of

information to patients, namely 3D movement information, comments and instructions, and analysis results presented as a 3D visualisation.

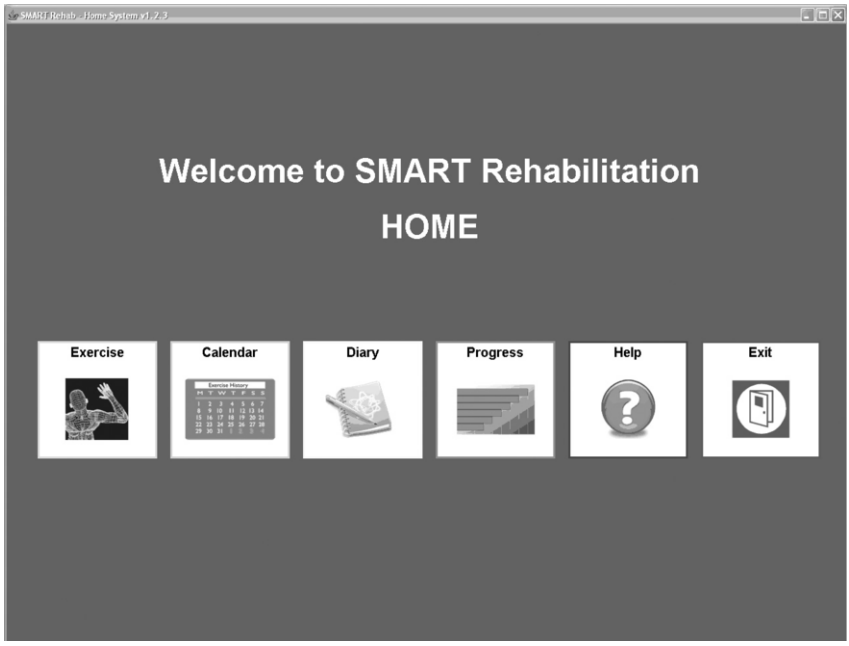


Figure 4: Welcome screen after user login. Menus are designed using large icons to accommodate for touch screen technology, and promote usability.

The visualisation replays the movement of rehabilitation exercises to users in a 3D environment using a virtual body and arm. Stored movement templates can be overlaid or mirrored on the screen as a reference to help the patient improve the exercise. The target movement template is personalised and adaptive, and is selected by the therapist according to the patient's rehabilitation progress (*Level 5*). Figure 5 displays an upper-limb arm movement during a reach exercise.

2.5. Web-Based System

All the SMART home-based systems are linked to a central server via the public Internet, using a virtual private network. Patients' movement data and comments or queries can be sent to the central server. A therapist can log into the central server to view his/her patients' movement data and their comments, provide instructions after analysing their exercises and provide feedback on their queries.

There are three processes that run on the server to manage the system; MySQL database server to control the storage of data, Apache TomCat web server (see section 2.3) to deal with Internet traffic and Matlab to provide a tool for analysing the movement data (*Level 3*).

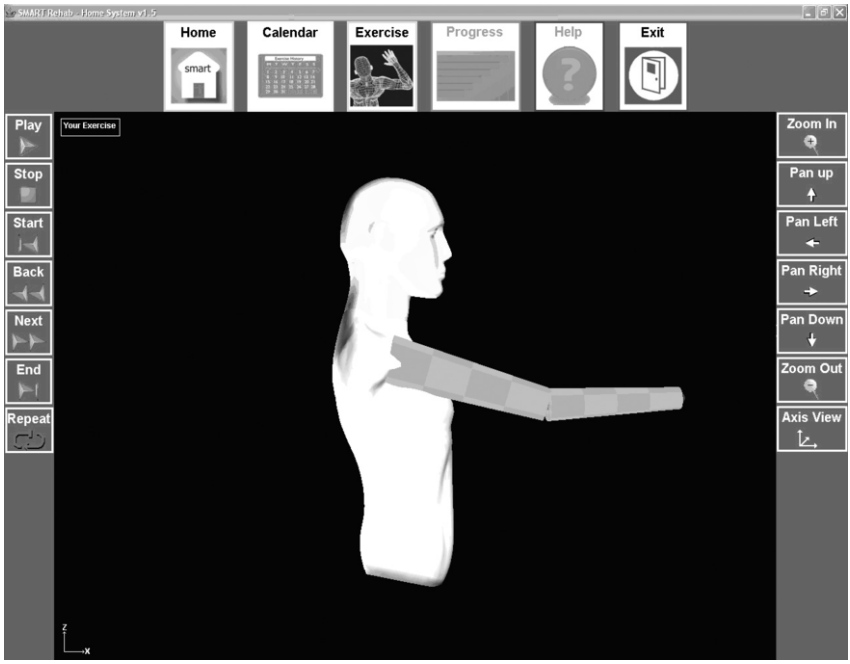


Figure 5: Software screen shot of 3D rendering providing personalized feedback to promote rehabilitation

2.6. User centered Design Strategy

The user groups involved in the rehabilitation system are post-stroke patients, their carers, and therapists. The system must be as simple as possible to use, and adaptable to individual needs. Special care must be taken with people who have had a stroke as they have complex impairments often incorporating cognitive difficulties such as problems with perception, attention, information processing, language and memory. In order to provide a user friendly rehabilitation system, the user-centered design strategy was applied throughout the projects development.

In the early stages of the project, focus groups were held with patients and healthcare professionals to ensure that proposed technical solutions, methodology and outputs were acceptable. In the later stages, a group of expert users provided feedback on key aspects of the system such as user interface, type of feedback and computer interface. These data were collected by qualitative researchers, summarised and fed back to the engineering teams.

3. Summary

In this review, we have examined computers and communication which are drivers for medical informatics and the eHealth paradigm. A conceptual model for medical informatics, proposed by van Bemmel, is applicable to eHealth. Having provided the

background on the technology, we use a case study of home-based tele-rehabilitation to illustrate the potential of eHealth in an era when conventional interventions for chronic conditions will become increasingly more difficult to support. The case study provides examples of hardware, software, secure communication over a network and a 3-tiered client server Internet based approach. It comprises many of the model sub components and indeed its purpose was to research the area of tele-rehabilitation and stroke (*Layer 7*), in order to provide new knowledge that can be shared with the multi-disciplinary team comprising biomedical engineers, computer scientists and healthcare professionals.

Acknowledgement

The SMART consortium consists of Royal National Hospital for Rheumatic Diseases, University of Bath, Sheffield Hallam University, University of Essex, University of Ulster and The Stroke Association. We would like to thank the patients and carers, who have given their time and support to the project. We would also like to thank Design Futures from Sheffield, Xsens, Charnwood Dynamics Ltd. and the Stroke Association for their support. SMART was funded under the EQUAL 4 (extend quality of life) initiative from the UK Engineering and Physical Sciences Research Council (EPSRC).

Glossary

3G	Third generation cellular telephony
ADSL	Asymmetric Digital Subscriber LIne
ALU	Arithmetic and Logic Unit
Bluetooth	Personal area network standard
CERN	Conseil Européen pour la Recherche Nucléaire
CPU	Central Processing Unit
DSS	Decision Support System
ECG	Electro-CardioGram
eHealth	Electronic Health
EPR	Electronic Patient Record
GPRS	General Packet Radio Service
GSM	Global System for Mobile communication
HCI	Human Computer Interface
HSPDA	High Speed Packet Download Access
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ICT	Information and Communication Technology
IEEE	Institution of Electrical and Electronic Engineers
IIS	Internet Information Services
I-Mode	Internet service designed for mobile phones
IrDa	The Infrared Data Association
IS-95	Cellular telephony system in USA
ISM	Industrial Scientific and Medicine
KB	Knowledge Base
LAN	Local area network

MAC	Message Authentication Code
MySQL	An Internet database system
NHSnet	UK's National Health Service private network
OFDM	Orthogonal Frequency Division Modulation
OS	Operating system
PDA	Personal Data Assistant
PSTN	Public Switched Telephone System
PUBMED	Publicly accessible database from US National Library of Medicine
RDBMS	Relational Database Management System
SMS	Short Message Service
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TCP/IP	Transport Control Protocol/ Internet Protocol
Tomcat	A web based server from Apache Corporation
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
USB	Universal Serial Bus (connectivity interface for PC)
VOIP	Voice over internet protocol
WAP	Wireless Access Protocol
WiFi	Wireless Fidelity, a wireless local area network
WWW	World Wide Web
XBus	Open Source routing and transformation system for connected devices
XML	eXtensible Markup Language
ZigBee	Personal area network technology

References

[1] D.L. Sackett, W.M.C. Rosenberg, J.A. Muir Gray, R.B. Haynes, W. Scott Richardson, Evidence based medicine: what it is and what it isn't. *BMJ* 312, (1996), 71-72.
[2] J.H. van Bemmel, The structure of medical informatics. *Med Informics* 9, (1984), 175-180.
[3] P. Wilson, C. Leitner, A. Moussalli, Mapping the Potential of eHealth: Empowering the citizen through eHealth tools and services. European Institute of Public Administration 2004.
[4] 4D Ultrasound, http://www.4dfetalimaging.com/, accessed Oct 2007.
[5] R. Morin, S. Langer. Speech Recognition System Evaluation. *Journal of the American College of Radiology* 2(5), 449-451, May 2005.
[6] Connecting for health, http://www.connectingforhealth.nhs.uk/ accessed Dec 2007.
[7] S. Kotzin, Medline and PubMed will be able to synthesise clinical data *BMJ*. 324 (2002), 791.
[8] PUBMED, National Institute of Health, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed, accessed Oct 2007.
[9] J.H. van Bemmel and M.A. Musen (eds.), Handbook of Medical Informatics, Springer 1997.
[10] W. Aspray, John Von Neumann and the Origins of Modern Computing. Cambridge, Massachusetts: The MIT Press, 1990.
[11] G.E. Moore, Cramming more components onto integrated circuits. *Electronics Magazine* 19 April 1965.
[12] R. Kurzweil, The Law of Accelerating Returns, http://www.kurzweilai.net/articles/art0134.html?printable=1, March 2001.
[13] Los Alamos National Laboratory, http://www.nigms.nih.gov/News/Results/FluModel040306, accessed Oct 2007.
[14] M.L. Baker Supercomputer Serves as Weapon in AIDS Fight, http://www.eweek.com/article2/0,1759,1972257,00.asp, accessed Oct 2007.
[15] Health Grids, http://www.healthgrid.org/, accessed Oct 2007.
[16] Biomedical Informatics Research Network, http://www.nbirn.net/, accessed Oct 2007.

[17] http://www.malariacontrol.net/, accessed Dec 2007.
[18] PDA software available from www.tucows.com, accessed Dec 2007.
[19] Apache web server project, http://httpd.apache.org/, accessed Dec 2007.
[20] EU Health portal, http://ec.europa.eu/health-eu/index_en.htm, accessed Dec 2007.
[21] http://www.stroke.org.uk/noticeboard/obesity.htm, accessed Oct 2007.
[22] M. Walker, Stroke rehabilitation. *The British Journal of Cardiology*, 9(1) (2002) 23-30.
[23] DoH (2000), National Service Framework for older people. London.
[24] J.M.Winters, Telerehabilitation research: emerging opportunities, *Annu. Rev. Biomed. Eng.* 4, (2002), 287-320.
[25] H.R.Miltner, H. Bauder, Effects of constraint-induced movement therapy on patients with chronic motor deficits after stroke, *Stroke* 30(1999), 586-92.
[26] P.M. Rossini, C. Calautti, Post-stroke plastic reorganisation in the adult brain. *The Lancet Neurology* 3(2003), 493-502.
[27] Natl. Inst. Disbil. Rehabil. Res. Request for applications for Rehabilitation Engineering Research Center on Telerehabilitation. 1998. Fed. Regist. June 12, 32526-39.
[28] H. Zheng, N.D. Black, N. Harris, Position-sensing technologies for movement analysis in stroke rehabilitation, *Medical & Biological Engineering & Computing* 43(4) (2005), 413-420.
[29] SMART PROJECT website, (http://www.shu.ac.uk/research/hsc/smart/), accessed December 2006.
[30] G.A.Mountain, P.M.Ware, J. Hammerton, S.J.Mawson, H. Zheng, R.Davies, N.D. Black, H. Zhou, H. Hu, N. Harris and C. Eccleston, The SMART Project: A user led approach to developing and testing technological applications for domiciliary stroke rehabilitation, 3rd Cambridge Workshop on Universal Access and Assistive technology (CWUAAT 2006).
[31] H. Zhou and H. Hu, Inertial motion tracking of human arm movements in home-based rehabilitation, *Proceedings of IEEE Int. Conf. on Mechatronics and Automation*, Ontario, Canada, 29 July – 1 August, 2005, 1306-11.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-140

III.2. Data Structures, Coding and Classification

Huiru ZHENG[a], Haiying WANG[a], Norman D. BLACK[a] and John WINDER[b]
[a]*Faculty of Engineering, University of Ulster at Jordanstown, Co. Antrim, N Ireland*
[b]*Faculty of Life and Health Sciences, University of Ulster at Jordanstown, Co. Antrim, N Ireland*

Introduction

Medical data are at the centre of many healthcare-related activities, ranging from providing medical care for each individual patient to supporting scientific research and education. This review offers an introduction to medical data structures and management. Examples of coding and classification systems are also presented.

1. Medical Data

Medical data include a broad range of data types, including text-based alphanumeric data, recorded physiological signals and medical picture/images. In addition to traditional clinical data, recent progress in medical science and technology has led to the accumulation of a tremendous amount of genomic data in support of individualised healthcare and personalised therapies. In this section, the characteristics of typical clinical data are introduced, followed by a brief introduction to genomic data.

1.1. Clinical Data

In general, there are three types of clinical data in the practice of medical and health science 0, [2].

- Alphanumeric data are typically recorded as unstructured, free text in the medical record. They account for a large portion of the information that is gathered verbally in the care of patients, such as patient demographics, a description of patients' illnesses and responses to physician's questions. Results of physical examination such as vital signs and laboratory tests, and other narrative descriptions, such as treatment plans and surgical procedures, also belong to this category.
- Physiological signals, captured from the human body by using appropriate biomedical sensors, are usually represented by a series of values along two dimensions with the x-axis representing time and the y-axis displaying voltage values associated with the underlying physiological activity. Advances in physiological signal acquisition techniques have generated a vast quantity of

physiological recordings, including electrocardiogram (ECG), electroencephalogram (EEG), gait, respiration, blood pressure tracings, and oxygen saturation. As illustrated in Figure 1, an ECG records the electrical activity of the heart as a function of time. Each ECG beat represents one cardiac cycle. By convention it includes the P wave, the QRS complex. T wave, and sometimes a small U wave may be seen following the T wave.

- Medical images, as illustrated in Figure 2. With the ability to reveal the areas of interest in great detail, medical images are a valuable resource in clinical diagnosis and medical research. Radiological images based on X rays, computed tomography and magnetic resonance imaging techniques are the most well known examples. Using the electromagnetic spectrum, radiological images provide a means to visualize the condition of internal organs or an area that is not externally visible. Light images may also be produced by skin photography and endoscopy. Instead of using electromagnetic radiation, ultrasound imaging is based on mechanical energy in the form of high-frequency sound waves.

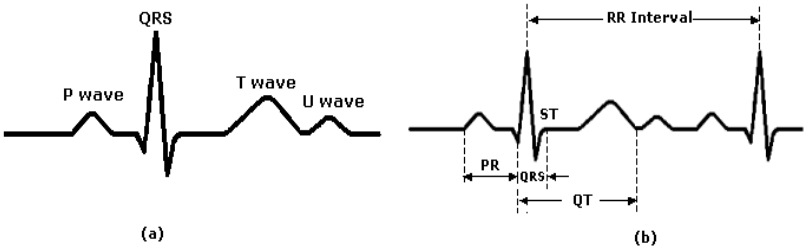


Figure 1. Illustration of an ECG signal: (a) a typical ECG wave, (b) ECG intervals. RR, QT and PR represent the intervals between two respective wave peaks.

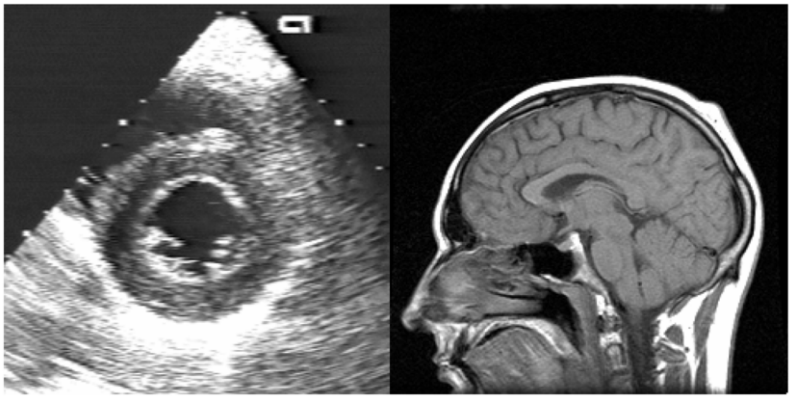


Figure 2. An example of ultrasound images and Magnetic Resonance (MR) images.

1.2. Genomic Data

Advances in biological science are fostering a new clinical practice, where clinical diagnosis and treatment will be supported by the information encoded in relevant genomic data, such as DNA (deoxyribonucleic acid) sequences, protein structures and microarray gene expression data. Examples of genomic applications in medical research can be found in the Programs for Genomic Applications (PGAs) funded by the National Heart, Lung and Blood Institute (NHLBI) of the NIH (National Institutes of Health) (http://www.nhlbi.nih.gov/resources/pga/).

Each type of genomic data has a unique data structure. For instance, DNA may be represented by a linear sequence from the alphabet {A, G, T, C}. Each of these characters refers to a nucleotide. A typical microarray experiment dataset includes expression levels of thousands of genes in a number of experimental samples (conditions). These samples may correspond to different conditions or serial time points taken during a biological process. An expression dataset can be summarised by a matrix, in which the horizontal rows represent genes, one gene in each row, and the vertical columns contain the various samples corresponding either to the time points or to various conditions, with one sample in each column. Thus, various models of electronic medical records have been developed to integrate traditional clinical data and diverse genomic data 0.

The strong interest in integrating genomic data into clinical research is driven by the hope that it will result in the prevention and diagnosis of complex genetic disease and the design of highly targeted therapies [4]. For example, with DNA microarray technology, scientists are able to measure gene expression of thousands of genes under selected physiological conditions at a given time. Thus, by comparing expression profiles of an individual gene under various conditions or by comparing expression profiles of samples to understand their similarities and differences and to find the relationships among genes and samples, many important physiological and medical phenomena can be studied at the molecular level.

1.3. Characteristics of Medical Data

Apart from privacy and confidentiality requirements, medical data have important and unique features. For example, a patient record is a mixture of opinion and factual data generated from a wide range of clinical professionals. Some of these data, such as observations and interpretations, are difficult to record in a structured form. Recorded physiological signals and medical images are contaminated by noise generated from the machine recording the data or from the patient. For instance, an ECG record could be contaminated by physiological variability, baseline wander noise, 50 or 60 Hz power line interference, and muscle noise 0. Due to the quantum noise inherent in photo detection, images generated by computed tomography-based techniques such as positron emission tomography (PET) and single photo emission tomography (SPECT) have a high noise content 0.

2. Data Structures and Management

The characteristics of medical data pose a great challenge to efficient data storage and management. In this section, electronic patient records are described. The examples of

management and storage of medical images and physiological signals are introduced respectively.

2.1. Electronic Patient Records

A patient record contains all the clinical data relevant to the care of one patient. Given the increasing volume and scope of medical data available nowadays, traditional paper-based patient record systems are not sufficient to address many issues related to efficient medical data management. The Electronic Patient Record (EPR) is revolutionising storage, management and communication of medical information. With EPR comes all related medical data, including patient history, physical examination, drug prescriptions, and recorded biological signals, linked into one record that can be easily accessed for different purposes. The ability to store all patient-related medical information in one record, which may be accessed at any time or from any place, ultimately contributes to improvements in healthcare delivery.

It has been shown that an EPR-based system offers several significant advantages in the process of data management and analysis 0]. One of the most obvious benefits is that the speed of data retrieval is significantly improved. A physician should be able to search through thousands of records for specific attributes such as medication or diagnosis with little effort. With structured data entry, medical data collected in an EPR are well structured and organised, allowing the electronic sharing of patient recording among different clinical professionals. Moreover, an EPR-based record system has the inherent flexibility to provide task-specific views on all the available patient data to meet the needs of different clinical tasks, as illustrated in Figure 3.

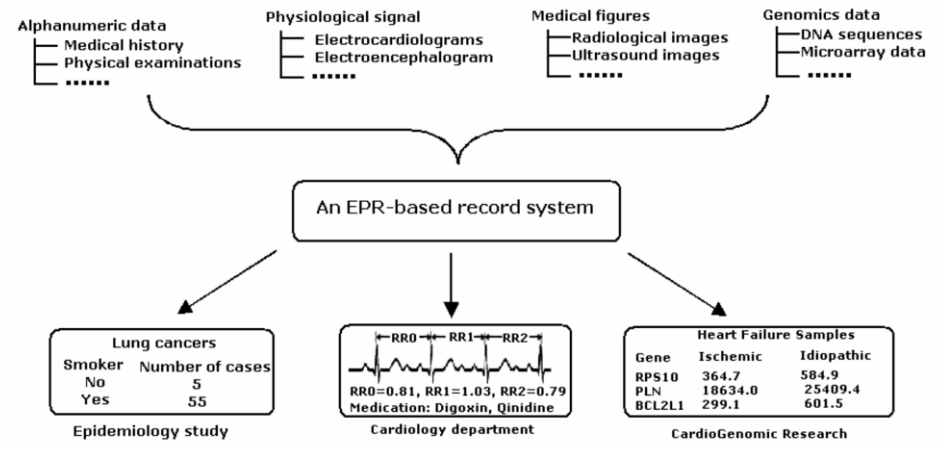


Figure 3. An EPR provides an integrated view of patient data and has the flexibility to generate task-specific views on all the available data to meet the needs of different tasks.

One of the key components in an EPR is a database management system (DBMS) 0. DBMS is a software system designed to manage the storage, retrieval, and organization of data in a database. Different data models have evolved in recent years to structure the logical view of the DBMS, including relational, hierarchical, text-

oriented, and object-oriented data models. Among these, the relational data model is most widely used.

A relational data model, developed by E.F. Codd [8], is one in which the data and relations between them are organised into a series of tables. Each table is made up of columns and rows. All values in a given columns have the same data type, which could be numeric or textual. All values in a row conceptually belong together. Typically, each row, known as a record, is uniquely identified by a primary key which may contain a single column or multiple columns in combination. Figure 4 shows an example of a relation database with two tables: *PatientDemographics* and *PhysicalExamination* tables.

PatientDemographic table:

PatientID	First Name	Second Name	Gender	Date of Birth	Phone Number
17	Edward	Garber	Male	03/11/1956	(44)2892321123
35	Sara	Gardner	Female	21/03/1972	
5	Alan	Barnett	Female	04/05/1992	
102	Suzan	Coronel	Male	29/12/1942	(86)59423234121

1:n

PhysicalExamination Table:

TestID	Date	Weight (kg)	Height (cm)	HeartRate (BPM)	BloodPressure (mmHg)	PatientID
03	01/05/2006	68	171	75	120/65	17
07	02/05/2006	61	168	65	110/62	35
11	03/07/2006	67	171	70	110/60	17
13	04/07/2006	45	135	80	110/74	5

Figure 4. Example of a relational database with two tables.

2.2. Medical Images - DICOM

Digital Imaging and Communications in Medicine (DICOM) is the healthcare industry standard for medical image data file storage and transfer. It was first conceived in 1983 by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) and first known as the ACR-NEMA standard. This standard was revised in 1993 and named DICOM. The standard defines principles and methods for the storage and transfer of medical images (mainly radiological) in a multi vendor environment (http://medical.nema.org). Virtually all radiological imaging modalities are now digital and a large teaching hospital may generate around 2-3 Tera (10^{12}) bytes of image data per annum. The use of medical imaging is increasing, as is

the resolution of imaging systems, therefore we can expect to see the computer network and storage requirements also increase.

The main DICOM functions are to transmit and store medical images, to enable searching and retrieval, to facilitate printing of hard copy, to support workflow management and to ensure quality and consistency of image presentation 0. The DICOM file format is well defined and contains two specific sections. The first part of the file, called the header, holds patient demographics, hospital information and details of the referring physician. Also, the header contains detailed data on the actual specific imaging modality parameters for each patient scan. For example, in a magnetic resonance (MR) image data file, the slice thickness, field of view, pulse sequence details, table position, and scan date are stored. The header also contains hidden fields used by the equipment vendor to store proprietary information on image reconstruction algorithms. Another feature of the DICOM header is that it contains a unique identification number that identifies the scanner manufacturer, the individual scanner, the modality types and the patient study number. This is called the unique identifier (UID). DICOM data files written to storage media like DVD or to another computer will be named using the UID as part or the entire file name.

Figure 5 shows a screen shot of part of a DICOM image header from OSIRIS developed by The University Hospital of Geneva, Switzerland. Osiris is a free DICOM viewer available at http://www.sim.hcuge.ch/osiris/01_Osiris_Presentation_EN.htm. The second part of the DICOM data file is the image pixel data where pixel values are stored using between 4 and 32 bits, depending on the imaging modality. A typical file size for MR is 0.5 MBytes (one patient examination may contain up to 1000 images), whilst a Computed Radiography chest X-ray may be 32 MB. The DICOM standard includes the functionality to compress medical images using lossless or "lossy" techniques 0. JPEG 2000 is the common standard achieving a lossless compression of approximately 3:1. Further compression is possible, however, there is some reluctance to reduce file size further as diagnostic integrity of the original image may be lost after significant compression.

The benefit of DICOM is the ability to connect a number of different vendor medical imaging modalities to a Picture Archive and Communication System (PACS). This is the main image database and transmission system in the health environment. This system enables multiple copies of images to be available at any time and at any place that has the appropriate network connections and permissions. PACS also enables the creation of large digital libraries for teaching and research.

2.3. SCP-ECG

The SCP-ECG (the Standard Communications Protocol for Computer-Assisted Electrocardiography) was developed by the Comité Européen de Normalisation Technical Committee 251 (CEN/TC251) project team in 1993 exclusively for the purpose of the exchanging, encoding and storage of 12-lead ECGs 0. The data level in the SCP-ECG standard includes the ECG signal data, patient demographics, ECG measurements, ECG interpretation results, as well as schemes for ECG waveform compression. In the SCP-ECG standard, various parts of data related to an ECG are specified in different data sections with different encoding forms as illustrated in Figure 6.

MR Acquisition Type
2D
Repetition Time
22.212000
Echo Time
3.750000
Echo Train Length
0
Number of Averages
2.000000
Imaging Frequency
63.907560
Imaged Nucleus
1H
Echo Number
1
Magnetic Field Strength
1.500000
Spacing Between Slices
11.000000
Number of Phase Encoding Steps
192
Percent Sampling
75.000000
Percent Phase Field of View
100.000000
Low R-R Value
0
High R-R Value

Figure 5. Typical information stored in the DICOM file header.

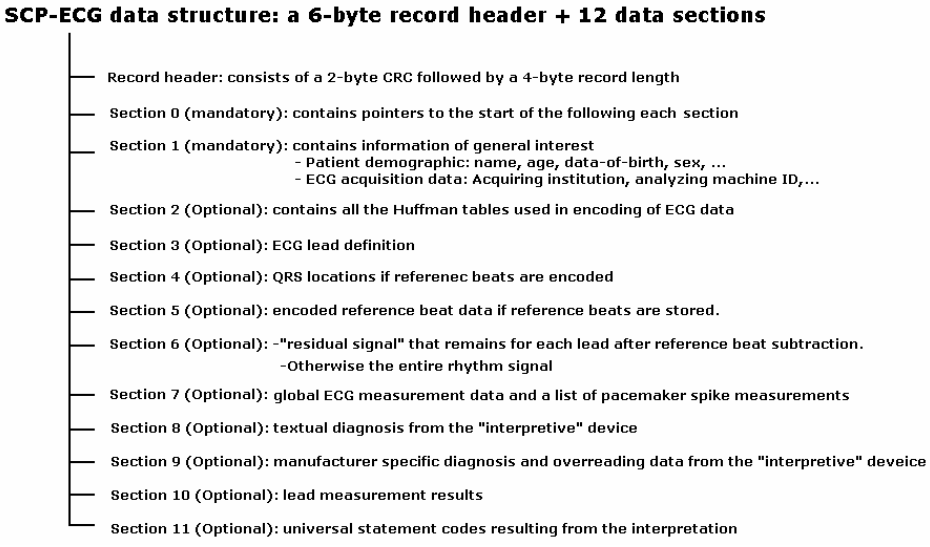


Figure 6. Structure of a SCP_ECG record

2.4. XML based systems

Due to its ability and flexibility to define markups for specific type of data, XML (eXtensible Markup Language, http://www.w3.org/xml/) and its related techniques have gained great attention in medical data representation and management. Currently, various organizations within healthcare such as HL7 (health level 7) and ASTM (American Society for Testing and Materials) are working on recommendations of XML-based e-healthcare technologies within the medical domain. For example, HL7 has published the Clinical Document Architecture (CDA), which uses XML as the representation format for medical and administrative data (http://xml.coverpages.org/ni2004-08-20-a.html). Along with the CDA, HL7 version 3 messaging standard has also been developed using XML as the standard exchange format. The use of XML syntax for the exchange of electronic patient records was illustrated in the EU project Synapses (https://www.cs.tcd.ie/synapses/public/) and its implementations [12]. The application of XML in DICOM can be found in several related projects [13]. The *U.S Food and Drug Administration* (FDA) Centre for Drug Evaluation and Research has proposed recommendations for the XML-based techniques for the exchange of time-series data. In December 2003, the *FDA XML* format was finalized and the Annotated ECG (aECG) format is now part of the HL7 family of standards (http://www.hl7.org/v3annecg/foundationdocuments/welcome/index.htm).

3. Coding and Classification

Clinical concepts can be represented and described in many ways. Consequently, standards for codes/terminology have become an essential element in the development of electronic patient records. In this section, several widely-used coding and classification systems are introduced.

3.1. The International Classification of Diseases

The International Classification of Disease (ICD) is a coding system published by the World Health Organization (WHO, http://www.who.int/en/). Its main purpose is to allow morbidity and mortality data to be collected from around the world in a standard format. Since its first edition was published in 1900, ICD has become the most widely used pathology classification system and *de facto* reference point for many healthcare terminologies 0. The code has been updated approximately every 10 years to reflect the medical advances in the past years and the current version is ICD-10.

The classification of ICD-10 is structured into 21 chapters, identified by 21 Roman numerals: I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XVII, XVIII, XIX, XX, and XXI. Among these, Chapters I-V, XV-XVII and XIX cover special diseases, such as neoplasms, pregnancy, childbirth and the puerperium, and injuries, poisoning and certain other consequences of external agents. Diseases associated with body systems are found in Chapters VI to XIV. For example, Chapter VI covers diseases of the nervous system, and diseases of the genitourinary system are assigned to Chapter XIV. Symptoms, signs and abnormal clinical and laboratory findings are given in Chapter XVIII. Other external factors affecting morbidity and mortality are included in Chapters XX and XXI.

Each chapter is divided into blocks and categories. Unlike ICD-9, in which diseases are coded with three digits, category codes in ICD-10 start with an alphabetical character followed by two numeric characters, ranging from A00 to Z99. For instance, code I01 represents rheumatic fever with heart involvement. This greatly increases the number of available codes, which can be used to code common diseases in more detail.

Most 3-character categories in ICD-10 are further subdivided into up to 10 subcategories by adding a fourth digit after a decimal point. This can be used, for instance, to classify varieties of a disease, e.g. as illustrated in Figure 7, *acute rheumatic pericarditis* is coded with I01.0, while *acute reheumatic endocarditis* is represented by I01.1.

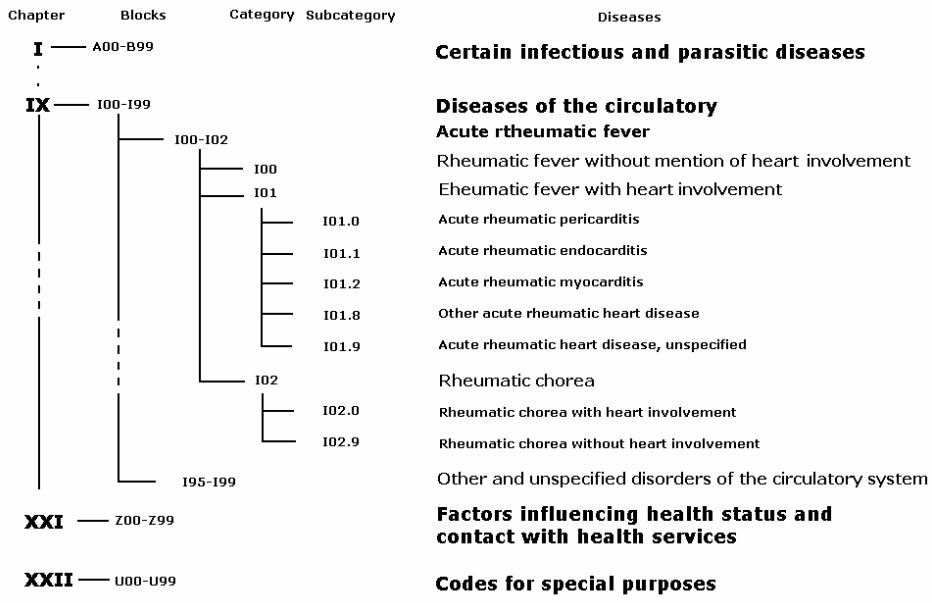


Figure 7. Example of ICD-10 code.

3.2. The Clinical Terms (Read Codes)

Read Codes were initially introduced into the UK in the middle 1980s as a 4-byte set terminology to summarise patient treatment, including clinical process and administrative data, in primary care. This version only consists of about 40000 codes. Its subsequent Version 2 released in 1990 was based on the content of ICD-9 and Classification of Surgical Operations and Procedures, Fourth Revision (OPCS-4). Read Codes in this version were extended to 5-digit alphanumeric characters, each representing one level of hierarchy. While being easy to understand and implement, such a rigid hierarchical classification structure poses some difficulties in the

representation of clinical terms. For example, without giving two Read Codes, a clinical term can only be located within one hierarchy.

Due to increasing demands from clinical professionals, Read Codes were further expanded by the Clinical Terms Project, a joint project established by the UK's Conference of Medical Royal Colleges and the government's National Health Service (NHS). As a result, the Clinical Term Version 3 (CTV3) was developed and first released in 1994.

In CTV3, a concept refers to any clinical disorder, procedure or observation and can be described by more than one clinical term. Like its previous version, CTV3 adopts a 5-digit alphanumeric code, giving more than 650 million possible codes to represent clinical concepts. However, CTV3 implements a new, more flexible structure to address the problems caused by its early version. For example, the coded-based hierarchical structure has been abandoned. A Read Code is only used to uniquely identify a concept and no longer represents its hierarchy position. Instead, hierarchical relationships between concepts are defined by a set of parent-child links described in a Hierarchy file.

Core release files included in CTV3 are Concept file, Term file, Descriptions file, Hierarchy file, key file, Specialty files, Cross-mapping files, Template file, Redundant codes mapping file and Description Change file. The structures and examples of Concept file (describing concepts), Term file (describing terms), Descriptions file (linking terms to Read Codes) and Hierarchy file (describing a hierarchy) are illustrated in Figure 8. The reader is referred to [14] for a more detailed description of CTV3.

3.3. Systematized Nomenclature of Medicine

The Systematized Nomenclature of Medicine (SNOMED), developed and maintained by the College of American Pathologists, is generally regarded as one of the most comprehensive standardized medical coding systems available today (http://www.snomed.org/). It uses a hierarchical, multi-axial method to encode clinical data, allowing for the representation of various aspects of a disease. For example, SNOMED International is organized around 11 independent axes: T (Topography), M (Morphology), L (Living organisms), C (Chemical), F (Function), J (Occupation), D (Diagnosis), P (Procedure), A (Physical agents, forces, activities), S (Social context), G (General syntactic). Each axis represents a unique hierarchical classification system. Thus, a disease in SNOMED can be described using a morphologic code, a topographic code, an etiological code, and a function code. Moreover, these codes across different axes can be cross-referenced, leading to a better understanding of each code. For example, the disease pneumonia could be coded as T-28000 (topology code for lung), M-4000 (morphology code for inflammation) and L-25116 (etiological code for Streptococcus pneumoniae).

To reflect the advances in medical informatics, SNOMED has evolved further with the release of SNOMED RT (Reference of Terminology) in 2000. To provide a unified international language that supports the electronic patient record and decision support system, the College of American Pathologists and the UK NHS announced their intention to unite SNOMED RT and CTV3. As a result, SNOMED CT (Clinical Term), which incorporates the content and structure of CTV3, was released in 2002.

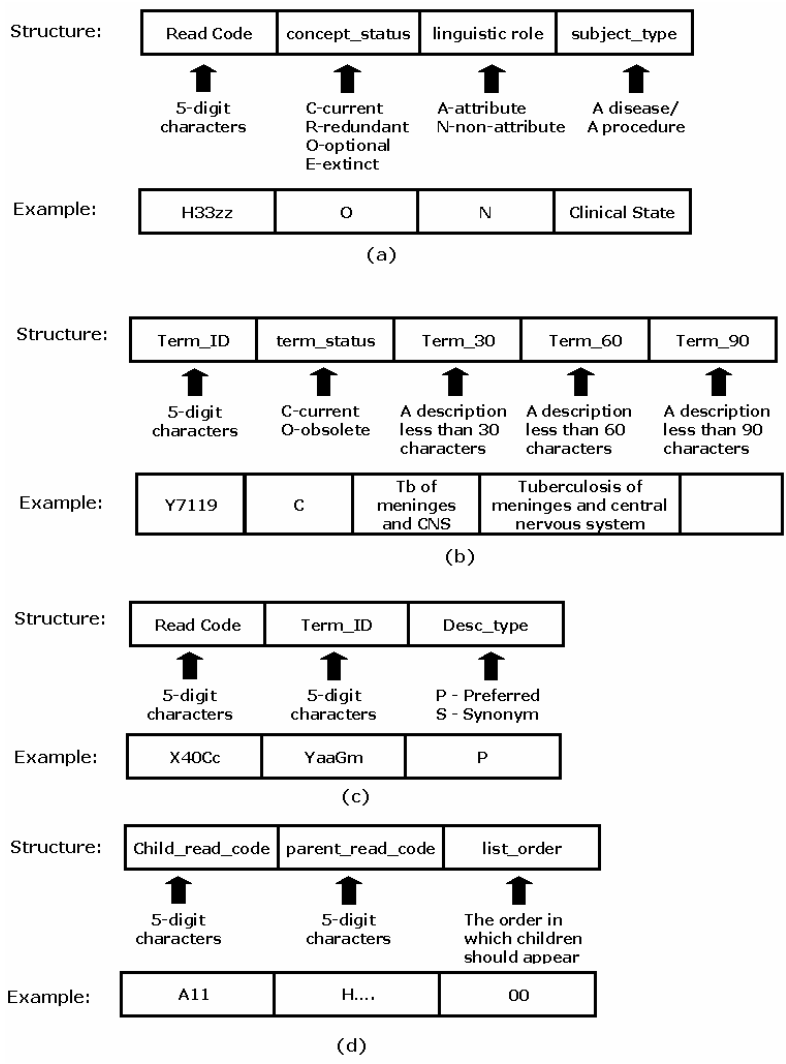


Figure 8. The structures and examples of (a) Concept file; (b) Term file; (c) Descriptions file; and (d) Hierarchy file.

One of the important features of SNOMED CT is that it utilizes description logic to describe the scope of a concept. Its core content includes concepts, descriptions and the relationships between them 0. As illustrated in Figure 9, a concept in SNOMED CT can be described by terms in one or more descriptions. SNOMED CT contains more than 300000 concepts with unique meanings and formal logic-based definitions, which are organized into top-level hierarchies such as *Clinical findings* and *Body Structure*. The relationships between concepts are described in relationship tables. There are two types of relationships in SNOMED CT. While *IS-A relationships* connect concepts within the same hierarchy, *Attribute relationships* describe the relationships between concepts located in different hierarchies. For example, the relationships between the

disease concepts: *Arthritis of knees*, *Arthritis* and *Arthropathy* belong to *IS-A relationships*. Examples of *Attribute relationships* in SNOMED CT include *Finding site*, *Procedure site*, *Associated morphology*, and *Method*. The relationship between the concepts *Appendicitis* that is a *disease* concept and *Inflammation* that belongs to the *Body structure* hierarchy can be described using *Associated morphology*.

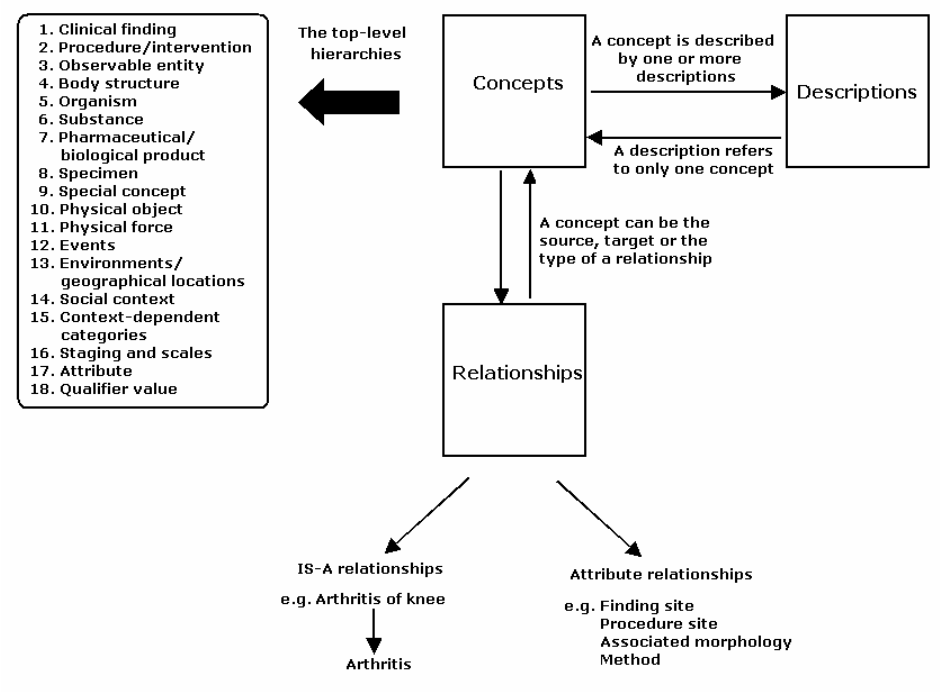


Figure 9. The outline of SNOMED CT core structure.

The current version of SNOMED CT includes more than 300,000 clinical concepts, 770,000 descriptions and 900,000 semantic relationships. SNOMED CT can also map to other medical classification systems such as ICD-10.

4. Case Study: An XML-based representation of ECG data

The ecgML model, a markup language for supporting representation and coding of ECG data, was first published in 2003 [15]. Based on XML, ecgML offers several advantages over existing ECG coding systems. For example, it includes all the components required for ECG representation. The terms and structure used can be further expanded or reviewed at any stage. This protocol can be used as a common coding platform between different ECG acquisition devices and visualization programs.

4.1. Hierarchical Presentation of ecgML

The main components included in ecgML are one *PatientDemographic*, optional element *MedicalHistory*, one or more *Record* components, and an optional *Diagnosis* element, as illustrated in Figure 10. To facilitate the inclusion of multiple time-related patient's ECG data, each *Record* element, which consists of zero-or-one *RecordingDevice*, zero-or-one *ClinicalProtocol*, and one-or-more *RecordData*, is uniquely identified by *AcquisitionDate* and *AcquisitionTime*. As a key component in ecgML, each RecordData includes three main subcomponents (*Waveforms*, *Annotations* and *Measurements*) to represent original ECG waveform data, the annotations and measurements respectively.

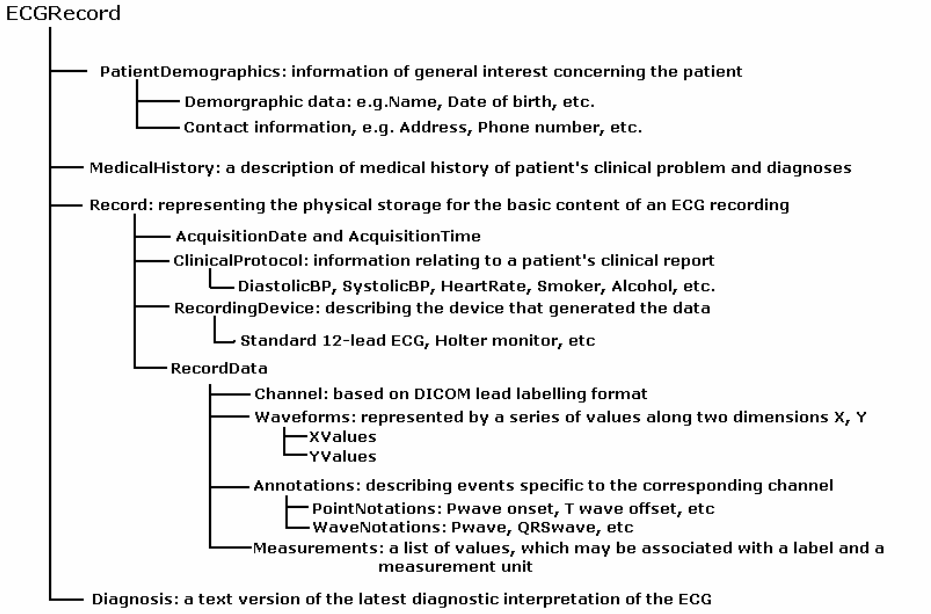


Figure 10. The main structure of the ecgML model.

ecgML incorporates several international standards to describe relevant information. For example, it applies the DICOM lead labeling format to define channel names associated with *RecordData* element. The *Unified Code for Units of Measure* (UCUM) scheme is used to define measurement units when appropriate. Moreover, different coding and classification schemes can be incorporated to define medical terms at different levels. One example is to utilize terminologies included in SNOMED CT to describe the diagnostic interpretation of the ECG (*Diagnosis* element) and medical history of patient's clinical problems (*MedicalHistory* element).

A portion of an ecgML-based ECG recording generated from the MIT-BIH Arrhythmia Database is illustrated in Figure 11.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ECGRecord SYSTEM "ecgML1.0.dtd">
<ECGRecord studyID="ECG0345">
    <StudyDate>2005-12-03</StudyDate>
    <StudyTime>11:23:01</StudyTime>
    <PatientDemographics>
        <Sex> male</Sex>
    </PatientDemorgraphics>
    <Record>
        <ClinicalProtocol>
           <Medication>Digoxin, Quinidine</Medication>
        </ClinicalProtocol>
        <RecordData>
            <Channel>MLII</Channel>
            <Waveforms>
                <XValues>
                    <XOffset dataType="time">00:00:00.000</XOffset>
                    <Duration dataType="time">00:30:06.000</Duration>
                    <SampleRate unit="Hz">360</SampleRate>
                </XValues>
                <YValues unit="mV">
                    <FileLink URL="http://www.physionet.org/">MIT-BIH ECG Database FileLink>
                    <RealValue>
                        <From dataType="time">00:00:00.000</From>
                        <To dataType="time">00:00:10.000</To>
                        <Data>0.350,0.350,0.020,-0.210,-0.330,-0.370...,-0.110,-0.130.</Data>
                        <Comment>this is the list of real value of the first second of recod 100,
                     separated by comma</Comment>
                    </RealValue>
                </YValues>
          </Waveforms>
          <Annotations>
              <WaveNotation>
                  <Pwave>
                      <Onset dataType="samples">162</Onset>
                      <Peak dataType="samples">173</Peak>
                      <Offset dataType="samples">196</Offset>
                      <Annotation>Normal</Annotation>
                  </Pwave>
                  <QRSwave>
                      <Onset dataType="samples">201</Onset>
                      <Peak dataType="samples">224</Peak>
                      <Offset dataType="samples">258</Offset>
                      <Annotation>Ventricular premature beat</Annotation>
                  </QRSwave>
              </WaveNotation>
          </Annotations>
      </RecordData>
    </Record>
</ECGRecord>
```

Figure 11. An example of an ecgML-based ECG representation.

4.2. Encoding ECG Waveform Data in ecgML

ECG waveform data are a key component of an ECG record. Given that a wide range of ECG recoding devices are available today, the size of waveform data can vary dramatically. Moreover, they are usually expressed in binary format. In order to accommodate the full spectrum of ECG data, ecgML provides the following three options to handle ECG waveform summarized in Figure 12.

- For large data files, typical of Holter recording, ECG data are maintained in an external file, which can be referenced in ecgML using a *FileLink* element;
- Using *RealValue* elements to directly include actual waveform values in ecgML as the content of the element;

- To encode binary waveform data using a *BinaryData* element. ecgML is based on XML techniques to encode and represent ECG records. Due to the valid-character restriction posed by XML specification, simply embedding binary ECG waveform data within ecgML may cause the parser to encounter invalid sequences and fail to achieve its intended goal. Thus, each *BinaryData* element is associated with a specified *encoding* scheme, which may be *Base64* or *hexadecimal*.

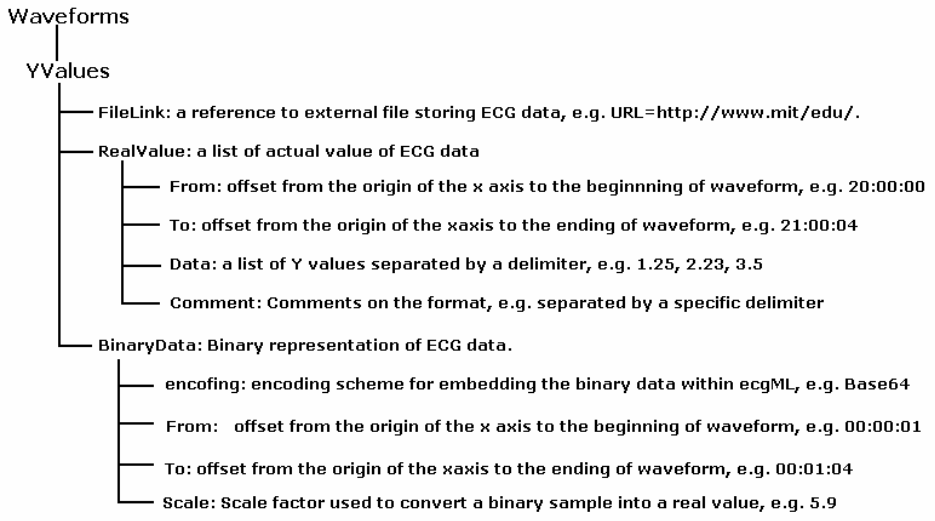


Figure 12. Framework for handling ECG waveform data in ecgML.

4.3. Accompanying Tools

A series of Java-based, user-friendly tools are being developed to assist users in exploiting ecgML-based application [16],[17]. These include ecgMLgenerator, ecgML parser and ecgMLBrowser.

The ecgMLgenerator produces ecgML-based ECG record from existing ECG databases. While the ecgMLparser allows the user to read the ECG record and access their contents and structure, ecgMLbrowser provides onscreen display of the collected waveform data as shown in Figure 13. The hierarchical structure of the ecgML-based ECG record is displayed on the left hand side. It can be expanded and shrunk at any level. The waveform data are shown on the right hand pane. The ecgMLeditor allows an authorized user to modify the contents of an ecgML-based ECG record. Some components such as the raw waveform data are not allowed to be changed.

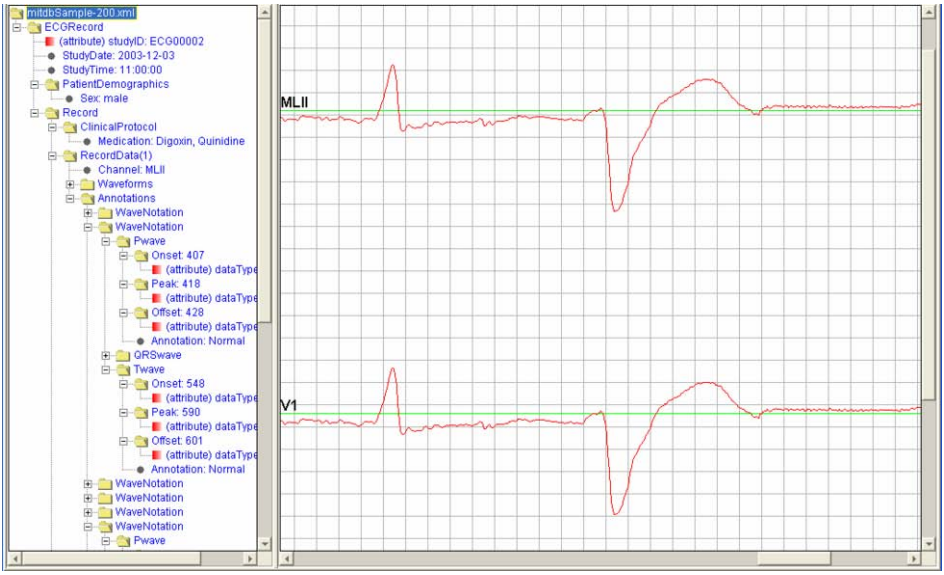


Figure 13. A screenshot of ecgMLbrowser.

5. Summary

The above coding systems are developed for a specific purpose - they are goal-oriented. For example, ICD takes a pathophysiological and aetiological view of medicine and classifies diseases primarily according to the organ system involved with some important exceptions such as infectious diseases [18] It is unable to fulfill the needs of all users.

Another problem is that the terms are changing over time, thus the meaning of a same term can be different over time, new terms are created and old terms may be replaced. When the meaning of a standard code is changed, the data will be interpreted incorrectly. When a standard code is removed from the coding system, the data are no longer interpretable. Additionally there is a lack of comprehensive, standardized medical terminology, and highly skilled experts are required to interpret the systems.

Despite these limitations, coding and classification systems have been successfully applied in health care. A widely accepted coding system is essential for storing and exchanging patient records efficiently.

Glossary

ACR	American College of Radiology
ASTM	American Society for Testing and Materials
CDA	Clinical Document Architecture
CEN/TC251	Comité Européen de Normalisation Technical Committee 251
CTV3	Clinical Term Version 3

DBMS	Database Management System
DICOM	Digital Imaging and Communications in Medicine
DNA	DeoxyriboNucleic Acid
ECG	ElectroCardioGram
EEG	ElectroEncephaloGram
EPR	Electronic Patient Record
FDA	U.S Food and Drug Administration
HL7	Health Level 7
ICD	International Classification of Disease
MR	Magnetic Resonance
NEMA	National Electrical Manufacturers Association
NHLBI	National Heart, Lung and Blood Institute
NHS	National Health Service
NIH	National Institutes of Health
OPCS-4	Classification of Surgical Operations and Procedures, Fourth Revision
PACS	Picture Archive and Communication System
PET	Positron Emission Tomography
PGA	Programs for Genomic Application
SCP-ECG	Standard Communications Protocol for Computer-Assisted Electrocardiography
SNOMED	Systematized Nomenclature of Medicine
SNOMED CT	Systematized Nomenclature of Medicine Clinical Term
SNOMED RT	Systematized Nomenclature of Medicine Reference of Terminology
SPECT	Single Photo Emission Tomography
UCUM	Unified Code for Units of Measure
UID	Unique IDentifier
WHO	World Health Organization
XML	eXtensible Markup Language

References

[1] E. Shortliffe, L. Perreault, G. Wiederhold, L. Fagan, eds., Medical Informatics: Computer Applications in Health Care and Biomedicine, New York: Springer-Verlag;Second Edition 2001.
[2] W. Tompkins, Biomedical Digital Signal Processing: C lauguage Examples and Laboratory Experiments for the IBM PC. Englewood Cliffs, N.J.; London : Prentice Hall, 1993.
[3] B. Robson, R. Mushlin, Genomic Messaging System and DNA Mark-Up Language for Information-Based Personalized Medicine with Clinical and Proteome Research Applications. J. Proteome Res. 3(5), (2004) 930 – 948.
[4] E. Coiera, Guide to Health Informatics. Second Edition. London: Hodder Arnold, 2003.
[5] Y. Pu, R. Patterson, Comparison of R-wave detection errors of four wireless heart rate belts in the presence of noise. Physiol. Meas. 24, 913-924, 2003.
[6] P. Razifar, M. Lubberink, H. Schneider, B. Langstrom, E. Bengtsson, and M. Bergstrom. Non-isotropic noise correlation in PET data reconstructed by FBP but not by OSEM demonstrated using auto-correlation function. BMC Medical Imaging, 5:3, (2005).
[7] J. H. van Bemmel, M.A. Musen, Handbook of medical informatics. Heidelberg, Germany ; AW Houten, Netherlands : Springer Verlag, 1997.
[8] E.F. Codd, A relational model of data for large shared data banks. Communications of the ACM, 13(6), (1970), 377–387.
[9] A.E. Flanders and J.A. Carrino, Understanding DICOM and IHE. Seminars In Roentgenology, 38(3), (2003) 270-281.

[10] Graham RNJ, Perriss RW, Scarsbrook AF. DICOM demystified: A review of digital file formats and their use in radiological practice. Clinical Radiology 60, (2005), 1133 – 40.
[11] ENV 1064 standard communications protocol for computer-assisted electrocardiography. European Committee for Standardisation(CEN), Brussels, Belgium, 1996.
[12] B. Jung, E.P. Andersen, J. Grimson, Using XML for Seamless Integration of Distributed Electronic Patient Records. In Proceedings of XML Scandinavia 2000 conference, Gothenburg, Sweden, May 2000.
[13] A. Tirado-Ramos, J. Hu, K.P. Lee. Information object definition-based unified modeling language representation of DICOM structured reporting: a case study of transcoding DICOM to XML. J Am Med Inform Assoc, 9(1), (2002), 63-71.
[14] http://www.connectingforhealth.nhs.uk/terminology/readcodes/, accessed Mar 2008.
[15] H.Y. Wang, F. Azuaje, B. Jung, N.D. Black, A markup language for electrocardiogram data acquisition and analysis (ecgML). BMC Medical Informatics and Decision Support, 3 (4), 2003.
[16] H.Y. Wang, B. Jung, F. Azuaje, N.D. Black. ecgML: Tools and technologies for multimedia ECG presentation. In the Proc. of XML Europe Conference, London, United Kingdom, May 2003.
[17] H.Y. Wang, F. Azuaje, G. Clifford, B. Jung, N.D. Black, Methods and tools for generating and managing ecgML-based information. In the Proc. of 2004 Computers in Cardiology, IEEE Press, Chicago, September 2004.
[18] http://www.opengalen.org/, accessed Mar 2008.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-158

III.3. Mining, Knowledge and Decision Support

Dewar D. FINLAY, Chris D. NUGENT, Haiying WANG, Mark P. DONNELLY and Paul J. McCULLAGH
School of Computing and Mathematics, University of Ulster, Shore Road, Belfast, Co. Antrim, N Ireland, BT37 OQB

1. Introduction to Decision Support

A decision support system (DSS) usually consists of a computer program which provides assistance in the decision making process. DSSs have found application in many domains and have seen particular prominence in medicine. This is an obvious application due to the large amounts of data (e.g. laboratory measurements such as blood pressure, heart rate, body-mass index) and information (e.g. patient history, population statistics based on age and sex) that must be considered before diagnosing any disease. As well as assisting in primary diagnosis, a DSS can reduce medical error, assist compliance with clinical guidelines, improve efficiency of care delivery and improve quality of care. Like their human counterparts, clinical DSSs can also be designed with medical specialities in mind.

In this review we look at the various traits of decision support in clinical practice. We examine two broad types of support, classification and prediction, and summarize various automated techniques that form the building blocks for these processes. We conclude with a discussion of some of the current trends in decision support along with a case study looking at how techniques commonly used in decision support can be employed in data mining and knowledge discovery.

1.1. The Modern Clinical DSS

Modern commercial clinical DSSs have evolved from the work of early investigators who set about developing systems that could flag the presence and genre of disease using the computational tools available at that time. An example of such a system was that developed by DeDombal in 1972 [1]. This system, designed to diagnose acute abdominal pain, used statistical techniques to suggest to which of 13 diagnostic categories a patient was most likely to belong. Although most early systems utilised statistical techniques, investigators soon began applying other methods. An example of this was in the implementation of MYCIN [2], a rule-based system designed to diagnose and recommend treatment for certain blood infections and other infectious diseases. In this system clinical knowledge was modelled as a set of IF-THEN rules with certainty factors attached to diagnoses. More recent systems have also been developed with the ability to provide support over a comprehensive range of conditions. The INTERNIST system [3] was capable of providing hundreds of

diagnoses in internal medicine, based on thousands of symptoms. This has since been superseded by Quick Medical Reference (QMR) which contains information on almost 700 diseases and more than 5,000 symptoms. Iliad, DXplain and DiagnosisPro are similar interactive consultation systems, which use clinical findings to produce a ranked list of diagnoses which might explain the clinical manifestations.

1.2. Categories of Decision Support System

In Table 1, adapted from [4], likely examples of DSS utility are presented. From this table it is evident that DSSs can be designed to provide a wide range of functionality and hence clinical support. We want, however, to give further consideration to two of the most common and hence important genres of functionality, namely "classification" and "prediction". These forms of support are analogous to the processes of diagnosis and prognosis respectively.

1.2.1. Classification

Classification is the term used to describe the process of analysing data to form decisions regarding the class membership of individual instances. An instance can be thought of as the individual case or record that is observed while a class relates to the possible categories to which the instance can belong. For example, consider a patient who has arrived at the emergency department with chest pain. This chest pain could result from some form of cardiac pathology or perhaps be attributed to less serious causes such as indigestion or a panic attack. The information collected through the medical assessment of the patient is used to describe the instance. Using this information, a diagnosis is made which places the instance into some predefined class, such as those outcomes outlined above. The actual diagnosis is formed by comparing and locating similarities from the details of the medical assessment with previously documented assessments for which there exists a known diagnosis.

1.2.2. Prediction

Similar to classification, prediction involves the analysis of measured information in order to form a decision. While classification is synonymous with diagnosis, prediction can be compared to prognosis. In this sense the analysis of information recorded to date is used to produce a forecast of the expected outcomes in the future or over time. When dealing with chronic diseases, for example, medical staff often refer to the records (instances) of patients who have, in the past, suffered from the same disease. These cases can then be used to predict, for example, the life expectancy or treatment outcomes for current patients.

2. Automated Decision Making

Regardless of functionality rendered, all decision support systems are based upon the same underlying processes. These processes are summarised in the block diagram presented in Figure 1, which consists of three stages. The first stage involves the measured data and its treatment and should result in useful 'features'. The second stage involves the interpretation of these features to provide some outcome. The last stage considers the relevance of the outcomes with some quantification of performance.

Table 1: Examples categories of clinical decision support systems, after Coeira [1]

Type of support	Clinical example
Alerts	Changes in a patient's vital signs can provide an alert, requiring immediate intervention, e.g. an ECG monitor on a ward can alert nursing staff to cardiac arrhythmia
Reminders	Digital storage of patient data can be used as an aide-memoir for doctors, e.g. during a surgical procedure the Electronic Patient Record (EPR) may generate a list of immunizations that each patient on the schedule requires based on medical guidelines
Diagnosis assistance	The DSS can use clinical findings (signs, symptoms, laboratory data) to produce a ranked list of likely diagnoses which might be associated with the clinical manifestations. Rare presentations of common illnesses, or illnesses which require specialised expertise can be flagged. Indeed further tests which would confirm or eliminate a diagnosis can be suggested.
Critiquing	The DSS can be used to critique the professional, e.g. if a patient's haemoglobin level is above the normal transfusion threshold, a clinician requesting a transfusion should receive a critique and must justify the procedure, by providing additional information such as 'patient has active bleeding' [2]
Therapy planning	Planning systems can propose a treatment based on a patient's condition and accepted treatment guidelines
Prescribing	Prescribing systems can assist with complex drug interactions, dosage errors, and contraindications such as allergies
Information retrieval	Intelligent information retrieval systems can assist with the generation of clinical questions and act as information filters to reduce the number of documents found in response to a query
Image recognition	Image processing algorithms can detect potentially abnormal images for more detailed human attention and interpretation, e.g. high density irregularities in mammograms can prompt further investigation

In the following paragraphs, we describe the various stages beginning with a description of features and their manipulation. We then introduce some of the fundamentals of automated interpretation algorithms. This is a short description, as we devote the entire following section to a more comprehensive description of widely used algorithms. We conclude this section by introducing common performance metrics.

2. 1. Features

Automated classification relies on a useful set of measurements or variables which accurately describe the prediction problem at hand. We refer to such measurements or variables as 'features'. In some, usually very simple, prediction problems we can easily identify the variables that are most useful and those which are not. Using the detection of diabetes as an example, we know that blood glucose level is a fairly sound indicator. It is also fairly obvious, even to those with the most basic clinical knowledge, that looking at the colour of a patient's hair is unlikely to inform us of the presence of this particular disease. In more complex problems, however, there may be many

measurements or variables and therefore many possible features. Additionally, there may also be little understanding of how these measurements relate to the likely disease. In such cases, where there are many poorly understood measurements, care must be taken that the actual interpretation algorithm is not overloaded with redundant information. The terms 'feature extraction' and 'feature selection' both refer to the process of reducing the number of variables which are to be considered in an interpretation problem. As illustrated in Figure 1, we refer to these processes collectively as 'feature processing'.

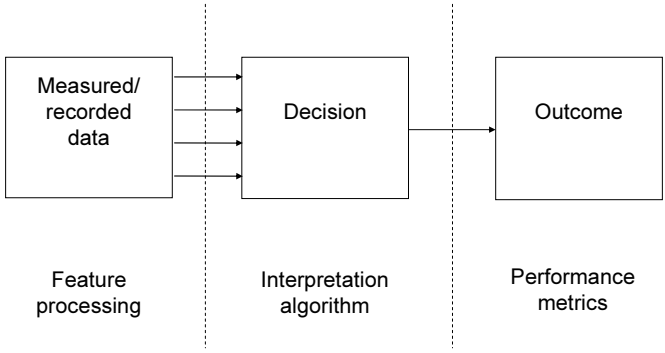


Figure 1. Steps involved in decision support

2.1.1. Feature Extraction

In feature extraction, a new set of features is obtained from the original features through some functional mapping [5], [6]. In this technique, the original features are not retained, but are transformed into a smaller feature space. Although feature extraction techniques are suited to many classification problems, there is a disadvantage in that when original features are transformed to a smaller feature space, all appreciation for the original features is lost. This is a particular disadvantage if the significance or contribution of the original features to the classification outcome is of interest. This will be exemplified in the case study at the end of this section.

2.1.2. Feature Selection

Feature selection works in a similar way to feature extraction by reducing the number of features considered in classifier development. In feature selection, however, the aim is not to transform the original features into a smaller feature space, but is to eliminate redundant features, whilst retaining those which provide the most discriminatory information. This process is said to result in the 'best subset' of original features [7]. This is particularly useful because as well as giving a useful insight into the nature of the prediction problem at hand [8], determining the most relevant features can provide an indication of which features should initially be measured.

2.2. Interpretation Algorithms

The interpretation algorithm can be considered the core component of a DSS as it must take the presented features and make a decision based upon their values. Thinking again of the simple example of analysing blood glucose levels for the presence of diabetes, a very simple algorithm could be implemented that would flag the presence of diabetes if a certain threshold has been reached. The threshold which dictates the presence of this disease can easily be specified by anyone with a prior knowledge of the domain. In more complex diseases, where there are perhaps many features and a less comprehensive knowledge of the disease process, development of the decision algorithm is more difficult. Specifically, in cases where the meaning of the features and the relationships between them are not obvious, it is necessary to develop an algorithm that will learn how to make decisions based on a pool of exemplar data. In this case, we say that the knowledge that is contained within the decision algorithm is 'learned' from the existing data. The process of 'learning' knowledge involves exposing an algorithm to a set of 'example cases' that represents the classification task at hand. Subsequently, when unseen data in the form of 'test cases' are presented, the algorithm will suggest which group or category each case belongs to, based on the previous learning. There are many ways in which interpretation algorithms can learn, these are broadly categories as either 'supervised' or 'unsupervised'.

2.2.1. Supervised Learning

In the supervised approach, the algorithm is exposed to a number of example cases. Each will consist of a set of features along with the actual outcome that was observed when those features were measured. After application of a suitable number of training examples, the algorithm is considered 'trained'. Subsequent to this, when the algorithm is exposed to a new example case which will consist of an unseen set of features, it should be able to indicate what the output would be. The number of training examples required to adequately train the algorithm relies on many factors. These include the type of algorithm used, the number of features and the overall complexity of the problem. The final performance of the algorithm is also dependent on many such factors. A diagram illustrating supervised learning is displayed in Figure 2.

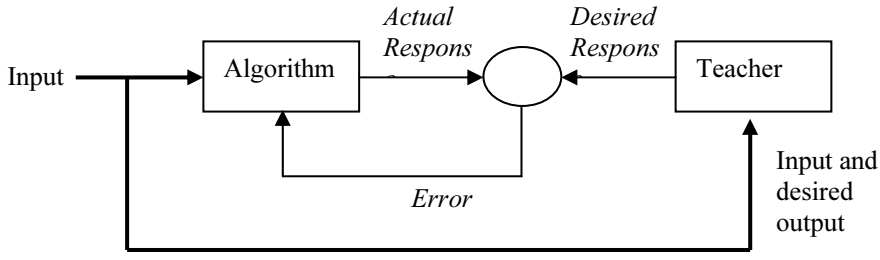


Figure 2. Diagrammatic representation of the supervised learning approach (Σ represents summation)

2.2.2. Unsupervised Learning

In unsupervised learning, the algorithm is also presented with a number of example cases, each consisting of the relevant features. As well as being more difficult to conceptualise and visualise, a big difference with this approach is that the algorithm is not given the corresponding outcome for each training example. Instead, based on the presented features, examples are clustered. When new unseen cases are subsequently presented, these are added to an already existing cluster or group. The algorithm itself determines possible segregating properties.Even though unsupervised learning does not require details of the desired output, some methods require details or guidelines on how the final groupings are to be organised, to avoid the situation where the end result may not be successful.

2.3. Performance Metrics

2.3.1. Accuracy

When trying to understand how well a given interpretation algorithm performs it is desirable to quantify how well it predicts the presence or absence of disease. This quantification is based on a comparison with the actual or true outcomes. The easiest way to achieve this is to calculate the percentage of cases that have been correctly classified. Refereed to as 'accuracy' (ACC), this can be defined as [9]:

$$ACC(\%) = 100 \times \frac{x}{N} \tag{1}$$

where x is the number of subjects that have been correctly classified and N is the total number of subjects in the dataset. Although this measure gives an indication of overall interpretation performance, no information can be attained from this metric regarding the breakdown of the performance of a particular group within the dataset, e.g. how many 'normals' (i.e. people without the disease) have been correctly assigned, or how many 'abnormals' (i.e. people with the disease) have been correctly assigned. To achieve this further measures of 'sensitivity' and 'specificity' are required.

2.3.2 Sensitivity and Specificity

Sensitivity (SEN) is a measure of how well a given interpretation algorithm can predict the presence of disease. Specificity (SPE), on the other hand, is how well a given interpretation algorithm can predict the absence of disease (SPE).

These are defined as [10]:

$$SEN(\%) = 100 \times \frac{a}{a+c} \tag{2}$$

$$SPE(\%) = 100 \times \frac{d}{b+d} \tag{3}$$

where a is the number of true positives, indicated as positive by the given algorithm, b is the number of false positives (i.e. true negatives that are indicated as positive by the given algorithm), c is the number of false negatives (i.e. true positives that are indicated as negative by the given algorithm), and d is the number of true negatives indicated as

negative by the given algorithm. In general a high level of SEN will be exhibited by an algorithm that performs well in the detection of presence of disease, a high level of SPE shall be exhibited by a classifier that performs well in the detection of absence of disease.

3. Interpretation Algorithms

In this section, we look at actual interpretation algorithms. We begin by discussing common algorithms that avail of supervised learning.

3.1. Examples of Supervised Methods

3.1.1. Statistical

The main objective of the statistical approach is the allocation of a test case to one of a number of possible diagnostic categories, with minimum probability of miscalculation [11]. When using such a classification technique, the background knowledge which is known is used to manipulate the data into a number of groups within the data set as opposed to using the knowledge to design the 'rules' for the algorithm itself. Once the data has been arranged, it is then possible to calculate a number of statistical measures relating to how likely a specific feature is in belonging to a specific category. These types of 'probability' measurements form the underlying basis of the statistical interpretation process.

3.1.2. Induction Algorithm

An induction algorithm builds a set of rules from a set of training examples which can then be used to assign unseen examples. After development, these rules are often represented as a 'decision tree' structure. In this type of system, a hierarchy of decision nodes connected by branches forms a decision tree. Each node represents a binary test, the result of each test determines the path through the tree network. After the rule set has been established, this type of algorithm is easy to implement, although the induction algorithms themselves are complex. The main disadvantage of the resulting rule set is their rigid nature i.e. only a very strict set of conditions can be tested for with sharp boundaries between normal and abnormal.

3.1.3. Neural Networks

Neural Networks (NNs) are vast interconnections of elements called neurons. The biological neuron itself is a building block of the brain and in NNs an attempt is made to replicate the behavior of these neurons. An artificial neuron consists of a number of inputs which are weighted before reaching a main body or processing unit, the signal is processed and passed to an output. The operation of the NN depends greatly on how the neurons are interconnected. After design, a NN must be trained. During this process the NN adjusts its own internal parameters causing the network to converge to the desired response. Although in this section we are focusing on supervised techniques, NNs can also be configured to work in an unsupervised manner.

3.1.4. Combining Classifiers/Mixture of Experts

In this approach, several interpretation algorithms are employed simultaneously with the aim of providing superior performance in solving complex problems. The first of two such approaches is known as 'Ensemble Combination' [12]. Here, each algorithm in the ensemble grouping attempts to solve the same problems and derives its own solution. A voting system is then employed to determine which outcome was the most popular among the majority of techniques. This outcome is then taken forward as the final accuracy.

In another approach, known as modular learning [12], each algorithm is presented with a different subsection of a particular problem. In this approach, each algorithm usually only focuses on a relatively minor part of the signal or image being classified and this often provides superior accuracy because the overall problem is divided into several simpler problems.

3.2. Examples of Unsupervised Methods

Traditional unsupervised methods include hierarchical clustering, k-means and the Kohonen Self-Organizing Feature Maps (SOM) [13]. These common techniques are described below.

3.2.1. Hierarchical Clustering

The hierarchical clustering approach constructs clusters in either a bottom-up (*agglomerative*) or a top-down (*divisive*) manner. The agglomerative method starts with every data sample in a single cluster. Then, it repeatedly merges the closest pair of clusters into a single cluster until all data samples are in one cluster. Depending on the way of defining similarity between clusters, an agglomerative clustering approach can be implemented in several ways, such as single linkage, average linkage and complete linkage. As a top-down technique, a divisive method starts with all the data samples in one single cluster, and then successively splits them into smaller clusters until each cluster contains only one sample. This method is outlined in Table 2. Hierarchical clustering is conceptually simple and relatively easy to implement. By visualising a dendrogram, which diagrammatically summarize the whole hierarchal clustering process, basic relationships between all the data samples can be obtained. However, the deterministic nature of this technique makes it impossible to make any adjustments and corrections once a data sample is assigned to a node.

3.2.2. K-Means Clustering

A *k*-means method is an algorithm that divides a data set into *k* disjoint partitions such that a certain metric relative to the centroids of *k* clusters is minimized. The algorithm is composed of several steps. In the first step, *k* initial cluster centroids are randomly determined. Then each sample is assigned to the closest centroid based on a distance function. Once this is done, the new centroid is recalculated for each cluster by taking the mean of all the cluster members. The process is iterative until no change in the centroids is observed. For a large data set, *k*-means technique is computationally faster than hierarchical clustering. Nevertheless, the performance of *k*-means clustering greatly depends on selection of initial seeds. It could get stuck at a local minimum with poor quality. Moreover, the number of clusters needs to de defined at the onset. In

some cases, it could be difficult to predict what the number of k should be. Fixed number of clusters can make it difficult to predict what *k* should be.

Table 2: Basic Divisive Hierarchical Clustering Algorithm

1: Initialise: Start with one all-inclusive cluster
2: Repeat
3: Calculate pair-wise distances within each cluster
4: Find two data samples that are in the same cluster but have a largest distance
5: Use these two samples as seed points to create two new clusters
6: Assign all samples in the original cluster to two new clusters that have the closest seed
7: Until each cluster contains only one sample

3.2.3. Self-Organising Map

Based on the combination of a competitive learning principle with a topological structuring of neurons, the SOM is perhaps one of the most widely used unsupervised neural networks in the literature. It implements an ordered dimensionality-reducing mapping of high-dimensional data sets. A SOM generally involves an architecture consisting of a 2-dimensional Kohonen layer, in which each neuron receives the same input from an input layer, as illustrated in Figure 3. Upon completing a learning process, the SOM is able to provide a graphical representation of clustering distribution with neighbouring neurons tending to define related clusters. A detailed description of this algorithm, along with the selection of its learning parameters can be found in [13]. Similar to the *k*-means algorithm, the SOM inherently requires users to specify the network topology and the number of neurons in advance. Such a fixed, predetermined grid structure has negatively influenced its applications in the context of knowledge discovery.

3.2.4. Utility of Unsupervised Techniques

The potential for the use of unsupervised algorithms is often much less obvious than that of their supervised counterparts and in the following paragraphs we provide some examples of their utilisation as reported in the literature. The ability to reveal the patterns hidden in a given data set certainly lends unsupervised learning techniques to many applications. From a knowledge discovery perspective, various clustering methods have received great attention. Bortolan and Pedrycz [14] presented a SOM-based interactive framework for ECG signal analysis, which allows users to gain an insight into the structure of the data under investigation. Douglas *et al*. [15] used the *k*-means algorithm to perform automatic segmentation of magnetic resonance images (MRI) of the residual limb of trans-femoral amputees into limb and background. Other studies include the utilization of the SOM for spatiotemporal analysis and classification of body surface potential map (BSPM) data investigated by Simelius *et al.* [16]. This study indicates that, unlike traditional QRS integral maps which miss time-dependent

features of activation process and have to compress the spatiotemporal information of the whole QRS complex into one spatial distribution, the SOM-based approach can incorporate the temporal dynamics and thus take advantage of both the whole spatial distribution and the temporal development during the QRS.

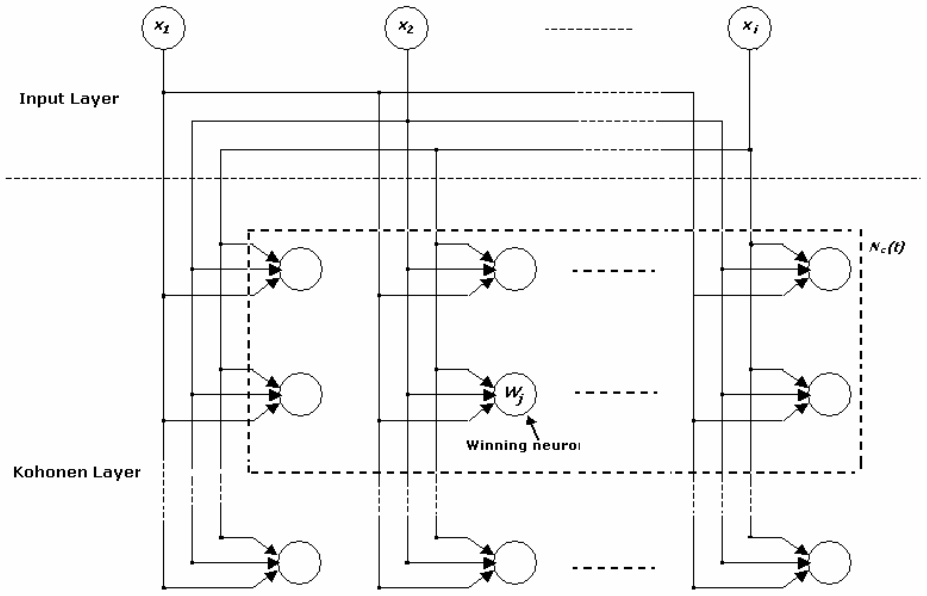


Figure 3. A typical SOM architecture consisting of two layers: an input layer with *i* input samples and a 2D layer which contains a grid of neurons. The rectangle in dotted lines represents a topological neighborhood, $N_c(t)$, of winning neuron, W_j, which is monotonically decreased as training progresses.

In bioinformatics, pioneering studies include investigations by Eisen *et al.* [17] and Golub *et al.* [18]. Eisen *et al.* [17] used an agglomerative hierarchical clustering method to group genes according to their gene expression similarity. By visualizing the output dendrogram, the inherent data structure encoded in microarray data can be conveyed in a form that is intuitive for biologists. Golub *et al.* [18] successfully applied SOMs to distinguish types and subtypes of leukemia based on gene expression patterns.

4. Future Trends and Possible Applications of DSS Techniques

In the introductory paragraphs of this article we have discussed DSSs mainly in the context of clinical care. There is also scope for utilizing DSSs in everyday life to support independent living [19]. We now live in an age where we can take more control of our own health by using readily available devices which support personal vital sign monitoring. The availability of such devices provides the opportunity for the deployment of decision support systems outside the boundaries of primary or secondary healthcare establishments. If we consider very basic home based measuring

devices such as the blood pressure monitor or blood glucose monitor, it is possible to add a simple algorithm which would sound an alarm if an excessive measurement were recorded. Obviously this intermediate form of assessment is merely a pre-cursor to assessment by a healthcare professional; however, it is a step forward in terms of the automated analysis of vital signs outside of the conventional clinical setting. In the following paragraphs we discuss three, albeit closely related, areas which we believe define new trends and challenges in decision support.

4.1. Wearable Computing

Recent technological advancements in the domain of 'wearable computing' have the potential to streamline the process of vital signs monitoring and subsequently decision support. In particular, developments in 'smart clothing' allow sensors and electrodes to be seamlessly woven into garments. This provides the ability to measure a wide range of vital signs from a subject when they wear a specially designed shirt. Obviously this streamlines the process of vital signs monitoring as the subject need not worry about the attachment of a number of discrete sensors and recording apparatus.This potentially provides much more freedom and scope for vital signs monitoring as much more information can be recorded in a less obtrusive manner whilst the patient goes about their everyday tasks. Although adding additional complexity to the design of the decision support system, there is now the potential to provide a fuller assessment of a patient's status.

4.2. Smart Homes

Whether it is through 'smart clothing' or conventional recording apparatus, vital signs monitoring has traditionally been an integral part of home based monitoring. However, it is now also possible to record more general activities of the person within their home and their interaction with domotic devices. It has emerged that much more can be gained by complementing and correlating traditional vital signs with this behavioral information. For example, a scenario of an increased heart rate should not be considered as alarming if it is also known that the person has just run up a flight of stairs. On the other hand, there may be reason for concern if a person's blood sugar levels are classified as being abnormal and the person has not been in the kitchen to cook food all day. Although this approach has the potential to define a new paradigm in healthcare delivery, a big challenge will be processing the endless information and the complex scenarios that are possible. It is believed that manipulation and further development of existing decision support systems will make this realizable.

4.3. Internet Based Healthcare

With all of the aforementioned scenarios, the decision support systems will be required to analyse information from a number of different sources. Given complementary advances within the telecommunications domain it is possible for all of the information to be analysed on the device or system within the home environment or relayed to some central healthcare facility. The advantage of the latter is the remote support which may be offered to a person within the comfort of their own home. For example, a decision support system based in a hospital may receive information regarding a possible alarm for a person's high blood pressure. In addition, the system may receive information

suggesting the person has not taken their anti-hypertensive medication. Based on this information the decision support system could raise an alarm resulting in a healthcare professional calling the person at home to check on their general status. Although this is a simplistic scenario, it conveys the end-to-end element of healthcare monitoring and delivery which is directing the future development and deployment of medical decision support systems.

5. Case study: Mining for Electrocardiographic Information

DSSs have the ability to assist the decision making process in many applications. These range from classical clinical decision making to more contemporary applications associated with independent living. The developments in DSSs have also spurred the development of elaborate computational techniques and algorithms which in turn have created new opportunities evident in the emergence of new domains such as knowledge discovery and data mining. These domains and the tools used have also opened up new possibilities for gaining an understanding of and streamlining the process of healthcare delivery. In this case study we look at how some of the techniques that form part of the DSS process can be utilized for solving less conventional problems. We approach this discussion not by focusing on the technical rigors of this application, but on the application and need itself.

A medical discipline that has gained a lot from these developments is cardiology, particularly electrocardiology. This particular area saw one of the first applications of computers in medicine [20], and computerized electro cardiology is now a domain in its own right [21]. The application of computers in this area is widespread with computerized enhancement of the acquisition process right through to computerized interpretation and diagnosis being commonplace. In more recent years computers and particular data mining techniques have been applied in an attempt to gain more of an understanding of the principles of electrocardiology and of the underlying electrophysiology. To provide a typical example of how these techniques can be applied we look at the application of data mining techniques in the selection of electrocardiographic recording sites. The most effective way to capture electrocardiographic information is to record information from recording sites which are positioned all over the torso. Such an approach is referred to as body surface potential mapping and can involve recording information from in excess of 200 recording sites. This technique is deemed to provide the most diagnostic information, as effectively all information as projected on to the body surface is recorded. Although comprehensive, this technique is seldom used in clinical practice due to the excessive number of recording sites required and the associated complexity of the hardware. A more favored technique is the 12-lead ECG which as the name suggests renders just 12 channels of information for interpretation. Although widely used, the 12-lead ECG is also met with some skepticism as it is well appreciated that the locations of the recording sites are known to be sub optimal. There therefore is a very valid research question of where should electrocardiographic information be recorded from a patient's torso to increase diagnostic yield. Assuming that something around the number of recording sites that are required to record the 12-lead ECG is suitable, many investigators have set about analyzing body surface potential maps to find which of the 200+ recording sites available yield the most information. Historically this analysis has

been conducted using statistical methods [22]. However, recently more contemporary techniques have been endorsed.

The problem of recording site selection can be described as trying to locate the leads in a body surface potential map that yield the most diagnostic information. If we take this a step further and think of this problem in the context of automated classification, we can think of the information from each electrocardiographic lead as a feature and we want to reduce the number of these features prior to classification. We have therefore defined a 'feature selection' problem. As described earlier, feature selection deals with the removal of redundant variables, and the retention of those which are deemed more useful. This is in contrast to feature extraction, which transforms the features and hence will not tell us which electrocardiographic leads we should measure.

Regardless of the problem domain, feature selection would involve exhaustively evaluating all possible combinations of input features and choosing the best subset. In reality, the computational cost will be prohibitive if there are many features to be considered. This is likely to be the case in the ECG lead selection problem where, for this example, we have in excess of 200 features and hence leads. For this reason, there has been an interest in developing algorithms and strategies that locate optimal features at low computational cost [23].

6. Summary

In this review, we have examined the role of decision support in medicine. Development of DSS is often undertaken by the biomedical engineer, medical physicist or medical information, in collaboration with clinicians. The transfer of domain knowledge is a key element to this process. Selection of important features for the decision making process is crucial to the success of this. Performance can be measured by statistical metrics: accuracy, sensitivity and specificity. Interpretation algorithms may use supervised or unsupervised learning. A case study to support decision making in the domain of electro cardiology has been presented.

Glossary

ACC Accuracy
BSPM Body surface potential map
DSS Decision support system
EPR Electronic patient records
MRI Magnetic resonance imaging
NN Neural Network
QMR Quick medical reference
SEN Sensitivity
SOM Self organizing map
SPE Specificity

References

[1] F.T. de Dombal, D.J. Leaper, J.R. Staniland, A.P. McCann and J.C. Horrocks, Computer aided diagnosis of acute abdominal pain, *British Medical Journal* 2 (1972), 9-13.

[2] E.H. Shortliffe, Computer Based Medical Consultations: MYCIN, Elsevier, North Holland, New York, 1976.

[3] J.D. Myers, The Background of INTERNIST-I and QMR, in A History of Medical Informatics, eds. Bruce I. Blum and Karen Duncan (New York: ACM Press, 1990), 427-433.

[4] E. Coiera, Guide to Medical Informatics, The Internet and Telemedicine, Chapman & Hall Medical, 1992.

[5] N. Wyse, R. Dubes, and A. K. Jain, A critical evaluation of intrinsic dimensionality algorithms, in *Pattern Recognition in Practice,* E. S. Gelsema, and A. K. Jain, Eds. Morgan Kaufmann Publishers, Inc, 1980, 415-425.

[6] L. Huan, and H. Motoda, Feature transformation and subset selection. *IEEE Intelligent Systems,* 13 (1998), 26-28.

[7] A. Jain, and D. Zongker, Feature selection: Evaluation, application, and small sample performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1998), 153-158.

[8] J. Reunanen, Overfitting in making comparison between variable selection methods, *Journal of Machine Learning Research,* 3 (2003), 1371-1382.

[9] C. D. Nugent, PhD. Dissertation, University of Ulster, 1998.

[10] P. M. Rautaharju, H. W. Blackburn, and J. W. Warren, Review: The concepts of sensitivity, specificity and accuracy in evaluation of electrocardiographic, vectorcardiographic and polarcardiographic criteria, *Journal of Electrocardiology,* 9 (1976), 275-281.

[11] J. A. Kors, J. H. van Bemmel, Classification Methods for Computerised Interpretation of the Electrocardiogram, *Methods of Information in Medicine*, 29 (1990), 330-336.

[12] J.C. Sharkley, Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. London: Springer, 1999.

[13] T. Kohonen, *Self-Organising Maps*. Heidelberg, Germany: Springer-Verlag, 1995.

[14] G. Bortolan and W. Pedrycz, An interactive framework for an analysis of ECG signals, *Artificial Intelligence in Medicine*, 24 (2002), 109-132.

[15] T. S. Douglas, S. E. Solomonidis, V. S. P. Lee, W. D. Spence, W. A. Sandham and D. M. Hadley, Automatic segmentation of magnetic resonance images of the trans-femoral residual limb, *Medical Engineering & Physics* 20 (1999), 756-763.

[16] K. Simelius, M. Stenroos, L. Reinhardt, J. Nenonen, I. Tierala, M. Mäkijärvi, L. Toivonen, and T. Katila, Spatiotemporal characterization of paced cardiac activation with body surface potential mapping and self-organizing maps, *Physiological Measurement* 24 (2003), 805-816.

[17] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA*, 95 (1998) 14863-14868.

[18] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J. P. Mesirov, H. Coller, M. L Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 286 (1999), 531-537.

[19] C. D. Nugnet, J Augusto, *Smart Homes and Beyond.* Amsterdam, Netherlands: IOS Press, 2006.

[20] R.J.A. Schijvenaars, Intraindividual variability of electrocardiogram. Assessment and exploiuting in computerized ECG analysis, PhD Dissertation, University Rotterdam, 2000.

[21] www.isce.org, accessed Mar 2008.

[22] F. Kornreich, P. M. Rautaharju, J. Warren, T. J. Montague, and B. M. Horacek, Identification of best electrocardiographic leads for diagnosing myocardial infarction by statistical analysis of body surface potential maps, *American Journal of Cardiology*, 56 (1985), 852-856.

[23] D. D. Finlay, C. D. Nugent, P. J. McCullagh, and N. D. Black, Mining for diagnostic information in body surface potential maps: A comparison of feature selection techniques, *Biomedical Engineering Online,* 4 (2005) 51.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-172

III.4. Remote Healthcare Monitoring and Assessment

Chris D. NUGENT[a], Dewar FINLAY[a], Richard DAVIES[a], Mark DONNELLY[a], Josef HALLBERG[b], Norman D. BLACK[a] and David CRAIG[c]

[a] *School of Computing and Mathematics, Faculty of Engineering, University of Ulster, Shore Road, Northern Ireland, BT37 0QB*
[b] *Luleå University of Technology, Sweden*
[c] *Belfast City Hospital/Queens University of Belfast, Northern Ireland*

1. Introduction to Remote Healthcare Monitoring

Remote healthcare monitoring is the process of assessing the well-being of a patient when the patient themselves and their healthcare professional are not physically together in the same room. Conventionally, patient assessment in Primary and Secondary healthcare provision involves a face-to-face consultation between the patient and the healthcare professional (General Practitioner, Consultant etc.). Advances in technology, specifically medical devices, sensors and high speed fixed and wireless communication networks have now made it possible to bring the assessment process to the patient, as opposed to limiting the assessment to the constraints of hospitals and doctors' surgeries. As a result, it is now possible for a patient to be assessed whilst in the comfort of their home or to have their vital signs monitored whilst, for example, shopping or at work. In addition, it is possible for patients to visit their local Doctor's surgery and benefit from expert consultants and receive their advice, without having to have a face-to-face meeting with them.

It is the aim of this review to introduce the concepts of remote healthcare monitoring, discuss the drivers which have lead to the realisation and uptake of this approach and to look at the benefits that evolving technologies can offer in this domain.

1.1. Changes in Population Demographics

We are at present witnessing a change in the way our population is represented, in terms of age profile. In the 1950s the largest percentage of the population was made up by those in the category of 0-9 years of age. At the turn of the century this category was amongst one of the largest, with a similar level of numbers being found in other age categories. Predictions for the next thirty years have shown that this trend will continue and infact result in a profile of age categories with almost similar sizes across all age groups (Figure 1). People are living longer and the total population size is increasing. In addition, the percentage of our population represented by those aged 65 and over is steadily increasing [1]. From a healthcare perspective this offers a number of challenges. In the first instance, we are faced by the simple fact that the population is

now larger and hence places a larger demand on healthcare services. Secondly, there is an increased prevalence of healthcare problems and instances of disease within the elderly and thirdly there are now fewer younger people to care for the elderly [2].

Taking all of this into account has resulted in healthcare providers, governments and members of society, searching for new paradigms of healthcare delivery.

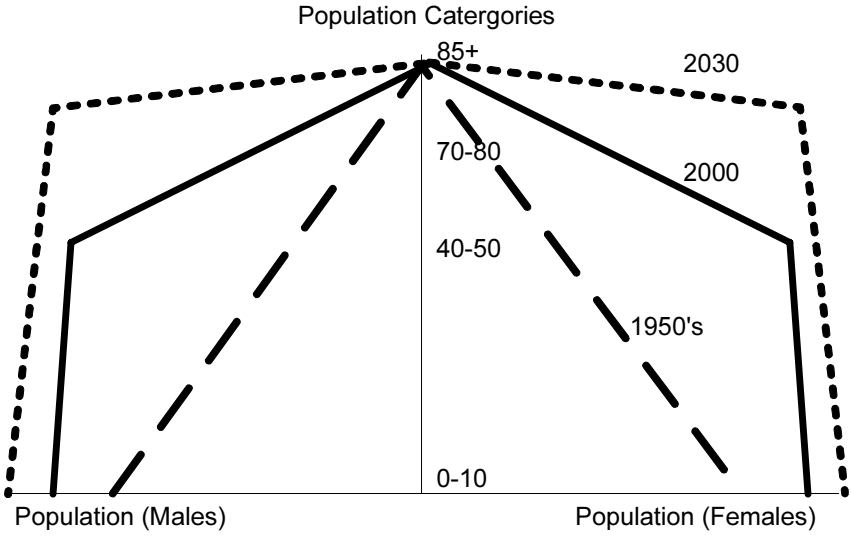


Figure 1 Estimated profile of age categories in 1950, 2000 and predictions for 2030.

1.2. The Impact of Technology from Social and Economic Perspectives

We have previously introduced the concept that the increase in the size of the population has a direct impact on the prevalence of instances of healthcare problems and of long term chronic diseases. In addition to placing a resource burden on primary and secondary healthcare organisations, this increase in population size has additional societal and economic implications. There are benefits from both perspectives which would support the introduction of remote healthcare monitoring.

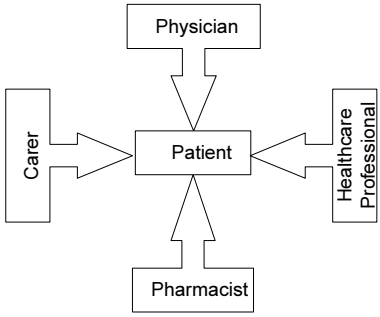
From a societal perspective, people are wishing to play a more active role in their own healthcare. This can be facilitated, at least notionally, via the realisation of remote healthcare monitoring where the patient is given a higher degree of responsibility to become more proactive with the management of their own healthcare and its assessment. Involvement in such a manner can also lead to the patient having a greater understanding of not only their condition, but also its means of assessment. In addition, there is the general appreciation that people recover more quickly in their own homes and would prefer to remain in their own homes, even if they are at greater risk [3]. Remaining at home for an extended period of time can also be linked to improvements in the patient's perceived quality of life. Taking this into account, the use of remote healthcare monitoring has the potential to avoid instances of institutionalisation when the patient requires high levels of healthcare assessment. It can also avoid lengthy and

inconvenient trips being made for the purposes of routine clinical assessments. Patients can also benefit from remote expert clinical diagnosis. For example, it is possible to send a patient's clinical details from one institution to another and receive expert clinical assessment via remote means.

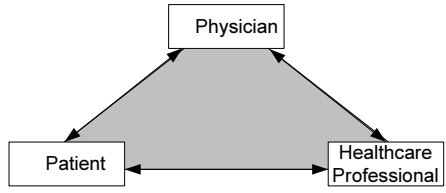
In addition to the societal benefits, deployment of technology to support healthcare assessment via remote means has a number of economic benefits. If a person can have their healthcare assessed within their own homes, removing the requirement of temporary or permanent institutionalisation, then significant healthcare savings can be gained. Additionally, those patients receiving clinical care within a hospital setting can be discharged in a much shorter space of time and continue to have routine assessments conducted via remote means. Finally, consultants can offer remote consultations to patients all over the world without having to leave their own office. This offers benefits to the patient in terms of them receiving treatment from 'world class' specialists who, because of geographical distances, would otherwise not be able to provide such treatment.

1.3. Models of Patient-Healthcare Professional Interaction

Traditionally, the means by which those involved in healthcare provision interact with their patients is based on a bi-lateral communication model (Figure 2 (a)). In this respect it is normal for the patient to have direct contact with a number of healthcare providers, however, commonly there is a potential lack of a unified infrastructure to support communication between healthcare providers. Deployment of technology and the establishment of care models facilitates communication channels to be established between both patients and healthcare providers as well as intercommunication between the healthcare providers (Figure 2 (b)). This offers an improved means of healthcare delivery from the patient's perspective as all those involved in the delivery of their care can interact.



(a) Interaction between stakeholders: the classic situation



(b) Enhanced stakeholder interaction facilitated via technological communications

Figure 2 Examples of Patient-Healthcare Professional interactions.

Although the aforementioned care models can be used to facilitate the deployment of technology, their conceptualisation can at times be faced with resistance. The deployment of technology within the realms of healthcare provision can offer benefits to the patients, however, the changes of practice which may be required from the healthcare providers' perspective, in addition to the problem of new service integration, creates major challenges concerning the widespread deployment of the technology. On the one hand, drivers such as the benefits of improved treatment results and alleviation of the problems associated with the shortage of physicians and nurses encourage such developments. Nevertheless, such innovations are faced with the constraints such as technology costs, reimbursement issues and legal issues, to name but a few. This inevitably results in a 'push-pull' scenario (Figure 3) from a technological perspective. The development of new technology results in new paradigms being pushed into the healthcare market. These can be considered as high risk solutions. If, on the other hand, a user driven approach is adopted, a pull effect can be witnessed. This may be viewed by some as the desired effect, where the end solution will be based on mature technology, which, from a technological perspective may not offer high levels of innovation, however, from the healthcare perspective can be seen as a low risk solution meeting the needs of the end users. In order to provide the patient with the best solution to their needs, a user centered design process offers many benefits. Adopting this approach allows the patient to be involved during the early stages of the design process.

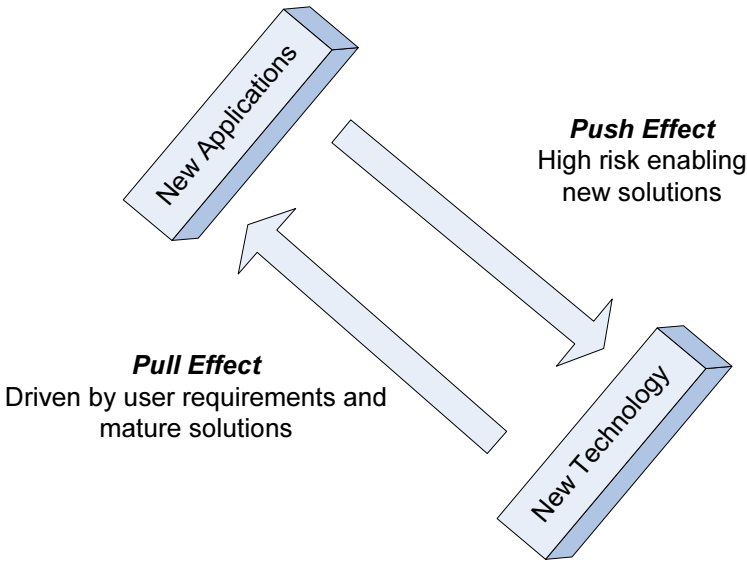


Figure 3 'Push-Pull' effect of technology within healthcare market place.

1.4. The impact of the Internet on Healthcare Delivery

The Internet is one of the most widely used and accepted forms of modern day communication, as well as a rich source of information. Patients can now seek healthcare advice, research their symptoms and determine possible diagnosis, in addition to researching available courses of treatments. Additionally, the Internet can be used as a forum to support communities offering discussion groups dealing with health and healthcare topics. Furthermore, the Internet offers great potential for distributed healthcare, which is important especially for sparsely populated rural areas, and for people with limited mobility.

The Internet makes it possible for user-centric healthcare, where the user's needs are in focus. More demands are also placed on being able to locate the necessary information and help in user understanding. This has created more demands on information and communication resources, as well as on available healthcare professionals. Nevertheless, this also means that for many health-related issues, the user can receive help from centralized healthcare resources. For example, a parent concerned about a child's skin rash could be taught how to acquire high-resolution pictures of the affected area and subsequently forward such pictures to a centralized healthcare service. This approach would offer a fast and efficient means of receiving direct feedback regarding the condition of the skin, without the necessity of the parent and child having to travel to the hospital for a physical examination.

2. Telemedicine

This Section aims to introduce the notion of Telemedicine and how technology can be deployed to link healthcare professionals together in an effort to improve the quality of the healthcare delivery process.

2.1. What is Telemedicine?

Telemedicine has received many definitions over the years. Most notably it can be described as the remote use of medical expertise at the point of need [4]. This results in the use of telecommunications to facilitate the sharing of medical knowledge and the delivery of healthcare over a distance [5]. With this approach medical expertise can be delivered to remote sites, rural areas or understaffed regions. lthough Telecare may also be considered as a form of Telemedicine, for the purposes of this article we will treat these two topics individually and focus within this Section on the element of Telemedicine and the associated technology which fosters a collaborative working environment between medical experts.

Telephony: A rudimentary means of providing telemedicine is via telephone communications. This can be in the form of two specialists discussing a patient's diagnosis or prognosis over large distances or similarly can relate to a consultation between a patient and a specialist. Other examples include automated services which patients access, via the telephone, to provide details regarding their current state of health. An example of this type of system is an automated diabetes service which permits patients to verbally enter information regarding their condition via an automated telephone system [6]. Such a system facilitates the routine assessment of a

patient's condition without the need for direct consultation thus alleviating pressure on healthcare services.

Video Conferencing Systems: Video Conferencing Systems have become of one the most popular platforms used to facilitate telemedicine and subsequently improve communications between healthcare professionals [7]. Such systems typically include the ability to send and receive video, audio, and text between one or more users. Additionally, some systems provide tools, such as shared whiteboards, which facilitate the communication of ideas and thoughts via informal means. The main purpose of using video conferencing systems in telemedicine is to provide a sense of presence through the provision of visual instructions, or demonstrations to patients, and to visually present or describe problems and symptoms to healthcare professionals.

Video conferencing provides a wide range of services that healthcare providers can avail of for example nurses can use video conferencing to make house-calls while keeping contact with the doctor, whilst patients can use the system to set up meetings with healthcare professionals from their home, instead of having to travel to the hospital.

2.1.1. Point-to-Point Ssystems

Similar to video conferencing, point-to-point systems offer commutations between stakeholders, however, they only provide communications between two points. At present, several such systems are freely available. While these systems are often sufficient in providing simple communication between a patient and the doctor, they do not offer the support group functionality that video conferencing systems do.

2.1.2. Remote Surgery

At the other end of the spectrum from telephony systems exists the concept of remote surgery. This technology provides the means to support the delivery of remote care through robotics [8]. An obvious advantage of this approach is that it permits world renowned surgeons from across the globe to provide patient care. The first remote surgery was conducted in 2001 and was referred to as "Operation Lindbergh". Through the use of three robotic arms setup in an operating theatre in France, surgeons located in New York where able to remotely perform a gallbladder operation over a fibre optic connection. Since this date, several other remote surgical procedures have taken place.

Store and Forward Services: In addition to real-time telemedicine systems such as those already described, offline telemedicine is also possible. Such care delivery does not require both parties to be present at the same time of communication. For example, technology within one consultant's office can be used to record some patient information which can be subsequently stored and forwarded at a set time to be reviewed offline by another consultant.

3. Telecare

This Section aims to introduce the notion of Telecare and how technology can be deployed to link remote patient information with healthcare providers in an effort to improve the quality and the healthcare delivery process.

3.1. What is Telecare?

Telecare combines the usage of sensing devices and telecommunications to support the remote monitoring of a patient within their home environment with the goal of providing a means of care support. Therefore, via remote patient monitoring it is possible for a Telecare service to react to abnormal situations which are a cause for concern. In response, a Telecare service can issue an alert to a caregiver or a family member. It is also possible for the person themselves to raise an alarm in instances when they may require support, for example, following a fall.

3.1.1. Technology Platforms Available for Home Based Monitoring

To provide a home based monitoring system to patients who require additional health care services requires exploiting current technology platforms that are already in place. These should be widely available and practical in terms of cost, performance and speed.

The Public Switched Telephone Network (PSTN) is a commonly used means of supporting the transmission of information from the patient's home to a healthcare organisation. The PSTN is available throughout the world and although originally based on analogue lines, is now almost entirely digital at its core. The primary use of the PSTN is the ability to connect people together to provide a phone based voice service. However, in more recent times, the PSTN has been used as a digital communication channel allowing information to be transferred to and from patients' homes. The PSTN itself has a number of drawbacks in comparison with more recent technologies based on the same infrastructure. It can be more difficult to set up and is expensive to run if the application is resource intensive and operates at relatively slow speeds making it suitable for only a small percentage of applications.

Broadband technology is another possibility. As its name suggests, broadband technology offers a wider band of information. One example of broadband is a Digital Subscriber Line (DSL) which encodes digital information at a higher frequency on one channel and sends voice on a lower frequency forming another voice based channel. DSL technology offers high availability, as it is based upon the already existing PSTN system which has worldwide coverage. There are a number of advantages with broadband technology. Firstly, it is easy to set up and easy to use and can even support wireless connectivity. Although dependent on line conditions such as the distance from the exchange, the speed and reliability are greatly improved. Broadband technology, and DSL in particular, are essential to providing the underlying communication infrastructure to allow home based services to become successful.

3.1.2. Telecare Services

One of the most common forms of Telecare services exist in the form of alarm based pendants. Alarm based pendants are essentially a device worn by a patient which can be used in either a passive or active manner. Their primary use is within the home environment and can be used to trigger various alarms. A passive pendant is one which involves no interaction between the patient and the device; this could take the form of a fall detector which passively monitors the orientation of the patient during home based activities. Falls are common place among elderly patients living alone and are the leading cause for such people having to enter into permanent institutional care. Such a device could alert a relative or health care personnel to a serious situation. A more obvious and active approach of an alarm pendant is one that is activated by the patient

whenever a security situation arises. Once the pendant has been activated, an emergency message is relayed to a healthcare professional or family member so that the necessary intervention can take place.

In addition to pendant alarms, Telecare services can be based upon a number of devices which may exist within the home environment, for example medication management devices, cognitive prosthetics, fire/smoke/water alarms and various door and device interaction sensors. The exchange of physiological data recorded from the patient in their home and transferred to a remote site can be referred to as Telehealth. Although the patient and the communication infrastructure for both Telecare and Telehealth are the same, it has been usual in the past to keep these two terms separate. Typical parameters which can be monitored within the auspices of Telehealth include blood pressure, electrocardiogram, weight and blood glucose levels.

4. Ambient and Pervasive Computing

The way in which healthcare can be delivered is changing rapidly. The convergence of mobile communications, decision support systems and Internet computing are offering a wealth of new healthcare paradigms. In this Section, we focus on a number of new and emerging technologies and demonstrate how they can be deployed within the realms of home based healthcare.

4.1. Intelligent Homes

Perhaps the most notable advance in care delivery has been through the introduction of the Intelligent Home environment [9]. This environment aims to promote the use of embedded technology to support a person within their own living environment. The desired goal is to offer a means of independent living and extend the period of time a person can remain in their own home prior to institutionalisation {2}. Within such an environment, it is common to find two types of devices; sensors and actuators. Sensors are the devices which can record information about the person in the environment. These may be motion sensors detecting which room the person is in, or they may be pressure sensors detecting whether the person is in, for example, a chair or their bed. Other types of sensors include temperature sensors, door sensors, water sensors, appliance sensors etc. Such sensors can provide sufficient information so that the current status of the person can be inferred and subsequently be reused to recommend how the environment should be modified or controlled. The environment itself is managed through the use of actuators. These can, for example, control the ambient temperature if the environment becomes too warm, or raise an alarm if a hazardous situation arises. For example, a person turns on the cooker and then turns on the taps in the bath. The challenge at present is to find the correct balance between the information gathered by the sensors and the ensuing processing to support the dynamic change of the environment through the use of the actuators to support the changing needs of the person.

4.2. Wearable Systems and Smart Clothes

The ability to monitor a number of vital signs from a patient in a pervasive manner has been realised through the introduction of Wearable Computing or Smart Clothes [10]. Smart clothing is a result of the consolidation of sensor technologies, advanced textiles and information and communication technologies [11]. Sensing devices can be embedded into clothing to offer a means of recording information such as heart rate, perspiration levels, body motion / movement and breathing rates. Smart clothing offers the ability to record from a larger surface area on the body, for example the torso, hence the limitations of taking recordings from the finger or the wrist can be avoided [12]. It is possible, during the manufacture of the garment, to fabricate the sensors by either coating a textile with a form of sensing material or to form a sensing material which can then be knitted or woven into the garment. Although many challenges now exist regarding which are the optimal places to locate the recording electrodes [13] and what information should be recorded, smart clothing provides a realistic solution to continuously monitor the person in an unconstrained manner and either provide a means of local processing or relay the information to a central location for assessment by healthcare staff. Smart clothing has huge potential in conjunction with the aforementioned intelligent environments as an augmented means, not only to assess the activities of the person, but also, their health status.

4.3. Data Processing and Context Awareness

We have described a number of paradigms whereby information relating to the patient can be shared between medical professionals or can be viewed remotely by a medical professional to offer a means of support. Given the ability to store large amounts of data about the person, their activities and their general health conditions through a number of vital sign markers, it has now become possible to consider deploying data processing systems with the ability to automatically understand the current status of the person. One of the biggest areas of interest at present is the ability to monitor changes in lifestyle and adapt the technology within the environment to support the changing requirements of the patient. As such, much effort is being directed towards the development of data processing systems which aim not only to detect trends within a person's behaviour but to try to understand the cause of such behaviour and attempt to correlate this with other social or environmental behaviours.

In addition to data processing, the introduction of Context Aware computing within healthcare and intelligent environments has also become more prevalent [14]. The term 'context' refers to any information which can be used to characterise the situation of a person or computational entity. Context aware computing is used to extract, interpret and then use contextual information in such a way that the system or environment can adapt its current state of operation to match the current context of its use.

5. Case Study: Home Based Medication Management

We will now show how some of the aforementioned topics have been addressed in a real project scenario – Home Based Medication Management. Within this context we present an exemplar of how the problem domain has been identified, how a technical

solution has been deployed and how emerging technologies have been integrated into the second generation of the system.

5.1. Problem Domain

It is a well recognised problem that patients do not fully adhere to their medication regimen having received a prescribed course of medication. This results in a significant burden being placed on the healthcare service due to extended patient welfare and healthcare costs. According to the World Health Organisation (WHO) adherence is defined as "the extent to which a person's behaviour – taking medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a healthcare provider" [15]. There have been numerous attempts to quantify the impact of non-adherence to medication, including:

- *one third of people take all their medication, one third take some and one third do not take any at all.*
- *as many as 125,000 deaths may occur each year in the US as a direct result of non-adherence to medication.*
- *up to 23% of admissions to nursing homes can be attributed to non-adherence.*

Adherence to medication is a complex issue and stems from problems such as the patient not understanding their medication regimen, perceived side effects and financial issues [16]. These can also be further complicated by any confusion which may arise through the context in which the medication is prescribed, delivered and taken by the patient.

5.2. Potential Deployment of Technology

Given the huge impact associated with non-adherence, many efforts have been made to deploy technology as a potential means to alleviate the effects of the problem [17]. Technology has been deployed in three different manners. In the first instance, pill holders can be offered to the patient which store the medication in labeled compartments for various times throughout the day. The second type of device is an extension of the first, however, electronic modules support the inclusion of pre-programmable alarms. The third is a monitoring device which provides a remote means to assess if the patient has taken their medication. This last type of device will be the focus of this Case Study.

5.3. The MEDICATE system

The aim of the MEDICATE system was to develop an Internet based care model and associated peripherals to support the needs of all stakeholders within the prescribe to intake chain of medication. The anticipated benefits of the system were the ability to improve and support the patient with the management of their medication in addition to providing a means to support communications between all stakeholders. To achieve this vision it was necessary to develop a suite of interfaces which could be used by each of the stakeholders. The interfaces took the form of both custom electronic devices and software interfaces all connected via an Internet based care model as shown in Figure 4.

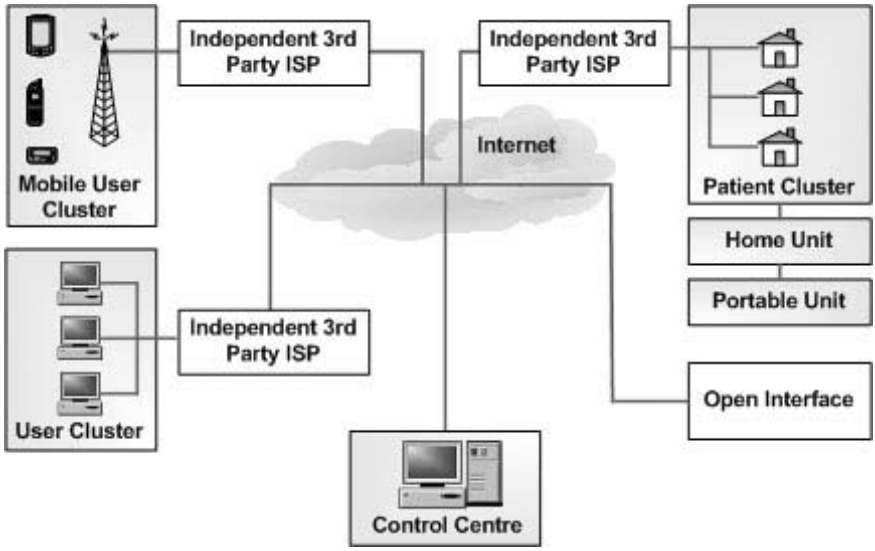


Figure 4 Overview of MEDICATE care model to support home based medication management.

This care model supported communication between the patient within their home and their remote formal or informal carer, in addition to a communication channel between healthcare professionals. Figure 5 below shows the various stakeholder interfaces. Briefly these include:

Patient: a mobile medication management system (Figure 5 (a)) and a base station acting as a reservoir of medication and a means to connect the patient to the Internet portal. These devices had the ability to store the medication and would remind the patient at the appropriate time to take their medication. Any instances of non-adherence would be recorded by these devices

Doctor: a web based interface (Figure 5 (b)) to support the prescribing of medication and assessment of patient adherence. This interface provides a means for the medication regime of the patient to be entered onto the system.

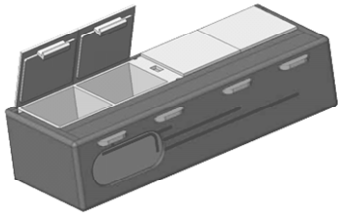
Pharmacist: a software interface (Figure 5 (c)) to support the filling of medication containers to be used by the patient's medication device according to their prescribed medication regimen. This system has the ability to retrieve the information entered onto the system by the Doctor.

Caregiver: this software interface (Figure 5 (d)) provides a means to allow the patient's adherence to their medication regimen to be monitored in real time and, in instances of non-adherence, will raise the attention of the care giving staff.

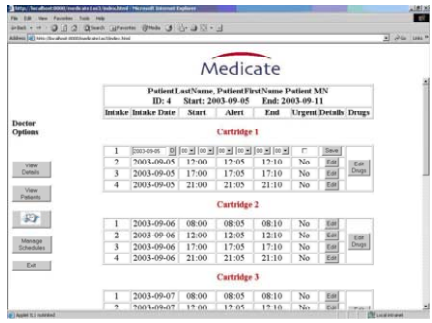
Within this scenario we have demonstrated how the basic concepts of utilising modern day communications have supported the integration of a number of stakeholders in the supply-to-intake chain of medication. Evaluation of this system has demonstrated its usefulness in terms of supporting a patient's adherence to medication.

Nevertheless, it has also highlighted the importance which should be given to the integration of such systems into existing practice.

The natural evolution of this system has involved the deployment of the concepts of medication management onto mobile phone based platforms [17]. These facilitate a similar means of providing management and of monitoring of a patient's intake of medication, in addition to maintaining the link with healthcare professionals. An additional extension to the system would involve integration with context aware services, where the means by which the reminder was offered to the patient would be dependent upon their current activity.

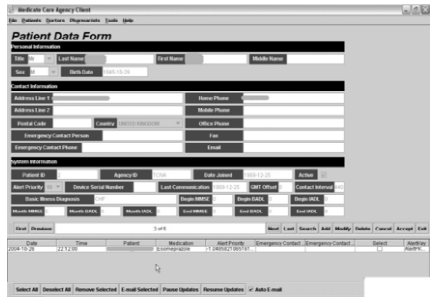


a: Mobile medication device



b: Doctor's interface

c: Pharmacists's interface



d: Caregiver's interface

Figure 5 Stakeholder interfaces for the MEDICATE system.

5.4. Conclusions

Overall, within this review we have shown how technology can be used to establish a number of care paradigms, all of which aim to improve the delivery of care provided to the patient and to overcome geographical boundaries. In some cases existing technological infrastructures may be used and in others new forms of technology are required. Prior to widespread deployment, a number of challenges still remain. These challenges are only partly related to the technology. Other challenges are more related to organizational issues and issues relating to who is actually going to pay for the

technology and its associated services. Nevertheless, it is widely accepted that there are benefits to be accrued, hence it is anticipated that further effort will be directed towards these areas in future years.

Glossary

PSTN Public Switched Telephone Network
DSL Digital Subscriber Line
WHO World Health Organisation

References

[1] Ageing and life course, World Health Organisation, http://www.who.int/ageing/en/ (accessed May 2008).
[2] M.E. Pollack, Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment, *AI Magazine*, (2005), 9-24.
[3] D.J. Cook and S.K. Das, How smart are our environments? An updated look at the state of the art, *Journal of Pervasive and Mobile Computing*, 3 (2007), 53-73.
[4] L. Belochi, Telemedicine Glossary: Glossary of Concepts, Technologies, Standards and Users, 5th ed., European Commission, Brussels, 2003.
[5] H.R. Fischer, S. Reichlin, J.P. Gutzwiller, A. Dyson, C. Beglinger, Telemedicine as a new possibility to improve health care delivery, in M-Health: Emerging Mobile Health Systems, eds. RSH Istepanian, S Laxminarayan, CS Pattichis, (Springer, 2006), 203-218.
[6] L.A. Black, C. McMeel, M. McTear, N. Black, R. Harper, M. Lemon, *Journal of Telemedicine and Telecare*, 11 (2005), 6-8.
[7] M. Drugge, J. Hallberg, P. Parnes, K. Synnes, Wearable systems in nursing home care: prototyping experience, *IEEE Pervasive computing*, (2006), 86-91.
[8] J. Rosen and B. Hannaford, Doc at a distance, *IEEE Spectrum*, 43 (2006), 34-39.
[9] J.C. Augusto and C.D. Nugent, Designing Smart Homes: The Role of Artificial Intelligence, LNAI 4008, Springer, 2006.
[10] F. Axisa, P.M. Schmitt, C. Gehin, G. Delhomme, E. McAdams and A. Dittmar, Flexible technologies and smart clothing for citizen medicine, home healthcare, and disease prevention, *IEEE Transactions on Information Technology in Biomedicine*, 9 (2005), 325-336.
[11] C.D. Nugent, P.J. McCullagh, E.T. McAdams and A. Lymberis, Personalised Health Management Systems: The Integration of Innovative Sesning, Textile, Information and Communication Technologies, (Amsterdam: IOS Press, 2005).
[12] S. Brady, L. Dunne, A. Lynch, B. Smyth and D. Diamond, Wearable Sensors? What is there to sense?, in Personalised Health Management Systems: The Integration of Innovative Sensing, Textile, Information and Communication Technologies, (Amsterdam: IOS Press, 2005), 80-88.
[13] M. Donnelly, C. Nugent, D. Finlay, P. McCullagh and N. Black, Making smart shirts smarter: Optimal electrode placement for cardiac assessment, *International Journal of Assistive Robotics and Mechatronics*, 8 (2007), 53-60.
[14] S.S. Intille, Designing a home of the future, *IEEE Pervasive Computing*, 1 (2002), 76-82.
[15] World Health Organisation (WHO), Adherence to Long-Term Therapies, Evidence for Action, 2003.
[16] L. Blonde, Removing polytherapy as a barrier to adherence, Managed Care, Compliance and Persistence with Medication Therapy, 9 (2000), 1.
[17] C.D. Nugent, D. Finlay, R. Davies, M. Mulvenna, J. Wallace, C. Paggetti, E. Tamburini, N. Black, The next generation of mobile medication management solutions, *International Journal of Electronic Healthcare*, 3 (2007), 7-31.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-185

Introduction to Chapter IV: Biomaterials and Tissue Engineering

Fergal J. O'BRIEN and Brian O'CONNELL (eds.)

The goal of tissue engineering is to develop cell, construct, and living system technologies to restore the structure and functional mechanical properties of damaged or degenerated tissue. The term "tissue engineering", which is interchangeably used with the more recent term of regenerative medicine, was officially coined at a National Science Foundation workshop in 1988 to mean "the application of principles and methods of engineering and life sciences toward fundamental understanding of structure-function relationships in normal and pathological mammalian tissues and the development of biological substitutes to restore, maintain or improve tissue function." While the field of tissue engineering may be relatively new, the idea of replacing tissue with another goes as far back as the 16th century when an Italian, Gasparo Tagliacozzi (1546-99), Professor of Surgery and Anatomy at the Bologna University described a nose replacement that he had constructed from a forearm flap in his work 'De Custorum Chirurigia per Insitionem" (The Surgery of Defects by Implantation) which was published in 1597. In modern times, the techniques of transplanting tissue from one site to another in the same patient (an autograft) or from one individual to another (transplant or allograft) have been revolutionary and lifesaving. However major problems exist with both techniques. Harvesting autografts is expensive, painful, constrained by anatomical limitations and associated with donor-site morbidity due to infection and haemorrhage. Transplants have serious constraints. The major problem is accessing enough tissue and organs for all of the patients who require them. Transplants are strongly associated with rejection by the patient's immune system and they are also limited by the potential risks of introducing infection or disease.

The field of tissue engineering is highly multidisciplinary and draws on experts from mechanical engineering, materials science, surgery, genetics, and related disciplines from engineering and the life sciences. Tissue engineering technologies are based on a biological triad and involve the successful interaction between three components: (1) the scaffold that holds the cells together to create the tissue's physical form i.e. acts as a template for new tissue formation, (2) the cells that synthesise the tissue and (3) signalling mechanisms (i.e. mechanical and/or chemical signals) that direct the cells to express the desired tissue phenotype. The specific properties of each of these three components is critically important in determining the quality of the

engineered tissue and therefore the potential for clinical success when the engineered tissue has been implanted into an injured site in vivo. For example, the choice of biomaterial and the fabrication process used to produce the porous scaffold which will act as the template for in vitro tissue formation should be selected based on the specific tissue type and anatomical site in which it will be implanted. Similarly, consideration must be given to the choice of cell type used, for example, whether the cells should be autologous or allogeneic and whether stem cells or primary mature cells should be used. Finally, the process of cellular differentiation into a specific phenotype followed by the production of extracellular matrix is stimulated by either the provision of specific growth factors and cytokines and/or by subjecting a cell-seeded construct to biophysical stimulation in the form of a bioreactor. The types of growth factors, choice of bioreactor and duration of exposure all play a key role in determining the success of the engineered tissue. The mechanical properties of the scaffold and engineered tissue will govern the ability of the construct to function in vivo and to allow infiltration of host cells and in most tissues (cartilage being an exception), vascularisation. It is therefore important to develop techniques to test biomaterial scaffold and tissue properties prior to implantation and to be able to predict how the constructs will behave when implanted in vivo.

This chapter focuses on all three areas of the tissue engineering triad: Scaffolds & Surfaces, Cellular & Molecular Biomechanics and Bioreactors in Tissue Engineering. In addition, the fourth part of the chapter is devoted to Characterisation and Testing of Biomaterials. It is pitched at a level that should allow either an engineer or a clinician to understand the basic principles of tissue engineering and the fundamental interaction between cells & scaffolds and their response to biophysical stimuli and the different types of bioreactor that exist. The chapter finishes with a section which discusses the fundamentals of materials testing which will be of interest to clinicians and biologists involved not only in tissue engineering, but in any area of biomechanical analysis. The contributors are all experts in their fields and are involved in research in tissue engineering and a number of cognate disciplines and come from all across Europe. They include an anatomist/bioengineer (O'Brien), dentist (O'Connell), materials scientist (Partap), orthopaedic surgeon (Lyons), physiologist (Campbell), engineers (Plunkett, Dendorfer and Hammer) and trauma surgeon (Lenich).

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-187

IV.1. Scaffolds & Surfaces

Sonia PARTAP [a,b], Frank LYONS [a,b] and Fergal J. O'BRIEN [a,b]
[a] *Department of Anatomy, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland*
[b] *Trinity Centre for Bioengineering, Department of Mechanical Engineering, Trinity College Dublin, Dublin 2, Ireland*

Introduction

Every day thousands of clinical procedures are performed to replace or repair tissues in the human body that have been damaged through disease or trauma. Current therapies are focused on the replacement of the damaged tissue by using donor graft tissues (autografts, allografts or xenografts). Problems associated with this approach include shortage of donors or donor sites, the volume of donor tissue that can be safely harvested, donor site pain and morbidity, the possibility of harmful immune responses, transmission of disease and rejection of grafts [1]. Alternatively, the field of tissue engineering (a phrase that is interchangeably used with regenerative medicine) aims to regenerate damaged tissues instead of replacing them (with grafts) by developing biological substitutes that restore, maintain or improve tissue function [2, 3].

In native tissues, cells are held within an extracellular matrix (ECM) which guides development and directs regeneration of the tissue, serves to organise cells in space and provides them with environmental signals to direct cellular behaviour. The goal of tissue engineering is to synthesise substitutes that mimic the natural ECM to help guide the growth of new functional tissue *in vitro* or *in vivo*. At a simplistic level, biological tissues consist of cells, signalling mechanisms and extracellular matrix. Tissue engineering technologies are based on this biological triad and involve the successful interaction between three components: (1) the scaffold that holds the cells together to create the tissues physical form, (2) the cells that create the tissue and, (3) the biological signalling molecules (such as growth factors) that direct the cells to express the desired tissue phenotype (Figure 1). Tissue engineering is a multidisciplinary field that harnesses expertise and knowledge from a variety of fields, including those of the medical profession, materials scientists, engineers, chemists and biologists.

Why are Scaffolds Required ? 2D vs. 3D Culture

There are differences in cell behavior in three dimensional (3-D) *vs.* two dimensional (2-D) environments. *In vitro* 3-D cell culture conditions more accurately model *in vivo* biological responses, as the conditions more closely resemble the natural structure and function of tissues *in vivo* [4]. These conditions can be created by using a 3-D scaffold that acts as a template, allowing cells to produce and deposit extracellular matrix (ECM) that would not be possible in 2-D environments. 2-D cell culture does not allow cells to move or assemble with the freedom they have *in vivo*, and thus cannot replicate

the effects of nutrient gradients, signal propagation or the development of bulk mechanical properties. Studying these cells in 3-D models allows us to better understand their biochemical and biophysical signaling responses as they would normally occur *in vivo,* particularly the external signals occurring in the ECM, as well as the mechanical and chemical signals arising from both adjacent and even distant cells [5]. This approach can lead to the generation of more accurate cell- based assays for engineering of suitable biomaterials that can be used to determine the cell-material interaction.

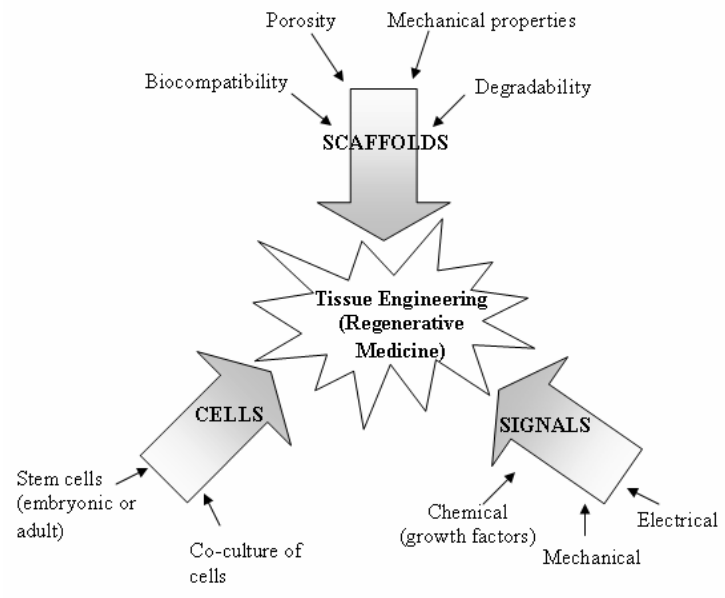


Figure 1. The tissue engineering triad; factors that need to be considered when designing a suitable structure for tissue engineering applications.

1. Properties of Scaffolds for Tissue Engineering

All scaffolds for tissue engineering applications are designed to perform the following functions: (1) to encourage cell-material interactions *i.e.* cell attachment, differentiation and proliferation, eventually leading to the deposition of extracellular matrix, (2) to permit the transport of nutrients, wastes and biological signalling factors to allow for cell survival, (3) to biodegrade at a controllable rate which approximates the rate of natural tissue regeneration, and (4) to provoke a minimal immune and/or inflammatory response *in vivo*. The following parameters must be considered when designing a scaffold for tissue engineering.

1.1. Biocompatibility

The implantation of a scaffold may elicit different tissue responses depending on the composition of the scaffold. If the scaffold is non-toxic and degradable, new tissue will eventually replace it; if it is non-toxic and biologically active then the scaffold will integrate with the surrounding tissue. However, if the scaffold is biologically inactive, it may be encapsulated by a fibrous capsule, and in the worst case scenario if the scaffold is toxic, rejection of the scaffold and localised death of the surrounding tissue can occur [6]. Biocompatibility is the ability of the scaffold to perform in a specific application without eliciting a harmful immune or inflammatory reaction. For a scaffold to positively interact with cells and with minimal disruption to the surrounding tissue, it should have an appropriate surface chemistry to allow for cellular attachment, differentiation and proliferation. Cells primarily interact with scaffolds *via* chemical groups on the material surface or topographical features. Topographical features include surface roughness and pores where cell attachment is favoured. Alternatively, cells may recognise and subsequently bind to the arginine-glycine-aspartic acid (RGD) cell adhesion ligand. Scaffolds synthesised from natural extracellular materials (*e.g.* collagen) already possess this specific ligands, whereas scaffolds made from synthetic materials may be designed to deliberately incorporate them.

1.2. Biodegradability

The severity of an immune or inflammatory reaction is not only determined by the actual scaffold itself, but is also dependent on the scaffold's degradation products. Ideally, scaffolds are designed to be completely replaced by the regenerated extracellular matrix by integrating with the surrounding tissue, eliminating the need for further surgery to remove it [7]. Scaffolds should degrade with a controllable degradation rate, (approximating the rate of natural tissue regeneration), as well as with controllable degradation products. As it degrades, the breakdown products should be non-toxic and easily excreted from the body *via* metabolic pathways or the renal filtration system [8].

1.3. Mechanical Properties

The scaffold provides structural integrity to the engineered tissue in the short term. Furthermore, it provides a framework for the three dimensional (3-D) organisation of the developing tissue as well as providing mechanical stability to support the growing tissue during *in vitro* and/or *in vivo* growth phases [9]. The mechanical properties of the scaffold should be designed to meet the specific requirements of the tissue to be regenerated at the defect site. Furthermore, at the time of implantation, the scaffold should have sufficient mechanical integrity to allow for handling by the clinician, be able to withstand the mechanical forces imposed on it during the implantation procedure and survive under physiological conditions. Immediately after implantation, the scaffold should provide a minimal level of biomechanical function that should progressively improve until normal tissue function has been restored, at which point the construct should have fully integrated with the surrounding host tissue.

1.4. Scaffold Architecture

Porous structures allow for optimal interaction of the scaffold with cells. The pore architecture is characterised by pore size and shape, pore interconnectivity/tortuosity, degree of porosity and surface area. The microstructure determines cell interactions with the scaffold, as well as molecular transport (movement of nutrients, wastes and biological chemicals *e.g.* growth factors) within the scaffold. Specifically, pore size determines the cell seeding efficiency into the scaffold [10]; very small pores prevent the cells from penetrating the scaffold, whilst very large pores prevent cell attachment due to a reduced area and therefore, available ligand density. Subsequently, cell migration within a scaffold is determined by degree of porosity and pore interconnectivity/tortuosity. A scaffold with an open and interconnected pore network, and a high degree of porosity (>90 %) is ideal for the scaffold to interact and integrate with the host tissue [11].

1.5. Manufacturing Technology

In order for a scaffold or engineered construct to become commercially available in a clinical setting, the cost effectiveness of it should be considered; particularly when it is to be scaled up from making one at a time in a research laboratory to a production process allowing small batch quantities of 100 to 1000 constructs to be made. In addition, as clinicians ideally would prefer “off the shelf” products that may be used routinely, it is important to take into consideration how the constructs will be transported and stored in clinical environments. The cost effectiveness will be determined by the choice of biomaterial, which will in turn affect the selection of fabrication method. Many different techniques have been used to fabricate scaffolds for tissue engineering (Figure 2). The following summarises the most commonly used methods.

1.5.1. Particulate Leaching Methods

Particulate leaching is a technique that uses solid particles of a particular size to act as a template for the pores; water soluble particles are frequently used as they can easily be leached out of the final product by simply washing the final product with water. In solvent casting-particulate leaching, a polymer dissolved in a solvent is mixed with salt particles in a mould; the solvent is then evaporated to give a polymer monolith embedded with the salt particles, these are then removed by washing the scaffold with water, resulting in the formation of a porous scaffold [12]. Another variation of this technique is melt moulding-particulate leaching: in this particular technique the polymer is cast into a mould with the embedded solid porogen. The polymer is set by applying heat and pressure, and again the porogen is leached away by washing the resulting product with water to yield a porous polymer scaffold [13].

1.5.2. Phase Separation

Various forms of phase separation techniques enable the creation of porous structures. A two phase polymer system that is homogenous can become thermodynamically unstable by altering the temperature leading to (1) liquid/liquid or (2) liquid/solid phase separations. In the first, a polymer is dissolved in a molten solvent, a liquid/liquid phase separation (where one phase is concentrated in polymer whilst the other is not) is achieved by lowering the temperature. The two phase liquid is quenched to yield a two

phase solid, and the solvent is then removed yielding a porous polymer, this is known as thermally induced phase separation (TIPS) [14]. In the second, a polymer is dispersed in a solvent which is then frozen to induce crystallisation of the solvent to form solvent crystals that act as templates for the pores. These crystals are then removed by freeze drying to yield a porous foam. Manipulation of the processing conditions enables the creation of different pore sizes and distributions [15].

1.5.3. Foaming

Foaming techniques use gaseous porogens that are produced by chemical reactions during polymerisation, or are generated by the escape of gases during a temperature increase or drop in pressure. Nam *et al.* 2000 [16] synthesised poly (lactic acid) [PLA] scaffolds using ammonium bicarbonate which acted as both a gas foaming agent and as a solid salt porogen, an increase in temperature caused the formation of carbon dioxide and ammonia to create a highly porous foam. Also, high pressure carbon dioxide can be used to foam polymers by saturating a prefabricated polymer monolith. A subsequent reduction in pressure causes a decrease in solubility of the carbon dioxide within the polymer, and as the carbon dioxide gas tries to escape it causes the nucleation and growth of bubbles resulting in a porous microstructure [17].

1.5.4. Emulsion Templating

Porous structures can also be obtained by using emulsion templating techniques. The internal phase of the emulsion acts as a template for the pores whilst polymerisation occurs in the continuous phase (in which the monomer is dissolved). After polymerisation, the internal phase is removed to give a templated porous material. The resulting porous microstructures are replicas of the internal phase droplets around which polymerisations were performed. The size of the emulsion droplets is preserved, producing polymer foams with approximately the same size and size distributions as that of the emulsion at the point of polymerisation [18].

1.5.5. Solid Free Form (SFF) Fabrication

Solid free form (SFF) fabrication or rapid prototyping (RP) technologies uses layer manufacturing techniques to create three dimensional scaffolds directly from computer generated files. There are a few techniques that come under this group including stereolithography, selective laser sintering, fused depositional modeling and three dimensional printing. However, all the techniques share the same principle where powders or liquids are solidified one layer at a time to gradually build a three-dimensional scaffold. The layering is controlled by computer assisted design (CAD) programs where the scaffold architecture is designed and modelled. Data collected from computed tomography (CT) or magnetic resonance imaging (MRI) scans may also be used to create CAD models that are specific to the tissue to be regenerated [19].

1.5.6. Combination of Techniques

The techniques discussed above can also be combined with each other depending on the exact requirements of the scaffold, *e.g.* phase separation (freeze drying) techniques can be combined with emulsion templating processes. Whang *et al.* 1995 created an emulsion that was quenched using liquid nitrogen, which was then freeze dried to produce porous PLGA polymeric monoliths [20].

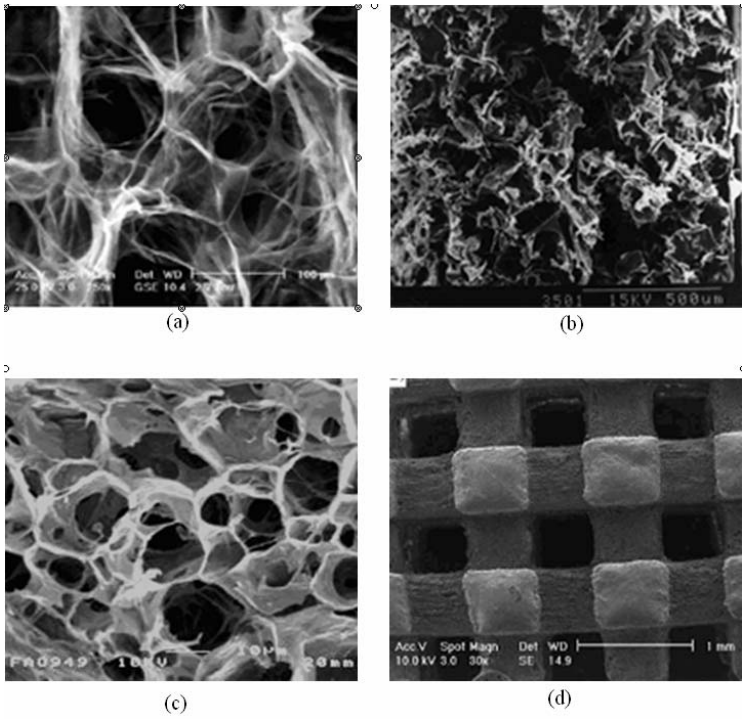


Figure 2 Scanning electron microscopy images of porous (a) collagen-GAG scaffolds made by freeze drying[15], (b) poly-L-lactide (PLLA) foams made by solvent casting-particulate leaching[12], (c) alginate scaffolds made by emulsion templating[18] and (d) polycaprolactone–calcium phosphate composites made by solid free form fabrication methods[51].

2. Biomaterials in Tissue Engineering

A number of different categories of biomaterials are commonly used as scaffolds for tissue engineering.

2.1. Ceramics

Ceramics (inorganic, non metallic materials) used within the biomedical field are classified as being either bioinert or bioactive. The bioinert ceramics include materials such as alumina and zirconia that are typically used as implants for musculoskeletal, oral and maxillofacial applications whilst the bioactive group include the calcium phosphates, the bioglasses and glass-ceramics [6]. All bioceramics are also further defined as being osteoconductive (supporting bone growth) or osteoinductive (stimulating bone growth); all types of bioceramics are osteoconductive as all support the formation of bone, but not all are osteoinductive. The calcium phosphate based bioceramics, bioglasses and glass-ceramics are commonly used as scaffolds for bone tissue engineering as they have a compositional similarity to the mineral phase of bone

[21]. Hydroxyapatite (HA) and tri-calcium phosphate (TCP) are two of the most commonly used calcium phosphate bioceramics in tissue engineering applications. TCP is used as a degradable scaffold, whilst HA, which is non-resorbable and has the added advantage of being osteoinductive, is typically used for coating biomedical implants to induce bone regeneration, allowing the implant to integrate with the surrounding tissue. For this reason, HA has shown much popularity for use as a scaffold for tissue engineering.

2.2. Synthetic Polymers

The mechanical, physical and biological properties of synthetic polymers can be tailored to give a wide range of controllable properties that are more predictable than materials obtained from natural sources. The advantage of using synthetic materials is that the resulting properties can be customised by adjusting the ratios of the monomer units (basic building blocks of the final polymer) and by the incorporation of specific groups (*e.g.* RGD peptide that cells can recognise). Also, the degradation rate and products can be controlled by the appropriate selection of the segments to form breakdown products that can either be metabolised into harmless products or can be excreted *via* the renal filtration system [8]. Among the many biodegradable synthetic polymers used for tissue engineering applications, there are numerous reports on the use of polylactic acid (PLA), polyglycolic acid (PGA) and their copolymers poly (DL-lactic-co-glycolic acid) (PLGA), which are approved by the US Food and Drug Administration (FDA). These polymers degrade by hydrolytic mechanisms and are commonly used because their degradation products can be removed from the body as carbon dioxide and water. However, a disadvantage is that there is a lowering of the pH in the localised region resulting in inflammatory responses when they do degrade. Polycaprolactone (PCL) has a very similar structure to PLA and PGA and is also degraded *via* hydrolytic mechanisms under physiologic conditions (Figure 3). In addition, it is degraded enzymatically and the resulting low molecular weight fragments are reportedly taken up by macrophages and degraded intracellularly. It is predominantly used for drug delivery devices because it has a slower degradation rate than PGA and PLA. However, more recently, it is increasingly finding applications in tissue engineering [22]. Traditionally, polyurethanes were used in the biomedical field as blood contacting materials for cardiovascular devices, and were intended to be used as non-degradable coatings. More recently they have been designed to be biodegradable by being combined with degradable polymers such as PLA for soft tissue engineering applications [14]. Poly(ethyleneglycol) [PEG] is a biocompatible, non-toxic, water soluble polymer that is a liquid at cold temperatures and elastic gel at 37 ^{o}C [23]. PEG based copolymers have been used as injectable scaffolds for bone as well as for drug delivery applications [24]. Also, copolymers of PEG and PLA have been created where the degradation rate and hydrophilicity could be controlled by adjusting the ratio of the hydrophilic (PEG) to hydrophobic (PLA) blocks.

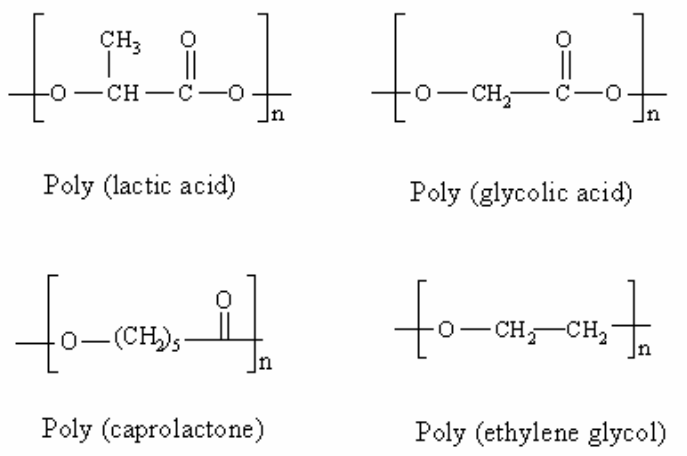


Figure 3 Chemical structures of some biodegradable synthetic polymers used as scaffolds in tissue engineering applications

2.3. Natural Polymers

Natural polymers offer an alternative to synthetic polymer systems (which intrinsically lack cell recognition signals) as they can more closely mimic the natural extracellular matrix of tissues. Alginate and chitosan are two natural polysaccharides that do not exist within the human body but have been investigated for tissue engineering applications because they are structurally similar to the glycosaminoglycans (GAGs) found in the natural extracellular matrix of tissues *i.e.* skin, bone, and blood vessels. Alginate originates from seaweed and is attractive because of its low toxicity, water solubility and its simple gelation chemistry with calcium ions. Alginate hydrogels have been investigated for use as scaffolds for cartilage [25] and liver regeneration [26], as well as for wound dressings [27]. Chitosan is a derivative of naturally occurring chitin which is found in the exoskeletons of crustaceans. It has a low toxicity and is biocompatible. Chitosan scaffolds have been investigated for skin and bone tissue engineering [28].

Given the importance of GAGs in stimulating normal tissue growth, the use of GAGs as components of a scaffold for tissue engineering appears to be a logical approach for scaffold development. Hyaluronic acid (sometimes referred to as hyaluronan) is one of the largest GAG components found in the natural extracellular matrix of all soft tissues and synovial fluid of joints [29]. The applications of pure hyaluronic acid in tissue engineering applications are limited because of its easy dissolution in water and fast biodegradation in biological environments. However, it can be chemically modified to produce a more hydrophobic molecule, thus reducing its solubility in water. Hyaluronic acid scaffolds are known to be biocompatible, and cells easily adhere to and proliferate on this material. Hyaluronic acid also plays a significant role in wound healing and can be modified for drug delivery applications.

Structural proteins such as fibrin are also utilised in tissue engineering applications. Fibrin can be used as a natural wound healing material, and has found applications as a sealant and adhesive in surgery. It can be produced from the patient's own blood, to be used as an autologous scaffold. However, the stability of the material is limited as it can be easily degraded unless apronitin, a protein inhibitor, is used to control the rate of degradation. Fibrin hydrogels have been used to engineer tissues with smooth muscle cells [30] and chondrocytes [31]. Alternatively, gelatin (a derivative of collagen) that is produced by altering the helical structure of the collagen molecule by breaking it into single strand molecules) has been investigated for cartilage tissue regeneration. [32]. However, as one of the main disadvantages of gelatin is its poor mechanical strength, it has also been cross-linked with hyaluronic acid for skin tissue engineering, and with alginate for wound healing applications [33]. Instead, collagen, the main component found in the extracellular matrix of mammalian connective tissues has found use in tissue engineering applications including skin substitutes [34], scaffolds for bone and cartilage, vascular applications and as drug delivery systems. As is typical of all natural polymers, collagen gels also display poor mechanical properties. However, these can be improved by employing both chemical and physical crosslinking methods. Physical crosslinking methods include UV radiation and dehydrothermal treatments, whilst cross-linking agents such as glutaraldehyde and carbodiimides (EDAC) can be used to produce chemically cross-linked collagen hydrogels with improved physical properties (Figure 4).

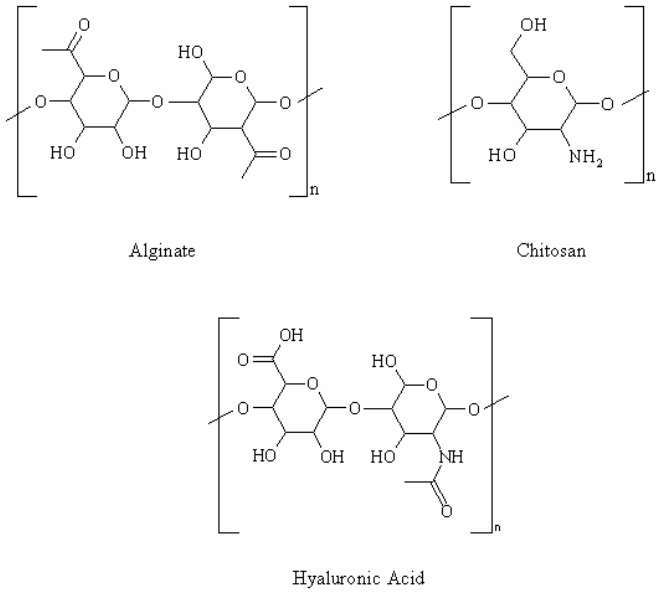


Figure 4 Chemical structures of some natural polymers used as scaffolds in tissue engineering applications

2.4. Composites

Due to some of the problems associated with using scaffolds synthesised from a single phase biomaterial (*eg.* poor mechanical properties and biocompatibility of natural and synthetic polymers respectively, and poor degradability of bioceramics), a number of researchers have developed composite scaffolds comprising of two or more phases to combine the advantageous properties of each phase. For example, polymer/ceramic composites of poly(lactic-*co*-glycolic acid (PLGA) and hydroxyapatite have been investigated for tissue engineering applications [35], whilst Cui *et al.* [36] have produced tri-phasic scaffolds by depositing nano-hydroxyapatite particles onto cross-linked collagen-chitosan matrices. However, even though composite scaffolds such as these have shown some promise as grafts for bone and cartilage, each one consists of at least one phase which is not found naturally in the body and therefore has problems with either biocompatibility or biodegradability or both. Table 1 summarises the different types of biomaterials described above and lists the advantages and disadvantages of each type for use as scaffolds in tissue engineering applications.

2.5. Case study: Collagen Scaffolds for Bone Tissue Engineering

From an engineering viewpoint, bone is a composite material made up of both organic and inorganic phases embedded with bone cells and blood vessels. The main components of the organic and inorganic phases are collagen and hydroxyapatite, respectively. The collagen fibres impart tensile strength to the bone whilst the HA crystals contribute to its stiffness. Based on this, collagen scaffolds are currently being investigated for bone tissue engineering applications. In our laboratory, we are currently using porous collagen-glycosaminoglycan (CG) composite scaffolds which are produced using a lyophilisation (freeze drying) process. The final pore microstructure of the scaffolds can be varied by controlling the rate and temperature of freezing during fabrication and the volume fraction of the precipitate [15]. We have shown that by varying the final freezing temperature during the lyophilisation process a homologous series of scaffolds with a constant composition and solid volume fraction with distinctly different pore sizes can be produced [10]. Additionally, experiments performed in our laboratory using osteoblasts demonstrated that the fraction of cells attaching to the scaffold decreased with increasing mean pore diameter, indicating that scaffold ligand density is affected by pore size where an increase in ligand density causes increased cell attachment. In another study, we have shown that collagen-based scaffolds seeded with rat mesenchymal stem cells promoted differentiation along osteogenic and chondrogenic lineages demonstrating their potential for orthopaedic applications [37]. There is also evidence to suggest that non-seeded collagen scaffolds with incorporated growth factors implanted into defects induce bone formation [38]. A problem with collagen-based scaffolds, as with most natural polymer scaffolds, is their poor mechanical properties. However, these can be improved through physical and chemical crosslinking methods [39], and allowing bone cells to produce osteoid on the scaffolds, enabling them to subsequently mineralise the scaffold *in vitro* prior to implantation, also leads to improved mechanical properties. Alternatively, as bioceramics are mechanically stronger and are known to enhance osteoblast differentiation and proliferation, they have been combined with collagen scaffolds to form mineralised collagen scaffolds that support cell growth [40, 41].

Table 1 Properties, advantages and disadvantages of biomaterials used as scaffolds in tissue engineering applications

Scaffold	Properties	Advantages	Disadvantages
Bioceramics			
Hydroxyapatite (HA)	Found naturally as a component of mineral phase of bone	Biocompatible Osteoinductive	Non resorbable Poor mechanical properties
Tricalcium Phosphate (TCP)	Compositional similarity to mineral phase of bone	Biocompatible Biodegradable	Poor mechanical properties
Synthetic polymers			
Poly(lactic acid), Poly(glycolic acid) and their copolymers	Mechanical and degradation properties can be tuned by varying polymer segments	Biocompatible	Degradation products are CO_2 and H_2O creating local acidic conditions
Poly(ethylene glycol)	Used as an injectable gel Mechanical and degradation properties can be tuned by varying polymer segments	Biocompatible Hydrophilic	Poor cell adhesion
Natural polymers			
Collagen	Component of natural extracellular matrix (ECM)	Biocompatible Good cell recognition	Poor mechanical properties
Hyaluronic acid	Plays role in natural wound healing Component of natural ECM	Biocompatible Easily functionalized Good cell recognition	Poor mechanical properties
Alginate	Originates from seaweed Structurally similar to natural glycosaminoglycan's (GAG)	Biocompatible Simplegelation methods	Poor mechanical properties
Composites			
Polymer - Ceramic	Natural or synthetic polymers combined with ceramics Often combined for bone tissue engineering applications	Ability to tailor mechanical, degradation and biological properties	Compromise between 'best' qualities of individual components with overall scaffold properties
Polymer - Polymer	Combinations of (1) synthetic-synthetic, (2) synthetic – natural and (3) natural – natural polymers possible	Ability to tailor mechanical, degradation and biological properties	Compromise between 'best' qualities of individual polymers with overall scaffold properties

There are also reports of triphasic scaffolds made from collagen, a bioceramic and a synthetic polymer. Scaffolds made from nano-HA, collagen and PLA were placed in defects of rabbit radius and they integrated with the defect site within 12 weeks [42]. These studies indicate that by finding an adequate balance between pore structure, mechanical properties and biocompatibility, a collagen-based construct can potentially support bone growth and may have real potential for bone tissue engineering.

3. Scaffolds: State of the Art and Future Directions

Economic activity within the tissue engineering sector has grown five-fold in the past 5 years. In 2007, approximately 50 companies offered commercially available tissue-regenerative products or services, with annual sales recorded in excess of $1.3 billion, whilst 110 development-stage companies with over 55 products in FDA-level clinical trials and other preclinical stages spent $850 million on development [43]. The tissue engineering approach was originally conceived to address the gap between patients waiting for donors and the amount of donors actually available. To date the highest rates of success have been achieved in the areas of skin regeneration where tissue-engineered substitutes have been successfully used in patients [44].

However, much research still remains to be performed in all aspects of tissue engineering [45]. Cellular behaviour is strongly influenced by signals (biochemical and biomechanical) from the extracellular matrix, the cells are constantly receiving cues from the extracellular matrix about their environment and are constantly remodelling it accordingly. Therefore, an appropriate three dimensional structure that is predominantly thought of as playing a mechanical role is not enough to promote the growth of new tissue. It is important that the scaffold provides adequate signals (*e.g.* through the use of adhesion peptides and growth factors) to the cells, to induce and maintain them in their desired differentiation stage, for their survival and growth [46]. Thus, equal effort should be made in developing strategies on how to incorporate the adhesion peptides and growth factors into the scaffolds, as well as in identifying the chemical identity of adhesion peptides and growth factors that influence cell behaviour, along with the distributions and concentrations required for successful outcomes. An example would be to incorporate angiogenic growth factors in scaffolds for different types of tissue in an attempt to generate vascularised tissues. Tissue vascularisation can be used to establish blood flow through the engineered tissues and strategies involving the incorporation of vasculature, as well as innervation will be of great importance [47]. Additionally, the incorporation of drugs (*i.e.* inflammatory inhibitors and/or antibiotics) into scaffolds may be used to prevent any possibility of an infection after surgery [48].

The field of biomaterials has played a crucial role in the development of tissue engineered products. An alternative to using prefabricated scaffolds is to use a polymer system that is injected directly into the defect site which is polymerised *in situ* using either heat [49] (thermoresponsive polymers) or light [50] (photoresponsive polymers). The advantages for the patient with this approach over current therapies are that injectable delivery systems fill both regularly and irregularly shaped defects (“get a custom fit”), they represent a minimally invasive procedure therefore avoiding surgery and the potential risks associated with it, eliminate the need for donor tissue or a donor site, and waiting time for treatment is reduced, as it can be used whenever treatment is required.

At present, there is a vast amount of research being performed on all aspects of tissue engineering/regenerative medicine worldwide. Thus, as the field progresses, one of the key challenges is to try to mimic the sophistication of the natural ECM more accurately in synthetic substitutes. As more advanced biomaterials and bioreactors are developed, and as research leads to more knowledge on the cell signaling mechanisms required to trigger the chain of tissue development, we will undoubtedly get closer towards our goal of reducing the number of patients waiting for donor tissues.

References

[1] R. Langer, Biomaterials in drug delivery and tissue engineering: One laboratory's experience, *Acc Chem Res* **33** (2000), 94-101.

[2] A. Atala, Tissue engineering and regenerative medicine: Concepts for clinical application, *Rejuvenation Res* **7** (2004), 15-31.

[3] L. J. Bonassar and C. A. Vacanti, Tissue engineering: The first decade and beyond, *J Cell Biochem Suppl* **30-31** (1998), 297-303.

[4] M. P. Lutolf and J. A. Hubbell, Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering, *Nat Biotechnol* **23** (2005), 47-55.

[5] L. G. Griffith and M. A. Swartz, Capturing complex 3d tissue physiology in vitro, *Nat Rev Mol Cell Biol* **7** (2006), 211-24.

[6] L. L. Hench, Bioceramics, J. Am. Ceram. Soc **81** (1998), 1705-28.

[7] J. E. Babensee, A. G. Mikos, J. M. Anderson and L. V. McIntire, Host response to tissue engineered devices, *Adv. Drug Del. Rev.* **33** (1998), 111-139.

[8] W. E. Hennink and C. F. van Nostrum, Novel crosslinking methods to design hydrogels, *Adv Drug Deliv Rev* **54** (2002), 13-36.

[9] D. W. Hutmacher, Scaffolds in tissue engineering bone and cartilage, *Biomaterials* **21** (2000), 2529-2543.

[10] F. J. O'Brien, B. A. Harley, I. V. Yannas and L. J. Gibson, The effect of pore size on cell adhesion in collagen-gag scaffolds, *Biomaterials* **26** (2005), 433-41.

[11] T. M. Freyman, I. V. Yannas and L. J. Gibson, Cellular materials as porous scaffolds for tissue engineering, *Prog. Mater Sci.* **46** (2001), 273-282.

[12] L. Lu, S. J. Peter, M. D. Lyman, H. L. Lai, S. M. Leite, J. A. Tamada, S. Uyama, J. P. Vacanti, R. Langer and A. G. Mikos, in vitro and in vivo degradation of porous poly(dl-lactic-co-glycolic acid) foams, *Biomaterials* **21** (2000), 1837-45.

[13] S. H. Oh, S. G. Kang, E. S. Kim, S. H. Cho and J. H. Lee, Fabrication and characterization of hydrophilic poly(lactic-co-glycolic acid)/poly(vinyl alcohol) blend cell scaffolds by melt-molding particulate-leaching method, *Biomaterials* **24** (2003), 4011-21.

[14] A. S. Rowlands, S. A. Lim, D. Martin and J. J. Cooper-White, Polyurethane/poly(lactic-co-glycolic) acid composite scaffolds fabricated by thermally induced phase separation, *Biomaterials* **28** (2007), 2109-21.

[15] F. J. O'Brien, B. A. Harley, I. V. Yannas and L. Gibson, Influence of freezing rate on pore structure in freeze-dried collagen-gag scaffolds, *Biomaterials* **25** (2004), 1077-86.

[16] Y. S. Nam, J. J. Yoon and T. G. Park, A novel fabrication method of macroporous biodegradable polymer scaffolds using gas foaming salt as a porogen additive, *J Biomed Mater Res* **53** (2000), 1-7.

[17] D. J. Mooney, D. F. Baldwin, N. P. Suh, J. P. Vacanti and R. Langer, Novel approach to fabricate porous sponges of poly(d,l-lactic-co-glycolic acid) without the use of organic solvents, *Biomaterials* **17** (1996), 1417-22.

[18] S. Partap, J. A. Darr, I. U. Rehman and J. R. Jones, "Supercritical carbon dioxide in water" Emulsion templated synthesis of porous calcium alginate hydrogels, *Adv. Mater.* **18** (2006), 501-504.

[19] E. Sachlos and J. T. Czernuszka, Making tissue engineering scaffolds work. Review: The application of solid freeform fabrication technology to the production of tissue engineering scaffolds, *Eur Cell Mater* **5** (2003), 39-40.

[20] K. Whang, C. H. Thomas, K. E. Healy and G. Nuber, A novel method to fabricate bioabsorbable scaffolds, *Polymer* **36** (1995), 837-842.

[21] K. A. Hing, Bioceramic bone graft substitutes: Influence of porosity and chemistry, *Int. J. Appl. Ceram. Technol*. **2** (2005), 184-199.

[22] L. Savarino, N. Baldini, M. Greco, O. Capitani, S. Pinna, S. Valentini, B. Lombardo, M. T. Esposito, L. Pastore, L. Ambrosio, S. Battista, F. Causa, S. Zeppetelli, V. Guarino and P. A. Netti, The performance of poly-epsilon-caprolactone scaffolds in a rabbit femur model with and without autologous stromal cells and bmp4, *Biomaterials* **28** (2007), 3101-9.
[23] P. J. Martens, S. J. Bryant and K. S. Anseth, Tailoring the degradation of hydrogels formed from multivinyl poly(ethylene glycol) and poly(vinyl alcohol) macromers for cartilage tissue engineering, *Biomacromolecules* **4** (2003), 283-92.
[24] F. Chen, T. Mao, K. Tao, S. Chen, G. Ding and X. Gu, Injectable bone, *Br J Oral Maxillofac Surg* **41** (2003), 240-3.
[25] W. J. Marijnissen, G. J. van Osch, J. Aigner, S. W. van der Veen, A. P. Hollander, H. L. Verwoerd-Verhoef and J. A. Verhaar, Alginate as a chondrocyte-delivery substance in combination with a nonwoven scaffold for cartilage tissue engineering, *Biomaterials* **23** (2002), 1511-7.
[26] J. Yang, M. Goto, H. Ise, C. S. Cho and T. Akaike, Galactosylated alginate as a scaffold for hepatocytes entrapment, *Biomaterials* **23** (2002), 471-9.
[27] D. Bettinger, D. Gore and Y. Humphries, Evaluation of calcium alginate for skin graft donor sites, *J Burn Care Rehabil* **16** (1995), 59-61.
[28] C. Mao, J. J. Zhu, Y. F. Hu, Q. Q. Ma, Y. Z. Qiu, A. P. Zhu,W. B. Zhao and J. Shen, Surface modification using photocrosslinkable chitosan for improving hemocompatibility, *Colloids Surf B Biointerfaces* **38** (2004), 47-53.
[29] J. L. Drury and D. J. Mooney, Hydrogels for tissue engineering: Scaffold design variables and applications, *Biomaterials* **24** (2003), 4337-4351.
[30] C. L. Cummings, D. Gawlitta, R. M. Nerem and J. P. Stegemann, Properties of engineered vascular constructs made from collagen, fibrin, and collagen-fibrin mixtures, *Biomaterials* **25** (2004), 3699-706.
[31] C. J. Hunter, J. K. Mouw and M. E. Levenston, Dynamic compression of chondrocyte-seeded fibrin gels: Effects on matrix accumulation and mechanical stiffness, *Osteoarthritis Cartilage* **12** (2004), 117-30.
[32] M. S. Ponticiello, R. M. Schinagl, S. Kadiyala and F. P. Barry, Gelatin-based resorbable sponge as a carrier matrix for human mesenchymal stem cells in cartilage regeneration therapy, *J Biomed Mater Res* **52** (2000), 246-55.
[33] Y. S. Choi, S. R. Hong, Y. M. Lee, K.W. Song, M. H. Park and Y. S. Nam, Studies on gelatin-containing artificial skin: Ii. Preparation and characterization of cross-linked gelatin-hyaluronate sponge, *J Biomed Mater Res* 48 (1999), 631-9.
[34] I. V. Yannas and J. F. Burke, Design of an artificial skin. I. Basic design principles, *J Biomed Mater Res* **14** (1980), 65-81.
[35] S. S. Kim, M. Sun Park, O. Jeon, C. Yong Choi and B. S. Kim, Poly(lactide-co-glycolide)/hydroxyapatite composite scaffolds for bone tissue engineering, *Biomaterials* **27** (2006), 1399-409.
[36] K. Cui, Y. Zhu, X. H. Wang, Q. L. Feng and F. Z. Cui, A porous scaffold from bone-like powder loaded in a collagen-chitosan matrix, *J. Bioactive and Compatible Polymers* **19** (2004), 17-31.
[37] E. Farrell, F. J. O'Brien, P. Doyle, J. Fischer, I. Yannas, B. A. Harley, B. O'Connell, P. J. Prendergast and V. A. Campbell, A collagen-glycosaminoglycan scaffold supports adult rat mesenchymal stem cell differentiation along osteogenic and chondrogenic routes, *Tissue Eng* **12** (2006), 459-68.
[38] M. Murata, B. Z. Huang, T. Shibata, S. Imai, N. Nagai and M. Arisue, Bone augmentation by recombinant human bmp-2 and collagen on adult rat parietal bone, *Int J Oral Maxillofac Surg* **28** (1999), 232-7.
[39] M. G. Haugh, Jaasma, M.J. and O'Brien, F.J., Effects of dehydrothermal crosslinking on mechanical and structural properties of collagen-gag scaffolds, *J. Biomed. Mater. Res.: Part A* **89** (2009), 363-369
[40] C. V. Rodrigues, P. Serricella, A. B. Linhares, R. M. Guerdes, R. Borojevic, M. A. Rossi, M. E. Duarte and M. Farina, Characterization of a bovine collagen-hydroxyapatite composite scaffold for bone tissue engineering, *Biomaterials* **24** (2003), 4987-97.
[41] A.A. Al-Munajjed, J.P. Gleeson and F. J. O'Brien, Development of a collagen calcium- phosphate scaffold as a novel bone graft substitute, *Stud Health Technol Inform* **133** (2008) 11-20.
[42] S. S. Liao, F. Z. Cui, W. Zhang and Q. L. Feng, Hierarchically biomimetic bone scaffold materials: Nano-ha/collagen/pla composite, *J Biomed Mater Res B Appl Biomater* **69** (2004), 158-65.
[43] Lysaght M.J., Jaklenec A. and D. E., Great expectations: Private sector activity in tissue engineering, regenerative medicine, and stem cell therapeutics, *Tissue Eng* **14** (2008), 305-315.
[44] I. V. Yannas, E. Lee, D. P. Orgill, E. M. Skrabut and G. F. Murphy, Synthesis and characterization of a model extracellular matrix that induces partial regeneration of adult mammalian skin, *Proc Natl Acad Sci U S A* **86** (1989), 933-7.
[45] J. P. Vacanti, Editorial: Tissue engineering: A 20-year personal perspective, *Tissue Eng* **13** (2007), 231-2.

[46] C. A. Pangborn and K. A. Athanasiou, Growth factors and fibrochondrocytes in scaffolds, *J Orthop Res* **23** (2005), 1184-90.
[47] R. Langer, Tissue engineering: Perspectives, challenges, and future directions, *Tissue Eng* **13** (2007), 1-2.
[48] M. V. Risbud and M. Sittinger, Tissue engineering: Advances in in vitro cartilage generation, *Trends Biotechnol* **20** (2002), 351-6.
[49] L. Klouda and A. G. Mikos, Thermoresponsive hydrogels in biomedical applications, *Eur J Pharm Biopharm* **68** (2008), 34-45.
[50] K. T. Nguyen and J. L.West, Photopolymerizable hydrogels for tissue engineering applications, *Biomaterials* **23** (2002), 4307-14.
[51] M. J. Mondrinos, R. Dembzynski, L. Lu, V. K. Byrapogu, D. M. Wootton, P. I. Lelkes and J. Zhou, Porogen-based solid freeform fabrication of polycaprolactone-calcium phosphate scaffolds for tissue engineering, *Biomaterials* **27** (2006), 4399-408

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-202

IV.2. Cellular & Molecular Biomechanics

Veronica A. CAMPBELL[a,b] and Brian O'CONNELL[a,c]
[a] *Trinity Centre for Bioengineering, Trinity College, Dublin 2, Ireland*
[b] *Department of Physiology, School of Medicine, Trinity College, Dublin 2, Ireland*
[c] *Dublin Dental School, Trinity College, Dublin 2, Ireland*

Introduction

Cells exist in a mechanical environment and must be able to elicit an appropriate response to strains evoked by fluid flow, compression, pressure and stretch. The ability to detect a range of forces (from 10^{-4} through to 10^{4} N m^{-2}, reflecting a faint sound and aortic pressure, respectively) is vital in order to elicit the appropriate cascade of molecular events that facilitate the physiological processing of sensory information. The detection of mechanical strain is mediated by mechano-sensitive components of cells and is a property of a wide range of tissue types. Mechanosensors in the cells of the skin and ear are critical for processing information about touch and hearing, while other mechanosensors respond to blood pressure (baroreceptors), muscle stretch (spindle receptors) and limb positions (proprioceptors). Bone tissue detects mechanical stress, and this informs the bone remodeling process, while endothelial cells lining blood vessels respond to shear stress evoked by fluid flow to initiate vascular remodeling. The ability to respond to mechanical stimuli also regulates fundamental cellular events such as proliferation, cell spreading, differentiation, motility and the maintenance of cell shape. The precise cellular mechanisms underlying mechanosensation remain to be fully resolved, but in recent years a number of surface proteins have been identified that serve a role as mechanoreceptors. Biological systems have evolved a complex array of mechanoreceptors to transduce force into an electrical and/or intracellular biochemical cascade. Such signaling pathways are intricately linked with cellular responsiveness and adaptation to mechanical stimulation.

1. Biomechanical Signal Transduction

1.1. Mechanosensitive (MS) Ion Channels

Mechanosensitive (MS) ion channels form pores that span the plasma membrane and are therefore ideally situated to respond to external forces applied to the cell. These channels serve as mechano-electrical switches converting a mechanical force into a change in the cellular electrochemical gradient (see Figure 1), and they play a role in coordinating diverse physiological events such as touch and hearing in mammals, to adaptation to turgor in bacteria [1, 2]. The patch clamp technique has facilitated the electrophysiological analysis of single MS channel currents and has demonstrated that these channels open in response to a hypo-osmotic environment which causes the

membrane to swell, or by the direct application of suction [3]. It has been suggested that MS channels may be gated directly by changes in the lipid bilayer tension, or in the 'tethered model' of gating, by displacement of the MS channel with respect to components of the cytoskeleton or extracellular matrix [1]. Amongst the mechanosensitive channels are the transient receptor potential channels (TRPs) [4] and stretch-activated ion channels (SACs) [5]. However, mechanosensitive channels also include other ion channels such as voltage-dependent Ca^{2+} channels [6], the NMDA receptor [7] and Ca^{2+}-dependent BK channels [8]. Many MS channels are permeable to cations in general, with highest permeability to Ca^{2+}, which results in an increase in intracellular Ca^{2+} concentration following channel opening. It is postulated that the Ca^{2+} then acts as a second messenger resulting in the appropriate cell response to mechanical stimulation [9]. Other classes of MS ion channels exhibit selective permeability for Na^{2+}, K^{+} or Cl^{-} ions. It is therefore apparent that changes in the mechanical environment initiate electrical and chemical changes in the cell which define the cellular response.

MS channels are also modulated by cellular components, such as the contractile actin cytoskeleton. The cytoskeleton provides support to the plasma membrane and forms part of the coupling mechanism to the extracellular matrix. It reacts and rearranges its configuration in response to alterations in the extracellular environment. The tensegrity model [2] proposes that the cytokeleton is the architectural basis for cellular mechanotransduction, and application of physical stress promotes changes in actin polymerization to control cell shape and realignment [10]. A direct coupling may exist between MS channels and the cytoskeleton since agents that disrupt the organization of actin filaments, such as colchicine and cytochalasin-D, affect MS channel activity. Studies suggest that the cytoskeleton exerts a tonic inhibition on MS channel activity and support the contention that MS ion channels are functionally linked with the dynamics of the intracellular cytoskeleton [11, 12].

1.2. Transient Receptor Potential Channels

Members of the transient receptor potential (TRP) superfamily of membrane-associated proteins are recognized to have role in mechanosensation. These ion channels are Ca^{2+}-permeable and lead to an increase in intracellular Ca^{2+} concentration upon their activation, either by facilitating Ca^{2+} influx, or by promoting the release of Ca^{2+} from intracellular stores [13]. TRP channels are widely expressed in the CNS and peripheral cell types and exhibit substantial evolutionary conservation. The TRP channels are critically involved in sensory function having an essential role in vision, hearing, taste, olfaction and mechnosensation. A number of possibilities for mechanical activation of TRP channels exist [4]. The channels may be activated directly by mechanical stimulation via forces conveyed through lipid tension or through structural proteins. Alternatively, these channels may be activated indirectly via another mechano-sensitive protein that is distinct from the TRP channel, or may be regulated by a diffusible second messager, such as arachidonic acid, generated by a mechano-sensitive protein linked to phospholipase A_2. The TRP channels are classified in to seven subgroups; TRPC, TRPM, TRPV, TRPN, TRPA, TRPP and TRPML. TRPV1, the vanilloid receptor, is responsible for sensing heat [14], yet is also proposed to have a role in mechanosensation associated with bladder filling [4]. TRPV4 is activated in response to cell swelling via an indirect mechanism involving second messenger pathways, such as arachidonic acid [15], whereas the TRPC1 channel is gated directly by the tension

that is generated in the plasma membrane [16]. In the nematode, C. elegans, the mechanosensitive TRPN channel has recently been found to be involved in stretch receptor-mediated proprioception to regulate sensory-motor integration during locomotion [17]. Thus, the TRP family of receptors are mechanically regulated via distinct mechanisms, and are involved in sensing changes in lipid tension precipitated by stretch and a hypotonic environment.

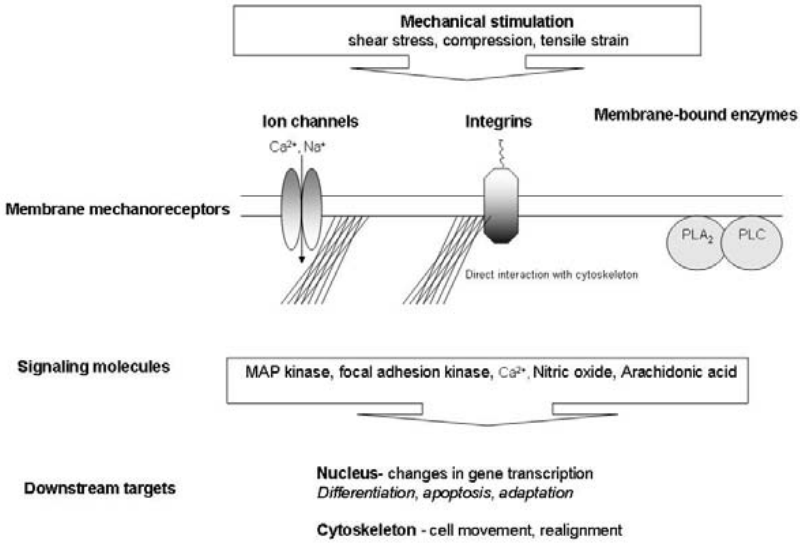


Figure 1. Cell pathways associated with mechanotransduction**.** Cells respond to mechanical stimulation via activation of membrane-associated mechanoreceptor proteins which undergo a conformation change resulting in the activation and recruitment of a cascade of intacellular signaling molecules. The signaling molecules may be released into the extracellular matrix (e.g nitric oxide) to influence neighbouring cells, or may target organelles such as the nucleus and cytoskeleton to evoke changes in gene transcription or cell motility, respectively.

1.3. Membrane –Associated Enzymes

Enzymes associated with the intracellular domain of the plasma membrane, such as phospholipase A_2, phospholipase C and tyrosine kinases have also been reported to have a mechanosensitive role [18]. Membrane stretch increases the production of prostaglandins via phospholipase A_2-mediated release of arachidonic acid [19]. In a number of cell types, flow and shear have been shown to induce activation of phospholipase C with subsequent activation of the phosphatidylinositol pathway. In glial cells exposured to pressure, cell proliferation is increased and this effect is blocked by genistein, the tyrosine kinase inhibitor, but not by blockers of mechanosensitive channel blockers, implicating tyrosine kinases as mechanosensitive elements in some cell types [20].

1.4. Intracellular Enzymes and Diffusible Messengers

Mechanical force has been shown to activate the mitogen-activated protein kinase (MAPK) superfamily [21]. The extracellular-regulated protein kinase (ERK) has been definitively implicated in mechanotransduction, but a role for p38 and c-jun N-terminal kinase (JNK) has also been reported [22]. ERK activation has been observed in response to steady and pulsatile stretch in the vasculature [23], renal mesangial cells [24], osteoblasts [25] and gingival fibroblasts [26] where downstream consequences of ERK activation include proliferation, immediate early gene expression and extracellular matrix accumulation. Furthermore, the strain-induced activation of ERK is thought to play a role in 'mechanical strain memory' in airway smooth muscle, whereby a mechanically-induced deformation of the smooth muscle cell produces a structural change which modulates mechanical performance in subsequent contractions [27].

Mechanical strain activates nitric oxide synthase in endothelial cells [28], stromal cells [29], chondrocytes [30], and osteocytes [31] leading to the formation of nitric oxide (NO), a gaseous diffusible second messenger that regulates vascular remodeling, increases expression of bone-related genes, regulates protoeoglycan synthesis, and alters cytoskeletal fibre alignment, respectively. The strain-induced changes in NO formation are coordinated by ERK signaling [32], demonstrating substantial crosstalk between these intracellular pathways.

1.5. Integrins and Cadherins

Integrins are transmembrane proteins that link components of the intracellular cytoskeleton to the extracellular matrix. These cell-matrix adhesion proteins can transmit information about membrane deformation, shear across the membrane, and pressure transients to the cytoskeleton and associated signal transduction pathways. Integrins faciltate the rearrangement of the cytoskeleton, and its associated organelles and nuclei, in order that the cell may be strengthened against mechanical distortion. Integrins also play a role in the functional adaptation of cells to matrix rigidity where the αvß3 integrin has been demonstrated to activate a ridigity response to regulate cell spreading [33]. Furthermore, ß1/3 integrins are upregulated in ventricular myocytes exposed to stretch, and the downstream signaling pathways, involving src and focal adhesion kinase (FAK), results in secretion of vascular endothelial growth factors which stimulates the induction of other gap junction proteins [34]. This pathway influences the pattern of cardiac cell-cell communication which is critical for optimal contractility of cardiac tissue. An integrin-dependent mechanotransduction pathway involving src signaling and induction of osteogenic proteins has also been reported in bone cells exposed to mechanical force, providing further evidence that mechanical force influences osseous regeneration [35]. Thus, mechnical activation of integrin receptors can regulate dynamic events such as reorganization of the cytoskeleton [33, 36] or the expression of genes encoding proteins associated with cell adhesion and differentiation [34, 37].

Cell-cell adhesion sites composed of cadherins also provide a mechanism for mechanotransduction. Cadherins mediate a force-induced increase in Ca^{2+} influx through fibroblast mechanosensitive channels [19] and can directly infuence gene transcription since components of the cadherin junctional complex, have deen demonstrated to translocate to the nucleus in osteoblasts [38].

2. Biomechanics of Differentiation and Maintenance of Tissue Phenotype

2.1. Stem Cells

A number of lines of evidence support the hypothesis that stem cells are responsive to mechanical stimulation, and that stem cell differentiation may be regulated by strain. Certainly during development the embryo is exposed to stresses imposed by the external environment of the uterus, as well as to stresses imposed by neighbouring cells. These growth-generated strains and pressures can regulate expression of a number of developmental genes [39]. Administration of a neuromuscular blocker to a chick embryo causes paralysis, and interferes with the development of the musculoskeletal system [40], thereby reflecting the importance of the local mechanical environment to tissue differentiation. At a cellular level, there is substantial evidence demonstrating that mechanical stresses exert a direct effect on stem cells. For example, mouse embryonic stem cells generate reactive oxygen species in response to mechanical strain and this pathway initiates differentiation into cells of the cardiovascular system [41]. Adult mesenchymal stem cells (MSCs) isolated from the marrow are multipotent cells that have the proclivity to differentiate along osteogenic, chondrogenic and adipogenic lineages [42]. Human MSCs exposed to cyclic uniaxial strain in the presence of osteoinductive factors increase the formation of bone-associated extracellular matrix [43], whilst exposure to low-intensity ultrasound promotes chondrogenesis [44]. MSCs isolated from adipose tissue respond to mechanical loading applied via pulsating fluid flow by increasing nitric oxide production and undergoing osteogenesis [45]. As well as controlling stem cell differentiation potential, mechanical signals are also important for regulating MSC proliferation via Ca^{2+} -dependent mechanisms [46]. Thus mechanical stimulation regimes may have potential to modulate stem cell proliferation and differentiation profiles for tissue engineering applications (Figure 2).

2.2. Vascular System

The walls of blood vessels react to the mechanical stimuli of pressure and shear stress evoked by the flowing blood. The responses to such mechanical events are necessary to control blood flow, or to adapt the vessel structure to its requirements. Shear stress is a critical biophysical phenomenon for determining vascular homeostasis, vascular remodeling and development of the cardiac system, as well as for governing the development of atherosclerotic lesions. The endothelial cells which line blood vessels are responsible for vascular responses to the shear stress that is evoked by fluid flow. Shear stress is transmitted from the apical surface of the endothelial cell through the cytoskeleton to points of attachments at cell-cell and cell-matrix junctions. These junctions, composed of proteins, such as platelet endothelial cell adhesion molecule (PECAM), transmit mechanical force via interaction with the intracellular signaling proteins, Src and P13kinase, leading to activation of the transcription factor, NFκB [47]. In laminar shear, this integrin pathway promotes endothelial cell alignment in the direction of flow, and a transient activation of the NFκB pathway. In contrast, in the face of disturbed shear, such as may occur at vessel branch points, bifurcations and regions of high curvature, the PECAM pathway is activated in a sustained manner. The sustained activation of NFκB leads to prolonged induction of NFκB-regulated genes, (e.g adhesion molecules, ICAM, VCAM, E-selectin) which are molecular hallmarks of

atherosclerosis-prone sites. The mechanosensory pathway involving the PECAM complex is therefore described as one of the earliest-known events in the development of atherosclerotic lesions. Thus, while this mechanosensory complex is important for physiological adaptation to fluid flow, emerging evidence suggests that overactivation of this pathway can lead to cardiovascular disease.

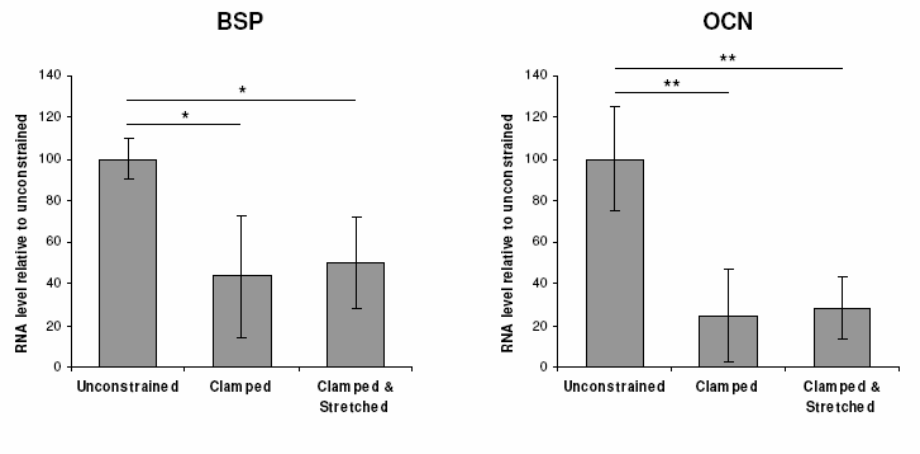


Figure 2. Bone-associated gene expression in stem cells is modulated by constraining the substrate or the application of uniaxial strain.Effects of one-dimensional clamping (Clamped) and 5% uniaxial cyclic strain (Clamped & Stretched) (1 Hz) on levels of bone-associated RNAs from mesenchymal stem cell-seeded collagen glycosaminoglycans scaffold constructs incubated for 7 days in the presence of osteogenic factors, when compared to unconstrained constructs (Unconstrained). BSP; bone sialoprotein, OCN; osteocalcin. *p<0.05, **p<0.01. Data courtesy of Dr. Elaine Byrne, RCSI

2.3. Skeletal System

The skeletal system can regulate bone density and architecture in order to meet the body's strucural and metabolic demands. While the metabolic demands of the skeleton are controlled by the calcitropic hormones, the responsiveness of the skeleton to mechanical load governs skeletal strucure. The cells of the skeleton – osteoblasts, osteocytes, osteoclasts, chondrocytes, as well as their progenitor bone marrow-derived cells – display mechnosensitivity which influences how these cells function, and thus the skeletal characteristics. At a macro-level it is clear that application of load is anabolic to bone [48], while un-weighting, such as occurs in the absence of gravity, induces bone loss [49]. At the micro-level a number of cells involved in the maintenance of bone may be regulated by mechnical force. The marrow stromal progenitor cells display alterations in proliferation, gene expression, and nitric oxide formation in response to mechnical load [32], and RANKL expression, a key factor in osteoclast formation, is increased in response to fluid shear stress [50].

Osteocytes are the principal cell in cortical bone and by extending cellular processes through a network of fluid-filled cannaliculi, they are ideally placed to monitor mechanical strain as it is transmitted to the cell surface. Application of mechanical load increases the production of bone-related proteins [39]. In relation to

bone mechanoreceptors, stretch-activated ion channels have been identified in bone cells [51] and L-type Ca^{2+} channel have a mechanosensitive role [52]. Mechanical stimuli across the cell membrane can be transmitted to the cytoskeleton via integrin receptors [53]. Osteoblasts, osteocytes and chondrocytes express integrin receptors, and fluid flow upregulates ß1 integrin expression and induces downsteam intracellular signaling pathways such as src [54, 55]. An additional signaling molecule involved in bone cell responsiveness to mechanical stimulation is focal adhesion kinase (FAK) which, upon phosphorylation, activates the mitogen-activated protein kinase (MAPK), one of the effectors of oscillatory flow in bone cells [56]. Activated MAPK can translocate to the nucleus to regulate gene transcription, and this represents a mechanism for long-term adaptation to mechanical stimulation since it involves *de novo* protein synthesis. The responsiveness of bone cells to mechanical strain, and thus bone remodeling, is likely to be affected by disease. In support of this, bone cells isolated from osteoarthritic donors displayed a reduced metabolic response to fluid shear stress compared to cells isolated from a patient with osteoporosis (see Figure 3) [57].

Cartilage homeostasis is also regulated by mechanical stress during embryonic limb development, as well as during fracture repair and skeletal remodeling in the adult [58]. Chondrocytes respond to mechanical strain that is evoked by deformation of the extracellular matrix, as well as to compressive and shear flow forces. Chondrocytes maintained in a 3-dimensional *in vitro* environment to mimic the cartilage matrix environment, and exposed to cyclic matrix deformation, exhibit changes in proliferation rate and synthesis of cartilage-specific proteins [59]. This stretch-induced change in chondrocyte proliferation and differentiation was affected differentially by the Ca^{2+} channel blocker, nifedipine, and the stretch-activated channel blocker, gadolinium, suggesting that stretch-induced matrix deformation regulates chondrocyte proliferation and differentiation via two distinct pathways. Integrin signaling has a role in the response of chondrocytes to dynamic compression [60], and chondrocytes are also responsive to shear stress, since flow-induced shear stress increases type II collagen deposition and tensile mechanical properties [61].

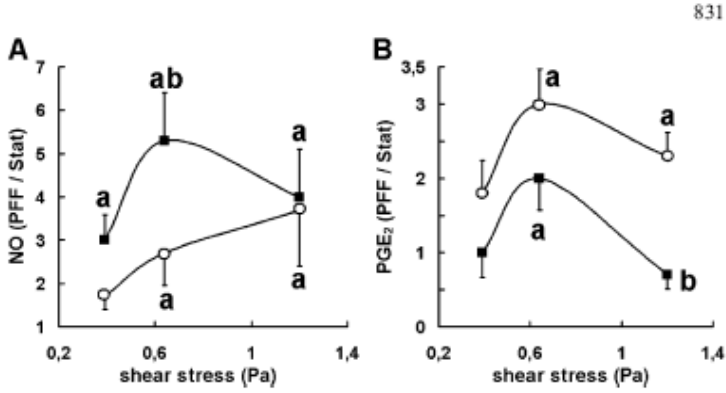


Figure 3. Comparison of osteoporotic and osteoarthritic cell responses to shear fluid flow Effect of pulsating fluid flow (PFF) at low (0.4±0.1 Pa), medium (0.6± 0.3 Pa), or high shear stress (1.2±0.4 Pa) on nitric oxide (NO) and prostaglandin E2 (PGE2) production by osteoporotic (OP, black squares) and osteoarthritic (OA, white circles) bone cells. Data are expressed as PFF-treated over static culture ratios calculated per individual donor. A; NO response, B; PGE2 response. PFF/Stat=1 means no effect. a implies a significant effect of PFF, b implies significantly different from OA cells, P<0.05. *Reproduced from [57].*

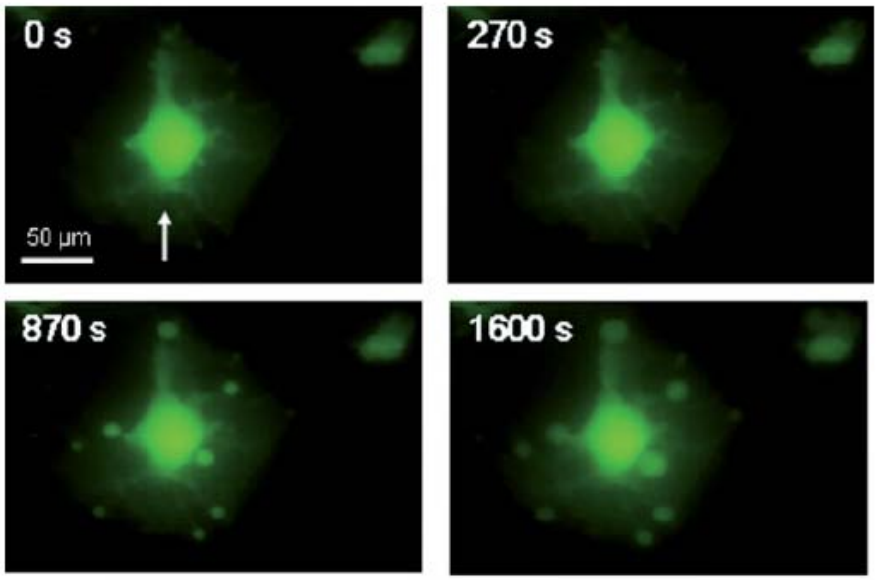


Figure 4. Generation of nitric oxide (NO) in a single MC-3T3 osteoblast cell in response to stimulation by atomic force microscopy. Time-series images of NO fluorescence intensity from 0 s to 1600 s. Periodic indentation with peak force of ~22nN was applied from 30 s until 60 s at the location indicated (white arrow*). Reproduced from [65].*

3. Experimental Models of Cellular Biomechanics

In order to understand the various mechanical transduction pathways mentioned above, various models have been developed at the molecular, cellular, organ and animal levels [62]. It is often helpful to use a number of complementary approaches to elucidate both what a molecule *can* do as well at what it *does* do [63]. In general, the study of isolated molecules provides precise information regarding their properties but may not provide the context or real life effect of their activation. On the other hand, whole animal models can give valuable clues to how cellular biomechanics plays out, but they are subject to a high level of background noise and sometimes difficult to interpret.

The mechanics of some molecules, such as actin and helicase, have been studied extensively in isolation so that their biological behaviour is fairly well understood. However, biomechanical responses involve many other molcules that are better examined in conditions that more closely resemble their native environment. For example, patch clamping is able to measure currents in membrane channels that are maintained in a lipid bilayer. Delevopments in atomic force microscopy mean that the mechanical properties of individual molecules can be explored on a cell's surface and the response of a living cell observed in real time. For example, we have characterized the geodesic F-actin structures that appear in rat mesenchymal stem cells [64] and have demonstrated that mechanical stimulation of the cytoskeleton in a single bone cell can lead to the mobilization of nitric oxide (Figure 4) [65].

For practical reasons, most biomechanical responses have been measured in cultured cell populations, rather than individual cells. This is because a cell response

(second messenger production, ion flux, protein phosphorylation) is often more readily and reliably measured from a large number of cells. Similarly, it may be more feasible to expose a population of cells to a particular stimulus (stretch, shear, pulsating flow) than a single cell. However, there are drawbacks with this approach, including the need to ensure a homogeneous cell population and the assumption that all cells in the experiment will experience precisely the same strain. Nonetheless, the observation of biomechanical responses in cells has been very useful in undertanding how cells are likely to behave in a variety of conditions, for example in coronary heart disease, bone fractures, during skeletal development or orthodontic tooth movement. An extension of the purely cell-based experiment is to use an entire tissue or organ in culture to study biomechanical responses. Isolated muscle, bone, cartilage, tendon or blood vessel may be maintained long enough *in vitro* to study cellular responses to applied strain over a period of hours to perhaps a few days. Improved systems for maintaining tissue viability, 'bioreactors', are discussed elsewhere in this text and they have been used, for example, to study the response of bone to the mechanical stimulation of artificial implants [66], the apoptosis of vertebral disc cartilage cells caused by compression [67] and the stimulation of growth at cranial sutures [68].

Biomechanical studies may be extended further, to the whole animal, to observe a wide variety of responses, particularly those that depend on the function of entire systems. For instance, rat hindlimb suspension has been used to compare the cellular (and matrix) response of bone to the loss of mechanical loading as it would be difficult to model this effect in an *in vitro* system [69]. Similarly, cellular changes have been studied resulting from shear stresses in arterial remodelling [70] and gene expression caused by mechanical stresses generated following lung injury [71]. Such data may clarify transduction pathways or even inform clinical decisions [72]. It is important to note that studies of cellular biomechanics do not necessarily follow the paradigm of application of stress followed by an observation of altered cell function. The advent of transgenic and knockout animal models means that key molecules in biomechanical pathways can be modified or ablated and any change in function revealed [73]. In this way, valuable new insights into cell function and disease may be gained.

4. Cellular & Molecular Biomechanics: State of the Art and Future Directions

The mechanical environment in which cells exist dictates the physiological attributes of the cell, and this in turn regulates the function and maintenance of the tissue in which the cell resides. It is clear that a number of cellular elements serve roles as mechanoreceptors, and that the intracellular signaling pathways that become activated by mechanical strain are diverse. Understanding exactly how cells respond to strain is an emerging area of research, and the knowledge gained from investigations in this area will be relevant to the tissue engineering field as we strive to develop paradigms with which to control cellular differentiation and the maintenance of tissue phenotype.

References

[1] B. Martinac, Mechanosensitive ion channels: molecules of mechanotransduction, *J Cell Sci.* **117** (2004), 2449-60.

[2] D.E. Ingber, Cellular mechanotransduction: putting all the pieces together again, *Faseb J.* **20** (2006), 811-27.
[3] B. Martinac and A. Kloda, Evolutionary origins of mechanosensitive ion channels, *Prog Biophys Mol Biol*. **82** (2003), 11-24.
[4] S.Y. Lin and D.P. Corey, TRP channels in mechanosensation, *Curr Opin Neurobiol*. **15** (2005), 350-7.
[5] F.B. Kalapesi, J.C. Tan, and M.T. Coroneo, Stretch-activated channels: a mini-review. Are stretch activated channels an ocular barometer? *Clin Experiment Ophthalmol.* **33** (2005), 210-7.
[6] A.J. el Haj, L.M. Walker, M.R. Preston, and S.J. Publicover, Mechanotransduction pathways in bone: calcium fluxes and the role of voltage-operated calcium channels, *Med Biol Eng Comput.* **37** (1999), 403-9.
[7] M. Casado and P. Ascher, Opposite modulation of NMDA receptors by lysophospholipids and arachidonic acid: common features with mechanosensitivity, *J Physiol.* **513** (1998), 317-30.
[8] T. Kawakubo, K. Naruse, T. Matsubara, N. Hotta, and M. Sokabe, Characterization of a newly found stretch-activated KCa,ATP channel in cultured chick ventricular myocytes, *Am J Physiol.* **276** (1999), H1827-38.
[9] J. Lammerding, R.D. Kamm, and R.T. Lee, Mechanotransduction in cardiac myocytes, *Ann N Y Acad Sci*. **1015** (2004), 53-70.
[10] M.J. Cipolla, N.I. Gokina, and G. Osol, Pressure-induced actin polymerization in vascular smooth muscle as a mechanism underlying myogenic behavior, *Faseb J.* **16** (2002), 72-6.
[11] F. Guharay and F. Sachs, Stretch-activated single ion channel currents in tissue-cultured embryonic chick skeletal muscle, *J Physiol.* **352** (1984), 685-701.
[12] X.Wan, P. Juranka, and C.E. Morris, Activation of mechanosensitive currents in traumatized membrane, *Am J Physiol*. **276** (1999), C318-27.
[13] S.F. Pedersen, G. Owsianik, and B. Nilius, TRP channels: an overview, *Cell Calcium*. **38** (2005), 233-52.
[14] M.J. Caterina, T.A. Rosen, M. Tominaga, A.J. Brake, and D. Julius, A capsaicin-receptor homologue with a high threshold for noxious heat, *Nature* **398** (1999), 436-41.
[15] H. Watanabe, J. Vriens, J. Prenen, G. Droogmans, T. Voets, and B. Nilius, Anandamide and arachidonic acid use epoxyeicosatrienoic acids to activate TRPV4 channels, *Nature* **424** (2003), 434-8.
[16] R. Maroto, A. Raso, T.G. Wood, A. Kurosky, B. Martinac, and O.P. Hamill, TRPC1 forms the stretch activated cation channel in vertebrate cells, *Nat Cell Biol*. **7** (2005), 179-85.
[17] W. Li, Z. Feng, P.W. Sternberg, and X.Z. Xu, A C. elegans stretch receptor neuron revealed by a mechanosensitive TRP channel homologue, *Nature* **440** (2006), 684-7.
[18] G. Apodaca, Modulation of membrane traffic by mechanical stimuli, *Am J Physiol Renal Physiol.* **282** (2002), F179-90.
[19] K.S. Ko and C.A. McCulloch, Partners in protection: interdependence of cytoskeleton and plasma membrane in adaptations to applied forces, *J Membr Biol*. **174** (2000), 85-95.
[20] Y. Oishi, Y. Uezono, N. Yanagihara, F. Izumi, T. Nakamura, and K. Suzuki, Transmural compression induced proliferation and DNA synthesis through activation of a tyrosine kinase pathway in rat astrocytoma RCR-1 cells, *Brain Res*. **781** (1998), 159-66.
[21] L.C. Martineau and P.F. Gardiner, Insight into skeletal muscle mechanotransduction: MAPK activation is quantitatively related to tension, *J Appl Physiol.* **91** (2001), 693-702.
[22] C. Yan, M. Takahashi, M. Okuda, J.D. Lee, and B.C. Berk, Fluid shear stress stimulates big mitogen activated protein kinase 1 (BMK1) activity in endothelial cells. Dependence on tyrosine kinases and intracellular calcium, *J Biol Chem*. **274** (1999), 143-50.
[23] S. Lehoux, B. Esposito, R. Merval, and A. Tedgui, Differential regulation of vascular focal adhesion kinase by steady stretch and pulsatility, *Circulation*. **111** (2005), 643-9.
[24] J.C. Krepinsky, Y. Li, D. Tang, L. Liu, J. Scholey, and A.J. Ingram, Stretch-induced Raf-1 activation in mesangial cells requires actin cytoskeletal integrity, *Cell Signal.* **17** (2005), 311-20.
[25] J.P. Hatton, M. Pooran, C.F. Li, C. Luzzio, and M. Hughes-Fulford, A short pulse of mechanical force induces gene expression and growth in MC3T3-E1 osteoblasts *via* an ERK 1/2 pathway, *J Bone Miner Res*. **18** (2003), 58-66.
[26] T.E. Danciu, E. Gagari, R.M. Adam, P.D. Damoulis, and M.R. Freeman, Mechanical strain delivers anti-apoptotic and proliferative signals to gingival fibroblasts, *J Dent Res*. **83** (2004), 596-601.
[27] H.R. Kim and C.M. Hai, Mechanisms of mechanical strain memory in airway smooth muscle, *Can J Physiol Pharmacol*. **8** (2005), 811-5.
[28] R.J. Hendrickson, C. Cappadona, E.N. Yankah, J.V. Sitzmann, P.A. Cahill, and E.M. Redmond, Sustained pulsatile flow regulates endothelial nitric oxide synthase and cyclooxygenase expression in cocultured vascular endothelial and smooth muscle cells, *J Mol Cell Cardiol.* **31** (1999), 619-29.

[29] X. Fan, J.A. Rahnert, T.C. Murphy, M.S. Nanes, E.M. Greenfield, and J. Rubin, Response to mechanical strain in an immortalized pre-osteoblast cell is dependent on ERK1/2, *J Cell Physiol*. **207** (2006), 454- 60.
[30] M. Matsukawa, K. Fukuda, K. Yamasaki, K. Yoshida, H. Munakata, and C. Hamanishi, Enhancement of nitric oxide and proteoglycan synthesis due to cyclic tensile strain loaded on chondrocytes attached to fibronectin, *Inflamm Res.* **53** (2004), 239-44.
[31] J.G. McGarry, J. Klein-Nulend, and P.J. Prendergast, The effect of cytoskeletal disruption on pulsatile fluid flow-induced nitric oxide and prostaglandin E_2 release in osteocytes and osteoblasts, *Biochem Biophys Res Commun.* **330** (2005), 341-8.
[32] J. Rubin, T.C. Murphy, L. Zhu, E. Roy, M.S. Nanes, and X. Fan, Mechanical strain differentially regulates endothelial nitric-oxide synthase and receptor activator of nuclear kappa B ligand expression via ERK1/2 MAPK, *J Biol Chem*. **278** (2003), 34018-25.
[33] G. Jiang, A.H. Huang, Y. Cai, M. Tanase, and M.P. Sheetz, Rigidity sensing at the leading edge through alphavbeta3 integrins and RPTP_, *Biophys J*. **90** (2006), 1804-9.
[34] K. Yamada, K.G. Green, A.M. Samarel, and J.E. Saffitz, Distinct pathways regulate expression of cardiac electrical and mechanical junction proteins in response to stretch, *Circ Res.* **97** (2005), 346-53.
[35] S.T. Rhee and S.R. Buchman, Colocalization of c-Src (pp60src) and bone morphogenetic protein 2/4 expression during mandibular distraction osteogenesis: in vivo evidence of their role within an integrinmediated mechanotransduction pathway, *Ann Plast Surg.* **55** (2005), 207-15.
[36] D. Riveline, E. Zamir, N.Q. Balaban, U.S. Schwarz, T. Ishizaki, S. Narumiya, Z. Kam, B. Geiger, and A.D. Bershadsky, Focal contacts as mechanosensors: externally applied local mechanical force induces growth of focal contacts by an mDia1 dependent and ROCK-independent mechanism, *J Cell Biol.* **153** (2001), 1175-86.
[37] M.E. Chicurel, R.H. Singer, C.J. Meyer, and D.E. Ingber, Integrin binding and mechanical tension induce movement of mRNA and ribosomes to focal adhesions, *Nature* **392** (1998), 730-3.
[38] S.M. Norvell, M. Alvarez, J.P. Bidwell, and F.M. Pavalko, Fluid shear stress induces beta-catenin signaling in osteoblasts, *Calcif Tissue Int*. **75** (2004), 396-404.
[39] H. Henderson and D.R. Carter, Mechanical induction in limb morphogenesis: the role of growth generated strains and pressures, *Bone.* **31** (2002), 645-53.
[40] B.K. Hall and S.W. Herring, Paralysis and growth of the musculoskeletal system in the embryonic chick*, J Morphol*. **206** (1990), 45-56.
[41] M. Schmelter, B. Ateghang, S. Helmig, M. Wartenberg, and H. Sauer, Embryonic stem cells utilize reactive oxygen species as transducers of mechanical strain-induced cardiovascular differentiation, *Faseb J*. **20** (2006), 1182-4.
[42] M.F. Pittenger, A.M. Mackay, S.C. Beck, R.K. Jaiswal, R. Douglas, J.D. Mosca, M.A. Moorman, D.W. Simonetti, S. Craig, and D.R. Marshak, Multilineage potential of adult human mesenchymal stem cells*, Science* **284** (1999), 143-7.
[43] A. Wiesmann, H.J. Buhring, C. Mentrup, and H.P. Wiesmann, Decreased CD90 expression in human mesenchymal stem cells by applying mechanical stimulation, *Head Face Med.* **2** (2006), 8.
[44] J.H. Cui, K. Park, S.R. Park, and B.H. Min, Effects of low-intensity ultrasound on chondrogenic differentiation of mesenchymal stem cells embedded in polyglycolic acid: an in vivo study, *Tissue Eng*. **12** (2006), 75-82.
[45] M. Knippenberg, M.N. Helder, B.Z. Doulabi, C.M. Semeins, P.I. Wuisman, and J. Klein-Nulend, Adipose tissue-derived mesenchymal stem cells acquire bone cell-like responsiveness to fluid shear stress on osteogenic stimulation, *Tissue Eng.* **11** (2005), 1780-8.
[46] R.C. Riddle, A.F. Taylor, D.C. Genetos, and H.J. Donahue, MAP kinase and calcium signaling mediate fluid flow-induced human mesenchymal stem cell proliferation, *Am J Physiol Cell Physiol.* **290** (2006), C776-84.
[47] E. Tzima, M. Irani-Tehrani,W.B. Kiosses, E. Dejana, D.A. Schultz, B. Engelhardt, G. Cao, H. DeLisser, and M.A. Schwartz, A mechanosensory complex that mediates the endothelial cell response to fluid shear stress, *Nature* **437** (2005), 426-31.
[48] C.T. Rubin and L.E. Lanyon, Regulation of bone mass by mechanical strain magnitude, *Calcif Tissue Int. 37* (1985), 411-7.
[49] G. Carmeliet, L. Vico, and R. Bouillon, Space flight: a challenge for normal bone homeostasis, *Crit Rev Eukaryot Gene Expr*. **11** (2001), 131-44.
[50] M. Mehrotra, M. Saegusa, S.Wadhwa, O. Voznesensky, D. Peterson, and C. Pilbeam, Fluid flow induces Rankl expression in primary murine calvarial osteoblasts, *J Cell Biochem*. **98** (2006), 1271-83.
[51] N. Kizer, X.L. Guo, and K. Hruska, Reconstitution of stretch-activated cation channels by expression of the alpha-subunit of the epithelial sodium channel cloned from osteoblasts, *Proc Natl Acad Sci U S A.* **94** (1997), 1013-8.

[52] J. Li, R.L. Duncan, D.B. Burr, and C.H. Turner, L-type calcium channels mediate mechanically induced bone formation *in vivo*, *J Bone Miner Res*. **17** (2002), 1795-800.
[53] J. Rubin, C. Rubin, and C.R. Jacobs, Molecular pathways mediating mechanical signaling in bone, *Gene* **367** (2006), 1-16.
[54] S. Kapur, D.J. Baylink, and K.H. Lau, Fluid flow shear stress stimulates human osteoblast proliferation and differentiation through multiple interacting and competing signal transduction pathways, *Bone* **32** (2003), 241-51.
[55] F.A. Weyts, Y.S. Li, J. van Leeuwen, H. Weinans, and S. Chien, ERK activation and alpha v beta 3 integrin signaling through Shc recruitment in response to mechanical stimulation in human osteoblasts, *J Cell Biochem.* **87** (2002), 85-92.
[56] T. Ishida, T.E. Peterson, N.L. Kovach, and B.C. Berk, MAP kinase activation by flow in endothelial cells. Role of beta 1 integrins and tyrosine kinases, *Circ Res*. **79** (1996), 310-6.
[57] A.D. Bakker, J. Klein-Nulend, E. Tanck, I.C. Heyligers, G.H. Albers, P. Lips, and E.H. Burger, Different responsiveness to mechanical stress of bone cells from osteoporotic versus osteoarthritic donors, *Osteoporos Int.* **17** (2006), 827-33.
[58] Y.J. Kim, R.L. Sah, A.J. Grodzinsky, A.H. Plaas, and J.D. Sandy, Mechanical regulation of cartilage biosynthetic behavior: physical stimuli, *Arch Biochem Biophys*. **311** (1994), 1-12.
[59] Q.Q. Wu and Q. Chen, Mechanoregulation of chondrocyte proliferation, maturation, and hypertrophy: ion-channel dependent transduction of matrix deformation signals, *Exp Cell Res*. **256** (2000), 383-91.
[60] [60] T.T. Chowdhury, R.N. Appleby, D.M. Salter, D.A. Bader, and D.A. Lee, Integrin-mediated mechanotransduction in IL-1 beta stimulated chondrocytes, *Biomech Model Mechanobiol.* **5** (2006), 192-201.
[61] C.V. Gemmiti and R.E. Guldberg, Fluid flow increases type II collagen deposition and tensile mechanical properties in bioreactor-grown tissue-engineered cartilage, *Tissue Eng.* **12** (2006), 469-79.
[62] T.P. Lele, J.E. Sero, B.D. Matthews, S. Kumar, S. Xia, M. Montoya-Zavala, T. Polte, D. Overby, N. Wang, and D.E. Ingber, Tools to study cell mechanics and mechanotransduction, *Methods Cell Biol.* **83** (2007), 443-72.
[63] D. Stamenovic and N. Wang, Invited review: engineering approaches to cytoskeletal mechanics, *J Appl Physiol*. **89** (2000), 2085-90.
[64] [64] P. Maguire, J.I. Kilpatrick, G. Kelly, P.J. Prendergast, V.A. Campbell, B.C. O'Connell, and S.P. Jarvis, Direct mechanical measurement of geodesic structures in rat mesenchymal stem cells, *HFSP Journal* Epub ahead of print (2007).
[65] J.G. McGarry, P. Maguire, V.A. Campbell, B.C. O'Connell, P.J. Prendergast, and S.P. Jarvis, Stimulation of nitric oxide mechanotransduction in single osteoblasts using atomic force microscopy, *J Orthop Res,* **26** (2008), 513-21.
[66] V. Mann, C. Huber, G. Kogianni, D. Jones, and B. Noble, The influence of mechanical stimulation on osteocyte apoptosis and bone viability in human trabecular bone, *J Musculoskelet Neuronal Interact*. **6** (2006), 408-17.
[67] K. Ariga, K. Yonenobu, T. Nakase, N. Hosono, S. Okuda, W. Meng, Y. Tamura, and H. Yoshikawa, Mechanical stress-induced apoptosis of endplate chondrocytes in organ-cultured mouse intervertebral discs: an ex vivo study, *Spine* **28** (2003), 1528-33.
[68] S.S. Tholpady, T.F. Freyman, D. Chachra, and R.C. Ogle, Tensional forces influence gene expression and sutural state of rat calvariae in vitro, *Plast Reconstr Surg*. **120** (2007), 601-11; discussion 612-3.
[69] D.A. Hardiman, F.J. O'Brien, P.J. Prendergast, D.T. Croke, A. Staines, and T.C. Lee, Tracking the changes in unloaded bone: Morphology and gene expression, *Eur J Morphol.* **42** (2005), 208-16.
[70] Y.H. Li, C.Y. Hsieh, D.L. Wang, H.C. Chung, S.L. Liu, T.H. Chao, G.Y. Shi, and H.L. Wu, Remodeling of carotid arteries is associated with increased expression of thrombomodulin in a mouse transverse aortic constriction model, *Thromb Haemost*. **97** (2007), 658-64.
[71] B.A. Simon, R.B. Easley, D.N. Grigoryev, S.F. Ma, S.Q. Ye, T. Lavoie, R.M. Tuder, and J.G. Garcia, Microarray analysis of regional cellular responses to local mechanical stress in acute lung injury, *Am J Physiol Lung Cell Mol Physiol*. **291** (2006), L851-61.
[72] M.C. Meikle, Remodeling the dentofacial skeleton: the biological basis of orthodontics and dentofacial orthopedics, *J Dent Res*. **86** (2007), 12-24.
[73] J.P. Schmitt, E.P. Debold, F. Ahmad, A. Armstrong, A. Frederico, D.A. Conner, U. Mende, M.J. Lohse, D. Warshaw, C.E. Seidman, and J.G. Seidman, Cardiac myosin missense mutations cause dilated cardiomyopathy in mouse models and depress molecular motor function, *Proc Natl Acad Sci U S A*. **103** (2006), 14525-30.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-214

IV.3. Bioreactors in Tissue Engineering

Niamh PLUNKETT [a,b] and Fergal J. O'BRIEN [a,b]
[a] *Department of Anatomy, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland*
[b] *Trinity Centre for Bioengineering, Department of Mechanical Engineering, Trinity College Dublin, Dublin 2, Ireland*

Introduction

What is a Bioreactor?

Bioreactors have been used for many years in areas other than tissue engineering. They have been used in diverse areas such as in fermentation, in water treatment, in food processing and in the production of pharmaceuticals [1]. All of these bioreactors are devices in which biological or biochemical processes develop under a closely monitored and tightly controlled environment. Bioreactors have been used in animal cell culture since the 1980s in order to produce vaccines and other drugs and to culture large cell populations. Bioreactors for use in tissue engineering have progressed from such devices.

A tissue engineering bioreactor can be defined as a device that uses mechanical means to influence biological processes [2]. In tissue engineering, this generally means that bioreactors are used to stimulate cells and encourage them to produce extra-cellular matrix (ECM). There are numerous types of bioreactor which can be classified by the means they use to stimulate cells. A number of these will be discussed below.

Why are Bioreactors Needed in Tissue Engineering?

Tissue engineering technologies are based on the biological triad of cells, signalling mechanisms and extracellular matrix. To simulate the development of tissues *in vitro*, tissue engineering aims to optimise cell growth, by providing regulatory signals in the form of growth factors and a regeneration template in the form of scaffold. Bioreactors may be used as an alternative to or in conjunction with growth factors in the signalling part of the triad. Defects requiring tissue-engineering solutions are typically many millimetres in size [3]. Scaffolds in such a size range are easily fabricated. However, problems arise when culturing cells on these scaffolds. Static culture conditions result in scaffolds with few cells in the centre of the construct [4]. This heterogeneous cell distribution is a major obstacle to developing any three-dimensional tissue or organ *in vitro*. It has been shown that despite homogeneous cell seeding, after long periods in culture, more cells are found on the periphery of demineralised trabecular bone constructs [4]. This is due to cell necrosis and cell chemotaxis. Necrosis occurs at the centre of the scaffold due to a lack of nutrient delivery to, and waste removal from that area. The only mechanism by which nutrients and waste can move when a scaffold is in

static culture is by diffusion. As the size of the scaffold increases, diffusion to the centre of the construct becomes more difficult. In addition, as cells on the periphery grow and secrete extracellular matrix, movement of fluid to the interior of the scaffold is further impeded. Chemotaxis of the cells from the interior towards the periphery occurs because of the concentration gradient in nutrients that has been set up [3]. Nutrient concentration is greater at the periphery so cells move along this gradient towards the periphery in order to obtain the nutrients they require. It has been found that mineralised bone matrix reaches a maximum penetration depth of 240 μm on poly(DL-lactic-*co*-glycolic acid) scaffolds seeded with stromal osteoblasts, which is far thinner than an ideal bone graft replacement [5].

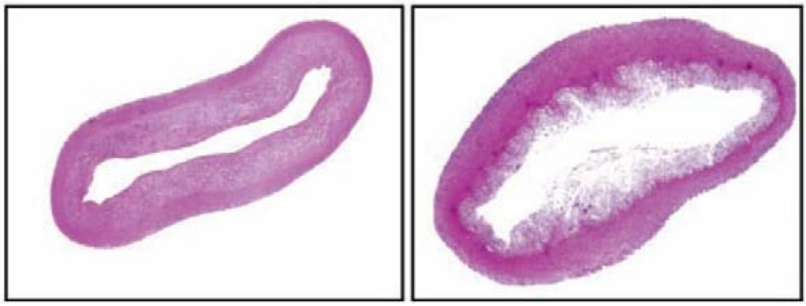


Figure 1. Histological sections of tissue engineered vascular tissue cultured using a bioreactor (on left) and statically (on right). Sections stained with Verhoff's elastin stain [6].

Figure 1 shows the distribution of smooth muscle cells in tissue engineered vascular tissue. Homogeneous distribution has been achieved by using a bioreactor [6]. In order to increase cell viability throughout a scaffold, fluid transport needs to be enhanced. A bioreactor may be used to achieve this aim. In addition to enhancing cell distribution, another important aspect of bioreactor use is cell stimulation. Cells respond to mechanical stimulation and bioreactors can be used to apply such stimulation. This can encourage cells to produce ECM in a shorter time period and in a more homogeneous manner than would be the case with static culture. For example, in comparisons between ECM protein levels after 5 weeks in culture, scaffolds cultured under hydrostatic pressure showed significant improvements over scaffolds cultured in static medium [7]. A benefit of ECM production is the increase in mechanical stiffness that it provides to the construct. A six-fold increase in equilibrium aggregate modulus (an intrinsic property of cartilage which is a measure of stiffness) was found after 28 days of culture in a compression bioreactor compared to free swelling controls [8].

Another important application of bioreactors is in cellular differentiation. Mechanical stimulation can be used to encourage stem cells down a particular path and hence provide the cell phenotype required. Bioreactors can provide biochemical and physical regulatory signals that guide differentiation [9]. There is great potential for using mesenchymal stem cells and other multipotent cells to generate different cell types and bioreactors can play an important role in this process.

Bioreactor Design Requirements

In general, bioreactors are designed to perform at least one of the following five functions: 1) providing uniform cell distribution, 2) maintaining the desired concentration of gases and nutrients in the medium, 3) providing mass transport to the tissue, 4) exposing tissue to physical stimuli, or 5) providing information about the formation of 3D tissue [10].

While the detailed requirements for bioreactor design are tissue- and/or application- specific, there are a few general principles which have to be adhered to when developing a bioreactor. The material selection is very important as it is vital to ensure that the materials used to create the bioreactor do not elicit any adverse reaction from the cultured tissue. Any material which is in contact with media must be biocompatible or bioinert. This eliminates the use of most metals, although stainless steel can be used if it is treated so that chromium ions do not leach out into the medium. Numerous plastics comply with this constraint but there are further limitations on material selection that must also be kept in mind. Materials must be usable at 37°C in a humid atmosphere. They must be sterilisable if they are to be re-used. Bioreactor parts can be sterilised by autoclaving or disinfected by submersion in alcohol. If they are to be autoclaved, materials that can withstand numerous cycles of high temperature and pressure must be used in bioreactor manufacture. Alternatively, some non-sterilisable disposable bioreactor parts may be used which can be replaced after each use of the bioreactor. Other material choices are between transparent or opaque and flexible or inflexible materials. Materials with different properties are needed for various components in the bioreactor. For example, transparent materials can be of benefit in allowing the construct to be monitored in the bioreactor during culture while flexible tubing can help with assembly of the bioreactor.

The design of the bioreactor should be as simple as possible, avoiding the introduction of, for example, machined recesses which could collect condensed steam during autoclaving and become breeding grounds for micro-organisms. Simplicity in design should also mean that the bioreactor is quick to assemble and disassemble. Apart from being more efficient, this ensures that cell-seeded constructs inserted into the bioreactor are out of the incubator for the minimum amount of time possible. This minimises the risk to the cells and the experiment being undertaken.

The specific application of the bioreactor must be kept in mind during the design process to ensure that all the design constraints are met. If various parameters such as pH, nutrient concentration or oxygen levels are to be monitored, these sensors should be incorporated into the design. If a pump or motor is to be used, it must be small enough to fit into an incubator and also be usable at 37°C and in a humid environment. The forces needed for cellular stimulation are very small so it is important to ensure that the pump/motor has the capability to apply small forces accurately. In any design involving fluids, problems can arise with leaking fluid seals and, if possible, the need for seals should be removed. However, in most cases, fluid seals are necessary and good design should decrease the problems with them. If a prototype bioreactor is being designed, it is worthwhile thinking about scale up opportunities for the bioreactor from the outset. This may mean designing a device that is relatively easy to enlarge without changing its characteristics or designing a simple device of which many more can be made so that numerous scaffolds can be cultured at one time.

1. Bioreactors in Tissue Engineering

1.1. Spinner Flask Bioreactor

Continuous stirred-tank reactors are commonly used in bioprocessing, for example in solid-state fermentation of organisms such as yeast [11] and spinner flask bioreactors for use in tissue engineering progressed from these devices. In a spinner flask (Figure 2), scaffolds are suspended at the end of needles in a flask of culture media. A magnetic stirrer mixes the media and the scaffolds are fixed in place with respect to the moving fluid. Flow across the surface of the scaffolds results in eddies in the scaffolds' superficial pores. Eddies are turbulent instabilities consisting of clumps of fluid particles that have a rotational structure superimposed on the mean linear motion of the fluid particles. They are associated with transitional and turbulent flow. It is via these eddies that fluid transport to the centre of the scaffold is thought to be enhanced [3]. Typically, spinner flasks are around 120 ml in volume (although much larger flasks of up to 8 litres have been used), are run at 50-80 rpm and 50% of the medium used in them is changed every two days [12]. Cartilage constructs have been grown in spinner flasks to thicknesses of 0.5 mm [13]. While this is an improvement on cartilage grown in static medium, it is still too thin for clinical use. Mass transfer in the flasks is not good enough to deliver homogeneous cell distribution throughout scaffolds and cells predominantly reside on the construct periphery [3].

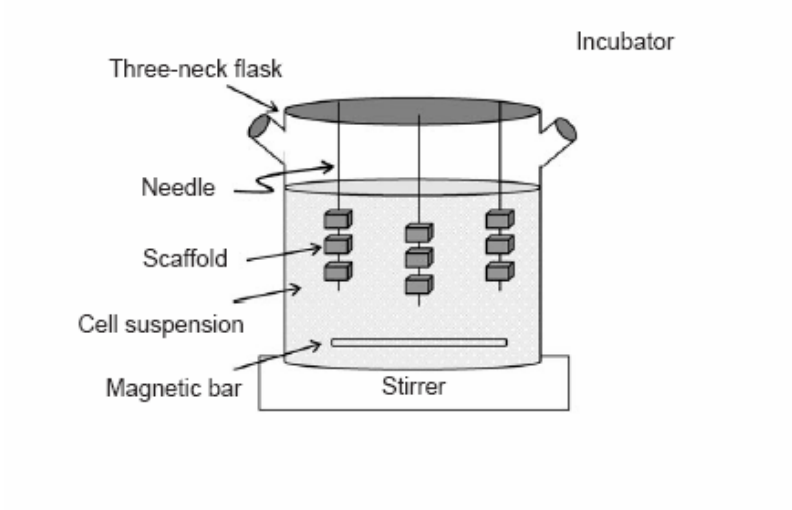


Figure 2. A spinner flask bioreactor. Scaffolds are suspended in medium and the medium stirred using a magnetic stirrer to improve nutrient delivery to the scaffold (Kim et al., 2005)

1.2. Rotating Wall Bioreactor

The rotating wall bioreactor was developed by NASA [14]. It was originally designed with a view to protecting cell culture experiments from high forces during space shuttle take off and landing. However, the device has proved useful in tissue engineering here

on earth. In a rotating wall bioreactor (Figure 3), scaffolds are free to move in media in a vessel. The wall of the vessel rotates, providing an upward hydrodynamic drag force that balances with the downward gravitational force, resulting in the scaffold remaining suspended in the media. Fluid transport is enhanced in a similar fashion to the mechanism in spinner flasks and the devices also provide more homogeneous cell distribution than static culture [3]. Gas exchange occurs through a gas exchange membrane and the bioreactor is rotated at speeds of 15-30 rpm. Cartilage tissue of 5 mm thickness has been grown in this type of bioreactor after seven months of culture [15]. As tissue grows in the bioreactor, the rotational speed must be increased in order to balance the gravitational force and ensure the scaffold remains in suspension.

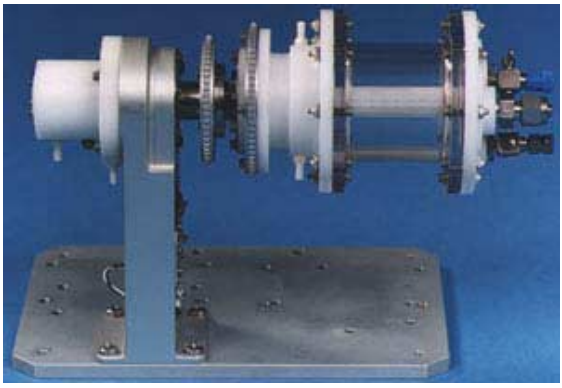


Figure 3. A rotating wall vessel bioreactor. Scaffolds are suspended in medium due to opposing gravitational and drag forces (www.che.msstate.edu/research/MDERL/demos/DEMoBioRctrExperiment)

1.3. Compression Bioreactors

Another widely used type of bioreactor is the compression bioreactor. This class of bioreactor is generally used in cartilage engineering and can be designed so that both static loading and dynamic loading can be applied. This is because static loading has been found to have a negative effect on cartilage formation while dynamic loading, which is more representative of physiological loading, has provided better results than many other stimuli [16].

In general, compression bioreactors consist of a motor, a system providing linear motion and a controlling mechanism. An example of such a system is shown in Figure 4, where a cam-follower system is used to provide displacements of different magnitudes and frequencies. A signal generator can be used to control the system and load cells and linear variable differential transformers can be used to measure the load response and imposed displacement respectively [8, 17]. The load can be transferred to the cell-seeded constructs via flat platens which distribute the load evenly, however in a device for stimulating multiple scaffolds simultaneously, care must be taken that the constructs are of similar height or the compressive strain applied will vary as the

scaffold height does. Mass transfer is improved in dynamic compression bioreactors over static culture (as compression causes fluid flow in the scaffold) and dynamic compression can also improve the aggregate modulus of the resulting cartilage tissue to levels approaching those of native articular cartilage [8].

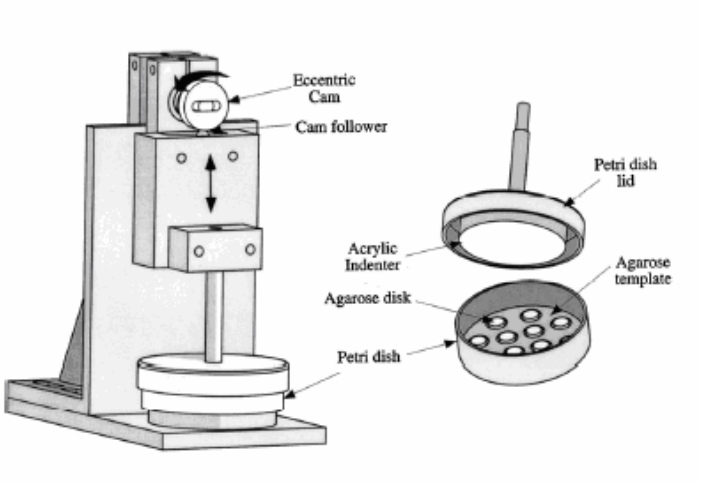


Figure 4. A compression bioreactor. Constructs are housed in a petri dish and subjected to compressive forces exerted on them via a cam and follower mechanism [41].

1.4. Strain Bioreactors

Tensile strain bioreactors have been used in an attempt to engineer a number of different types of tissue including tendon, ligament, bone, cartilage and cardiovascular tissue. Some designs are very similar to compression bioreactors, only differing in the way the force is transferred to the construct. Instead of flat platens as in a compression bioreactor, a way of clamping the scaffold into the device is needed so that a tensile force can be applied. Tensile strain has been used to differentiate mesechymal stem cells along the chondrogenic lineage. A multistation bioreactor was used in which cell-seeded collagen-glycosaminoglycan scaffolds were clamped and loaded in uniaxial tension [18]. Alternatively, tensile strain can also be applied to a construct by attaching the construct to anchors on a rubber membrane and then deforming the membrane. This system has been used in the culture of bioartificial tendons with a resulting increase in Young's modulus over non-loaded controls [19].

1.5. Hydrostatic Pressure Bioreactors

In cartilage tissue engineering, hydrostatic pressure bioreactors can be used to apply mechanical stimulus to cell-seeded constructs. Scaffolds are usually cultured statically and then moved to a hydrostatic chamber for a specified time for loading. Hydrostatic pressure bioreactors consist of a chamber which can withstand the pressures applied and a means of applying that pressure (Figure 5). For example, a media-filled pressure

chamber can be pressurised using a piston controlled by an actuator [16]. For sterility, the piston can apply pressure via an impermeable membrane so that the piston itself does not come into contact with the culture media. Variations on this design include a water-filled pressure chamber which pressurises a media-filled chamber via an impermeable film and is controlled using a variable backpressure valve and an actuator [20].

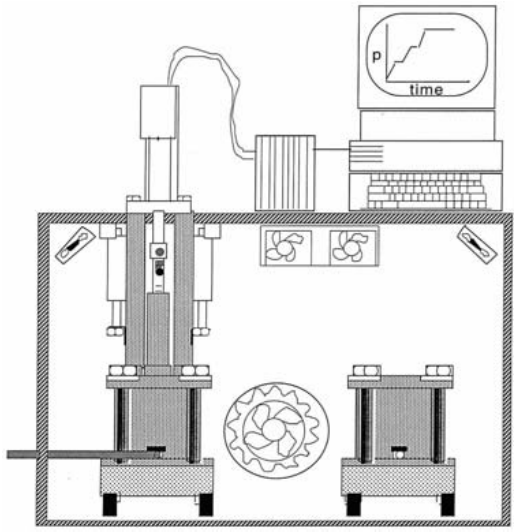


Figure 5. A hydrostatic pressure bioreactor. Constructs are placed in a chamber which is subsequently pressurised to the required level [16].

1.6. Flow Perfusion Bioreactor

Flow perfusion bioreactors generally consist of a pump and a scaffold chamber joined together by tubing. A media reservoir may also be present. The scaffold is kept in position across the flow path of the device (Figure 6). Media is perfused through the scaffold, thus enhancing fluid transport.

Culture using flow perfusion bioreactors has been shown to provide more homogeneous cell distribution throughout scaffolds. Collagen sponges have been seeded with bone marrow stromal cells and perfused with flow. This has resulted in greater cellularity throughout the scaffold in comparison to static controls, implying that better nutrient exchange occurs due to flow [21]. Using a biphasic calcium-phosphate scaffold, abundant ECM with nodules of CaP was noted after 19 days in steady flow culture [22].

In comparisons between flow perfusion, spinner flask and rotating wall bioreactors, flow perfusion bioreactors have proved to be the best for fluid transport. Using the same flow rate and the same scaffold type, while cell densities remained the same using all three bioreactors, the distribution of the cells changed dramatically depending on which bioreactor was used. Histological analysis showed that spinner flask and static culture resulted in the majority of viable cells being on the periphery of the scaffold. In

contrast, the rotating wall vessel and flow perfusion bioreactor culture resulted in uniform cell distribution throughout the scaffolds [3, 23]. After 14 days in culture, the perfusion bioreactor had higher cell density than all other culture methods [3].

In our laboratory, a flow perfusion bioreactor has been developed to examine the effects of different flow profiles on cell-seeded collagen-glycosaminoglycan scaffolds [24]. The scaffold chamber was specifically designed to ensure that the compliant scaffold was under perfusive flow. This involved using a scaffold of larger diameter than the flow path and using spacers to ensure the scaffold was under 10% compression during culture. A programmable syringe pump was used in order to stimulate the cell-seeded constructs using different flow profiles. As discussed below, this device demonstrated that intermittent flow perfusion is advantageous for mechanically stimulating osteoblasts while maintaining cell viability.

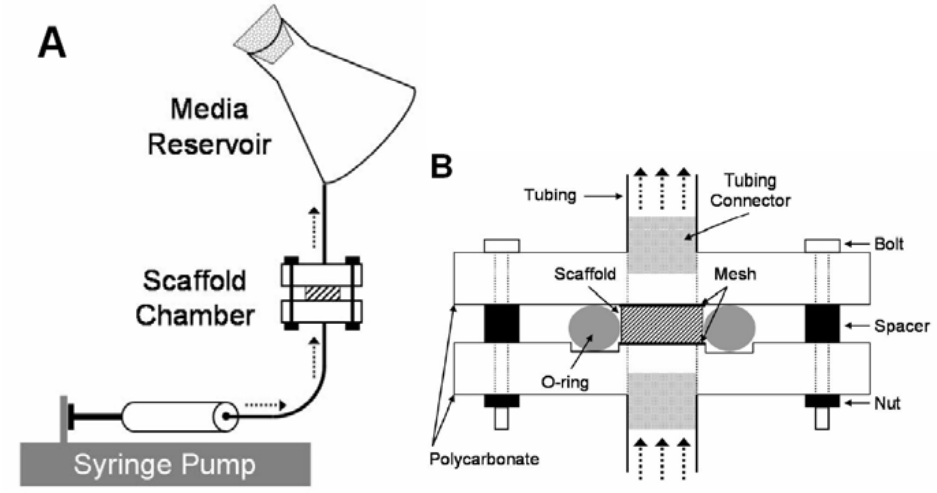


Figure 6. A flow perfusion bioreactor (A). Media is forced through the scaffold in the scaffold chamber (B) by the syringe pump [24].

1.7. Combined Bioreactors

In addition to the bioreactors mentioned thus far, numerous combinations of the different types of bioreactor have been used in order to better mimic the *in vivo* environment *in vitro*. In most cases, these more complicated bioreactors involve adding a perfusion loop on top of a standard bioreactor. Examples include compression, tensile strain or hydrostatic bioreactors with added perfusion [7, 20, 25]. These different bioreactors allow for nutrient exchange to take place due to perfusion while stimulation occurs due to a different mechanical stimulus. Bioreactors for engineering very specialised tissues have also been developed. One example of such a bioreactor is a combined tensile and vibrational bioreactor for engineering vocal fold tissue [26]. This bioreactor mimics the physiological conditions that human vocal folds experience in an

attempt to develop the tissue in the laboratory (Figure 7). Vocal fold tissue experiences vibrations of 100-1000 Hz at an amplitude of approximately 1 mm, which is a unique stimulus in the human body [27].

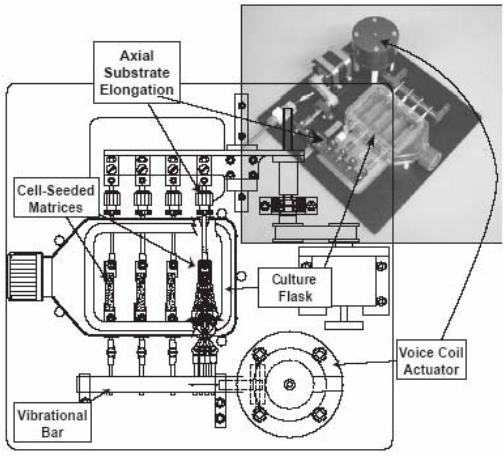


Figure 7. A bioreactor for vocal fold tissue engineering. Constructs can be stimulated using both tensile and vibrational forces [26].

2. Tissue Formation in Bioreactor Systems

Tissue engineering of all three dimensional tissues require homogeneous cell distribution in order for homogeneous tissue to develop. Therefore, there is a need for bioreactor culture in numerous different disciplines in tissue engineering. Bioreactor culture has been used in a diverse range of applications including skin, bladder and liver tissue engineering [1, 28]. Here, the use of bioreactor culture in bone, cartilage and cardiovascular tissue engineering will be discussed.

2.1. Bone

Bone is comprised of a mineral phase, collagen and cells. The mineral phase of bone is a hard, brittle material (hydroxyapatite) that is strong in compression but cannot withstand large shear or tensile loads. In contrast, collagen fibres are strong in tension and flexible in bending but weak in compression. Bone grafts are required to aid bone defect and non-union healing. The aim in using them is to restore function to the damaged area as quickly and completely as possible [29]. They are required in a number of procedures including, for example, replacing diseased bone, filling bone voids after non-unions or cyst removal, reconstructive surgery and in spinal fusion operations. The most commonly used graft in bone replacement is the autograft. Autografts are grafts taken from the patients themselves, usually from the iliac crest,

although the distal femur, the greater trochanter, and the proximal tibia can also be used [29]. This process is expensive and the size of graft that can be obtained is limited. Morbidity of the site the graft is acquired from is another problem and complications can arise due to infection and chronic pain [29]. An alternative to the autograft is the allograft. The term allograft is used for bone grafts which are obtained from an organ donor. A drawback to this option is the danger of infection. Xenografts, acquired from animal bone, are another alternative. However, the morphology of the bone is different to that of humans and the risk of cross-species infection exists. Coral has also been used but its structure is very different to that of bone and hence osteointegration of this type of graft has proved difficult [30].

The use of natural bone grafts has proved problematic and therefore attention has turned to tissue engineering. Engineered tissue must be strong enough to allow load bearing after implantation. Resorbability of the scaffold is another important issue. Ideally, as the scaffold is resorbed, bone should be deposited in its place, thus ensuring no loss in mechanical strength. A morphology that allows movement of cells and supports vascularisation is important for a scaffold material. A high porosity and controllable pore size can provide this. Cells must be able to penetrate into the core of the scaffold so that they are homogeneously distributed throughout the graft. There must be space to allow for transport of nutrients to and waste removal from cells, so the pores must be interconnected [31]. Thus, a balance must be struck between mechanical strength and porosity to develop an ideal scaffold for bone tissue engineering. The use of a bioreactor with such a scaffold should provide a homogeneous distribution of stimulated cells.

It has been shown that fluid flow can stimulate bone cells to increase levels of bone formation markers [4, 32-37] and its use could improve mineralisation of the scaffold on which cells are seeded. Flow perfusion bioreactors increase alkaline phosphatase (ALP) expression after 7 and 14 days more than spinner flasks or rotating wall vessels [3] and are more commonly used than any other bioreactor for use in 3-D stimulation studies. In one study, MC3T3-E1 cells were seeded on decalcified human trabecular bone, the flow rate of perfusion altered and the mRNA expression of Runx2, OC and ALP measured [4]. It was found that using a steady flow rate of only 1 mL/min killed nearly all the cells on the scaffold after 7 days in culture. However, a flow rate of 0.01 mL/min led to a high proportion of viable cells both on the surface and inside the scaffold. This compared favourably to static culture, where cells were predominantly on the periphery [4].

Using a CaP scaffold and a flow rate of 0.025 mL/min in a flow perfusion bioreactor, PGE_2 levels were found to increase over static controls. When a stimulus of 30 minutes of oscillatory flow at 1 Hz with a 40 mL/min peak was superimposed on the steady flow, PGE_2 levels increased further. The number of cells left residing on the scaffolds decreased due to this large dynamic stimulus but this decrease was not found to be statistically significant [38]. When different flow profiles were used intermittently to stimulate cells on a highly porous collagen-glycosaminoglycan scaffold in our laboratory, it was found that intermittent flow caused greater stimulation than a continuous low flow rate without a loss in cell number [39]. Cyclyoxygenase-2 and osteopontin expression increased due to culture in the bioreactor, as did prostaglandin E_2 production. This lends further backing to the hypothesis that the combination of a perfusion period (for nutrient delivery and waste removal) and a stimulation period may deliver enhanced fluid transport with enhanced stimulation of cells and may yet prove to be the optimum regime for bioreactor culture of bone cells.

2.2. Cartilage

Cartilage is a supporting tissue containing chondroitin sulphates, collagen and elastic fibres and cells. The cells present in cartilage are known as chondrocytes and they are situated in lacunae in the cartilage matrix. Cartilage is avascular and nutrient and waste product exchange occurs purely by diffusion through the cartilage matrix. When minor damage occurs to cartilage, it can repair by appositional growth, but when severe damage occurs, the body cannot replace the cartilage [40]. Cartilage tissue engineering may offer the solution to this problem. There are three types of cartilage: hyaline cartilage, elastic cartilage and fibrocartilage. Joints can contain both hyaline cartilage and fibrocartilage, with the more flexible hyaline cartilage covering the bone and the more durable fibrocartilage acting as a shock-absorber between bones. As a joint moves, there is motion between two articulating layers of cartilage. This deforms the cartilage, causes fluid flow within it and induces a hydrostatic pressure load on it. These mechanical forces affect the chondrocytes in the cartilage. The force applied, along with the length of time it is applied for and the frequency of application modifies the response of chondrocytes [41]. This is useful in bioreactor design and for use in cartilage tissue engineering; if the correct stimulation pattern is used, chondrocytes can be induced to produce more extracellular matrix and this can result in more cartilage-like tissue being formed.

After twenty weeks in static culture, the aggregate modulus of tissue-engineered cartilage was 179±9 kPa. This is 40% of the value reported for native cartilage [42]. At twenty five weeks, the modulus had not increased further so this may be the closest approximation to cartilage that can be cultured without the aid of a bioreactor. The most commonly used bioreactors in cartilage tissue engineering are compression bioreactors. When free swelling controls were compared to dynamically loaded agarose gels, after 28 days in culture, there was a six-fold increase in the equilibrium aggregate modulus for the loaded gels [8]. A sinusoidal strain of 10% at 1 Hz was applied to the gels for five days per week for a total of three hours per day with a rest period of one hour between each hour of loading. This complex loading pattern was deemed to be physiological and it resulted in increased glycosaminoglycan content over free swelling controls after 21 days in culture. The combination of increased modulus and increased glycosaminoglycan formation over free swelling controls after only four weeks in culture demonstrates the benefits of bioreactor culture in cartilage tissue engineering.

Compression bioreactors have also been used to examine the effect of loading on the differentiation of bone marrow mesenchymal stem cells down the chondrocytic lineage. Growth factors such as transforming growth factor (TGF-β) can also be used to encourage differentiation. In a study comparing the use of compressive loading, the use of TGF-β and the use of a combination of loading and TGF-β, it was found that compressive loading alone was just as effective at inducing chondrogenic differentiation as TGF-β or TGF-β plus loading [17].

2.3. Cardiovascular Tissue

There are three main types of tissue that are encompassed by the heading cardiovascular tissue, namely: vascular, cardiac muscle and heart valve tissue. These different tissues experience different mechanical environments in the body and therefore, in tissue engineering, they are cultured using different bioreactors and under

different loading conditions in an attempt to recreate physiological loading conditions in the laboratory.

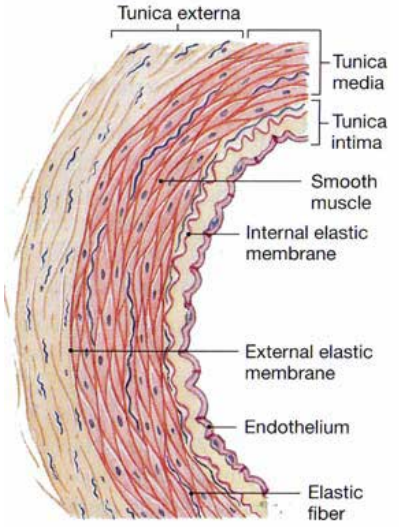


Figure 8. Structure of an artery. The three layers of the tissue (the tunical intima, the tunica media and the tunica externa) can be clearly seen [40].

2.3.1. Vascular tissue

Blood vessels must be flexible enough to move during physiological activities and yet tough enough to withstand the pressure changes that occur as the heart beats. Blood vessels consist of three layers of tissue: the tunical intima, the tunica media and the tunica externa. The intima (innermost part of the blood vessel) consists of an endothelial lining and a layer of connective tissue containing elastic fibres. The media consists of smooth muscle tissue and loose connective tissue and in arteries, the externa consists of connective tissue and collagen fibres (Figure 8). Arteriosclerosis is associated with half of all deaths in the United States each year [40]. It is a thickening and toughening of the arterial walls and can lead to coronary heart disease if it occurs in the coronary arteries or stroke if it occurs in the arteries supplying the brain with blood. Coronary heart disease can be treated by bypass surgery, when a section of artery or vein is removed from elsewhere in the body and used to bypass the blockage. Drawbacks with this procedure are that it creates another trauma site in the body and often, sufficient tissue for bypass is not available [6]. Research into tissue engineering of vascular tissue is proving promising for growing arteries *in vitro* that may be used instead of the patient's own blood vessels.

The ideal graft for vascular bypass consists of a confluent endothelium at the lumen and a smooth muscle layer surrounding it with sufficient mechanical integrity for suture retention and tolerance of arterial pressures. The endothelial layer is important as a confluent layer of endothelial cells inhibits the proliferation of smooth

muscle cells. The smooth muscle cells are therefore more likely to be quiescent and this decreases the risk of thrombosis formation and luminal occlusion [6]. In order to recreate the structure of blood vessels in vitro, co-cultures of different cell types must be used. Scaffolds can be cultured with smooth muscle cells initially and then a layer of endothelial cells can be seeded as the culture period nears its end.

The arteries nearest the heart experience the highest pressures and they have resilient walls in order to cope with these changing pressures. They are known as elastic arteries and their tunica media contains a high proportion of elastic fibres. This enables them to expand as the pressure inside the lumen increases and recoil when the pressure decreases, thus allowing them to deal with pressure changes but also to damp the pressure oscillations to make blood flow continuous rather than oscillatory in nature [40]. It is this expansion and recoil of blood vessels that is the inspiration for a number of bioreactors for use in vascular tissue engineering.

Some success has been achieved when culturing vascular tissue in both rotating wall vessels [13] and biomimetic bioreactors [6, 25]. Biomimetic bioreactors use combinations of strain and perfusion to stimulate vascular tissue development. In one such bioreactor, a tube of biodegradable polymer was seeded with smooth muscle cells and silicone tubing placed through the centre of the tube. The construct was maintained in a bath of stirred media and the highly distensible silicone tubing was pulsed at 165 beats per minute so that the scaffold experienced 5% radial strain over an eight week culture period [6]. The silicone tubing was then removed, an endothelial cell lining applied to the inner layer of the construct and culture medium perfused through the centre of the scaffold. This set up resulted in vessels with a rupture strength similar to that of native tissue and with a morphology that showed migration of smooth muscle cells throughout the construct. In contrast, scaffolds kept in static culture over the same culture period showed no migration of smooth muscle cells and therefore an inhomogeneous structure. A bioreactor has also been designed to apply both radial and axial strain to cell-seeded constructs, to mimic the environment in the body more closely [25]. The bioreactor has two perfusion loops that feed the internal and external parts of the scaffold. The internal perfusion can be pulsatile in nature and is of variable frequency.

2.3.2 Cardiac Muscle Tissue

Cardiac muscle cells are called cardiocytes or cardiomyocytes. They form extensive connections with each other and are almost totally dependent on aerobic metabolism for energy. When a myocardial infarction (heart attack) occurs, part of the blood supply to the heart muscle is blocked and cardiac muscle cells die due to lack of oxygen delivery [40]. Cardiac muscle only has a limited ability to regenerate and scar tissue formed may create additional problems by constricting other vessels in the damaged region [40, 43]. Cardiac muscle tissue engineering is therefore of interest in developing a method for myocardial repair. Oxygen delivery to cardiac cells is vital in order that a sufficiently thick layer of tissue can be grown in vitro. This means that the use of bioreactors that increase oxygen delivery to cells is important in cardiac muscle tissue engineering. The ideal tissue-engineered cardiac muscle tissue graft should have a dense, uniform distribution of cardiomyocytes, should contract in response to an electrical stimulus and have the mechanical integrity to allow implantation [43].

Polyglycolic acid (PGA) scaffolds have been seeded with cardiomyocytes and cultured in spinner flasks and rotating wall bioreactors [44, 45]. In both these culturing systems, a peripheral region densely populated with cells and a central region with a

sparse population of cells was apparent. Spontaneous contractions of the tissue occurred during culture using the spinner flask and impulse propagation occurred due to electrical stimulation, but only in the peripheral region [44]. Use of a perfusion system, however, delivered a much more uniform distribution of cells with higher cell number on perfused scaffolds than scaffolds cultured in an orbitally mixed dish at all time points up to 7 days [43]. Spontaneous contractions were also observed in the perfusion system but only up to 5 days after seeding. Perfused scaffolds contracted synchronously and at constant frequency upon electrical stimulation after seven days, while dish-cultured scaffolds exhibited an arrhythmic contraction pattern [43].

2.3.3 Heart Valves

There are four valves in the heart that prevent backflow of blood from the ventricles to the atria or from the pulmonary trunk and aorta into the ventricles. Problems with these valves decrease the heart's ability to maintain circulatory flow [40]. In severe cases, valve replacement is the only option. However, thromboembolism is a substantial risk when mechanical valves are used to replace native valves and xenografts or other non-viable tissue-based grafts often fail due to calcification [46, 47]. Patients with congenital malformations of the valves often require a durable replacement that can grow as the patient does, and a tissue-engineered valve may therefore be the best option.

Heart valves experience a complex mechanical environment including high bending stress and high shear stress *in vivo* [10]. When grown in static culture, valve leaflets are fragile, have a rough surface and a low suture retention strength. In bioreactor culture under pulsatile flow, however, leaflets that are intact, mobile, pliable, competent during closure and have a tensile strength greater than that of native tissue are formed after 14 days [46]. In this case, the pulse was applied to the scaffolds by pumping air into a chamber connected to the scaffold chamber via a silicone diaphragm. Scaffolds were seeded with myofibroblasts initially and after a four day culture period, with endothelial cells. The pressure at which the pulse was applied was increased over time, as was the flow rate of the media. Thus, as the scaffold became stronger, the forces applied to it were increased. The resulting valve leaflets were implanted into lambs and there was no evidence of thrombus formation up to 20 weeks after valve replacement. There was some evidence of pulmonary regurgitation at 16 and 20 weeks, however. Optimisation of many aspects of the culturing process remains to be completed but bioreactor culture of valve leaflets has enabled a functioning heart valve to be developed *in vitro* [46].

3. Bioreactors: State of the Art and Future Directions

The use of bioreactors has brought us a step closer to engineering numerous tissue types in the laboratory. At present, most bioreactors are specialised devices with low volume output. Their assembly is often time consuming and labour intensive. Many also exhibit operator dependent variability. While scaled-up versions of some devices may be useful for developing larger amounts of tissue, problems with process consistency and process contamination may persist. A better understanding of the different effects of mechanical stimulation on cell signalling and mechanotransduction is also needed. This can be achieved through the use of existing simple bioreactors in conjunction with numerical simulation of culture conditions to minimise the number of experiments needed.

In the future, ways of minimising the time and effort needed in order to form tissue must be found if costs are to be minimised and the use of engineered tissue is to become routine clinically. One way to do this is to automate the process. The ideal bioreactor would need autologous material and a scaffold as inputs and, after a defined culture period, would output the required amount of the required tissue. Automated systems for culturing cells already exist and work has begun on extending them to incorporate mechanical stimulation into the culturing process. Aastrom has developed a system that takes bone marrow as an input and expands the stem and progenitor cell population. The Automation Partnership has numerous systems for expanding cells. If systems such as these are used in conjunction with monitoring systems and a feedback loop, so that factors such as the temperature, oxygen level and pH can be regulated, cell culture can be optimised. Closed bioreactor systems for seeding and culturing skin grafts under perfusion were developed by Advanced Tissue Sciences before the company's liquidation in 2002. As this system demonstrated, the technology exists to enable the incorporation of mechanical stimulus into an automated cell culture system and this may be the future for bioreactors in tissue engineering.

References

[1] I. Martin, D. Wendt and M. Heberer, The role of bioreactors in tissue engineering, *Trends Biotechnol* **22** (2004), 80-6.
[2] E. M. Darling and K. A. Athanasiou, Biomechanical strategies for articular cartilage regeneration, *Ann Biomed Eng* **31** (2003), 1114-24.
[3] A. S. Goldstein, T. M. Juarez, C. D. Helmke, M. C. Gustin and A. G. Mikos, Effect of convection on osteoblastic cell growth and function in biodegradable polymer foam scaffolds,*Biomaterials* **22** (2001), 1279-88.
[4] S. H. Cartmell, B. D. Porter, A. J. Garcia and R. E. Guldberg, Effects of medium perfusion rate on cell-seeded three-dimensional bone constructs in vitro, *Tissue Eng* **9** (2003), 1197-203.
[5] S. L. Ishaug, G. M. Crane, M. J. Miller, A.W. Yasko, M. J. Yaszemski and A. G. Mikos, Bone formation by three-dimensional stromal osteoblast culture in biodegradable polymer scaffolds, *J. Biomed. Mater. Res*. **36** (1997), 17-28.
[6] L. E. Niklason, J. Gao, W. M. Abbott, K. K. Hirschi, S. Houser, R. Marini and R. Langer, Functional arteries grown *in vitro*, *Science* **284** (1999), 489-93.
[7] S. E. Carver and C. A. Heath, Semi-continuous perfusion system for delivering intermittent physiological pressure to regenerating cartilage, *Tissue Eng* **5** (1999), 1-11.
[8] R. L. Mauck, M. A. Soltz, C. C. Wang, D. D. Wong, P. H. Chao, W. B. Valhmu, C. T. Hung and G. A. Ateshian, Functional tissue engineering of articular cartilage through dynamic loading of chondrocytes seeded agarose gels, *J Biomech Eng* **122** (2000), 252-60.
[9] G. H. Altman, R. L. Horan, I. Martin, J. Farhadi, P. R. Stark, V. Volloch, J. C. Richmond, G. Vunjak-Novakovic and D. L. Kaplan, Cell differentiation by mechanical stress, *Faseb J* **16** (2002), 270-2.
[10] V. Barron, E. Lyons, C. Stenson-Cox, P. E. McHugh and A. Pandit, Bioreactors for cardiovascular cell and tissue growth: A review, *Ann Biomed Eng* **31** (2003), 1017-30.
[11] Y. Martin and P. Vermette, Bioreactors for tissue mass culture: Design, characterization, and recent advances, *Biomaterials* **26** (2005), 7481-503.
[12] R. I. Freshney, Culture of animal cells, Wiley-Liss, New York, 2000.
[13] L. E. Freed and G. Vunjak-Novakovic, Tissue engineering bioreactors. in: R. P. Lanza, R. Langer and J. Vacanti, (Eds.), Principles of tissue engineering, Academic Press, San Diego, CA, 2000, pp. 143-156.
[14] R. P. Schwarz, T. J. Goodwin and D. A. Wolf, Cell culture for three-dimensional modeling in rotating wall vessels: An application of simulated microgravity, *J Tissue Cult Methods* **14** (1992), 51-7.
[15] L. E. Freed, R. Langer, I. Martin, N. R. Pellis and G. Vunjak-Novakovic, Tissue engineering of cartilage in space, *Proc Natl Acad Sci U.S.A* **94** (1997), 13885-90.
[16] E. M. Darling and K. A. Athanasiou, Articular cartilage bioreactors and bioprocesses, *Tissue Eng* **9** (2003), 9-26.

[17] C. Y. Huang, K. L. Hagar, L. E. Frost, Y. Sun and H. S. Cheung, Effects of cyclic compressive loading on chondrogenesis of rabbit bone-marrow derived mesenchymal stem cells, *Stem Cells* **22** (2004), 313-23.
[18] L. A. McMahon, A. J. Reid, V. A. Campbell and P. J. Prendergast, Regulatory effects of mechanical strain on the chondrogenic differentiation of mscs in a collagen-gag scaffold: Experimental and computational analysis, *Ann Biomed Eng* **36** (2008), 185-94.
[19] J. Garvin, J. Qi, M. Maloney and A. J. Banes, Novel system for engineering bioartificial tendons and application of mechanical load, *Tissue Eng* **9** (2003), 967-79.
[20] S.Watanabe, S. Inagaki, I. Kinouchi, H. Takai, Y. Masuda and S. Mizuno, Hydrostatic pressure/perfusion culture system designed and validated for engineering tissue, *J Biosci Bioeng* **100** (2005), 105-11.
[21] J. Glowacki, S. Mizuno and J. S. Greenberger, Perfusion enhances functions of bone marrow stromal cells in three-dimensional culture, *Cell Transplant* **7** (1998), 319-26.
[22] F. W. Janssen, J. Oostra, A. Oorschot and C. A. van Blitterswijk, A perfusion bioreactor system capable of producing clinically relevant volumes of tissue-engineered bone: in vivo bone formation showing proof of concept, *Biomaterials* **27** (2006), 315-23.
[23] X. Yu, E. A. Botchwey, E. M. Levine, S. R. Pollack and C. T. Laurencin, Bioreactor-based bone tissue engineering: The influence of dynamic flow on osteoblast phenotypic expression and matrix mineralization, *Proc Natl Acad Sci* U S A **101** (2004), 11203-8.
[24] M. J. Jaasma, N. A. Plunkett and F. J. O'Brien, Design and validation of a dynamic flow perfusion bioreactor for use with compliant tissue engineering scaffolds, *J Biotechnol*. **133** (2008), 490.
[25] K. Bilodeau, F. Couet, F. Boccafoschi and D. Mantovani, Design of a perfusion bioreactor specific to the regeneration of vascular tissues under mechanical stresses, *Artif Organs* **29** (2005), 906-12.
[26] I. R. Titze, R. W. Hitchcock, K. Broadhead, K. Webb, W. Li, S. D. Gray and P. A. Tresco, Design and validation of a bioreactor for engineering vocal fold tissues under combined tensile and vibrational stresses, *J Biomech* **37** (2004), 1521-9.
[27] I. R. Titze, On the relation between subglottal pressure and fundamental frequency in phonation, *J Acoust Soc Am* **85** (1989), 901-6.
[28] M. C. Wallis, H. Yeger, L. Cartwright, Z. Shou, M. Radisic, J. Haig, M. Suoub, R. Antoon and W. A. Farhat, Feasibility study of a novel urinary bladder bioreactor, *Tissue Eng* (2007).
[29] C. R. Perry, Bone repair techniques, bone graft, and bone graft substitutes, *Clin Orthop Relat Res* **360** (1999), 71-86.
[30] J. Y. de la Caffiniere, E. Viehweger and A. Worcel, [long-term radiologic evolution of coral implanted in cancellous bone of the lower limb. Madreporic coral versus coral hydroxyapatite], *Rev Chir Orthop Reparatrice Appar Mot* **84** (1998), 501-7.
[31] F. J. O'Brien, B. A. Harley, I. V. Yannas and L. Gibson, Influence of freezing rate on pore structure in freeze-dried collagen-gag scaffolds, *Biomaterials* **25** (2004), 1077-86.
[32] N. N. Batra, Y. J. Li, C. E. Yellowley, L. You, A. M. Malone, C. H. Kim and C. R. Jacobs, Effects of short-term recovery periods on fluid-induced signaling in osteoblastic cells, *J Biomech* **38** (2005), 1909-17.
[33] M. R. Kreke, W. R. Huckle and A. S. Goldstein, Fluid flow stimulates expression of osteopontin and bone sialoprotein by bone marrow stromal cells in a temporally dependent manner, *Bone* **36** (2005), 1047-55.
[34] Y. J. Li, N. N. Batra, L. You, S. C. Meier, I. A. Coe, C. E. Yellowley and C. R. Jacobs, Oscillatory fluid flow affects human marrow stromal cell proliferation and differentiation, *J Orthop Res* **22** (2004), 1283-9.
[35] K. M. Reich and J. A. Frangos, Effect of flow on prostaglandin E2 and inositol trisphosphate levels in osteoblasts, *Am J Physiol* **261** (1991), C428-32.
[36] J. You, G. C. Reilly, X. Zhen, C. E. Yellowley, Q. Chen, H. J. Donahue and C. R. Jacobs, Osteopontin gene regulation by oscillatory fluid flow via intracellular calcium mobilization and activation of mitogen activated protein kinase in MC3T3-E1 osteoblasts, *J Biol Chem* **276** (2001), 13365-71.
[37] J. Klein-Nulend, E. H. Burger, C. M. Semeins, L. G. Raisz and C. C. Pilbeam, Pulsating fluid flow stimulates prostaglandin release and inducible prostaglandin g/h synthase mrna expression in primary mouse bone cells, *J. Bone Min. Res.* **12** (1997), 45-51.
[38] J. Vance, S. Galley, D. F. Liu and S. W. Donahue, Mechanical stimulation of mc3t3 osteoblastic cells in a bone tissue-engineering bioreactor enhances prostaglandin E_2 release, *Tissue Eng* **11** (2005), 1832-9.
[39] M. J. Jaasma and F. J. O'Brien, Mechanical stimulation of osteoblasts using steady and dynamic fluid flow, *Tissue Eng Part A* **14** (2008), 1213-23.
[40] F. Martini, Fundamentals of anatomy and physiology, Prentice Hall, 2002.

[41] R. L. Mauck, B. A. Byers, X. Yuan and R. S. Tuan, Regulation of cartilaginous ecm gene transcription by chondrocytes and mscs in 3d culture in response to dynamic loading, *Biomech Model Mechanobiol* **6** (2007), 113-25.
[42] P. X. Ma and R. Langer, Morphology and mechanical function of long-term in vitro engineered cartilage, *J Biomed Mater Res* **44** (1999), 217-21.
[43] M. Radisic, L. Yang, J. Boublik, R. J. Cohen, R. Langer, L. E. Freed and G. Vunjak-Novakovic, Medium perfusion enables engineering of compact and contractile cardiac tissue, *Am J Physiol Heart Circ Physiol* **286** (2004), H507-16.
[44] N. Bursac, M. Papadaki, R. J. Cohen, F. J. Schoen, S. R. Eisenberg, R. Carrier, G. Vunjak-Novakovic and L. E. Freed, Cardiac muscle tissue engineering: Toward an in vitro model for electrophysiological studies, *Am J Physiol* **277** (1999), H433-44.
[45] M. Papadaki, N. Bursac, R. Langer, J. Merok, G. Vunjak-Novakovic and L. E. Freed, Tissue engineering of functional cardiac muscle: Molecular, structural, and electrophysiological studies, *Am J Physiol Heart Circ Physiol* **280** (2001), H168-78.
[46] S. P. Hoerstrup, R. Sodian, S. Daebritz, J. Wang, E. A. Bacha, D. P. Martin, A. M. Moran, K. J. Guleserian, J. S. Sperling, S. Kaushal, J. P. Vacanti, F. J. Schoen and J. E. Mayer, Jr., Functional living trileaflet heart valves grown *in vitro, Circulation* **102** (2000), III44-9.
[47] R. J. Levy, F. J. Schoen,W. B. Flowers and S. T. Staelin, Initiation of mineralization in bioprosthetic heart valves: Studies of alkaline phosphatase activity and its inhibition by AlCl3 or FeCl3 preincubations, *J Biomed Mater Res* **25** (1991), 905-35.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-231

IV.4. Characterisation and Testing of Biomaterials

Sebastian DENDORFER[a], Joachim HAMMER[b] and Andreas LENICH[c]
[a] *AnyBody Technology A/S, Niels Jernes Vej 10, Aalborg, Denmark*
[b] *Laboratory for Materials Science, University of Applied Sciences Regensburg, Galgenbergstrasse 30, 93053 Regensburg, Germany*
[c] *Department of Trauma Surgery, Klinikum Rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany*

1. Introduction

The investigation of the structural mechanics of the entire body system is despite it's rather long history in science, compared to more recently developed research field like genetics, still extremely important and by no means outdated. Improved innovative treatments of the musculoskeletal system will require specifically designed implants and implant materials to guarantee optimal function of the injured segment and fast rehabilitation. As the actual loading situation of the implant-biomaterial compound is highly sensitive to implant-induced changes in the specific in vivo situation, the mechanical properties of living biological materials still require intense research activities, especially due to the large scatter in terms of biological variation.

Structural failure of biomaterials by plastic deformation or fracture is observed if the material dependent threshold strength is exceeded. This load inducing failure can be the result of a single load case or a superposition of different load combinations (i.e. monotonic or static loading). Furthermore, material failure can also be induced by repeated (cyclic) loading at loads which are significantly below the material or compound stability characterizing static strength. This failure type is generally the most common (already repeated load application in the order of 20 cycles can induce dynamic failure). It should be noted that the characteristic material or compound parameters are strongly affected and directly related to the material (micro-) structure. For biological materials this is generally described by biological scatter in terms of the mechanical morphogenesis. Therefore, any analysis concerning mechanical parameters or properties has to be interpreted in terms of the specific microstructure of the material.

The focus of this chapter is to give an introduction into the basic principles of load induced material behavior and failure under static and dynamic conditions with respect to specific material structure. Further information can be found in standard text books on biomechanics and solid mechanics [4], [14], [12],[13], [19], [20], [10], [24], [15], [16].

2. Mechanical Properties and Material Structure

Under in vivo conditions, materials and structures are exposed to a variety of different superimposed forces. Unfortunately, these in vivo loading conditions are rather complex and cannot be applied reproducibly in laboratory experiments. Therefore, it is common scientific practice to investigate the load bearing capability of materials or structures under idealized and standardized experimental conditions. These results are transferred to real components or segments with individual geometries by specific equations and computational methods.

How do biomaterials respond to specific types of loading due to external forces? Generally, all materials initially react on external forces (i.e. mechanical loads) by deformation followed by damage due to increasing deformation. The strength of materials is thus defined as the resistance of a material against deformation or fracture due to its atomic composition and its specific microstructure. The ratio of strength and acting forces has two aspects: The main function of hard materials (e.g. bones) is load bearing and thus they have to withstand high loads. For this purpose maximum stability is essential. On the other hand, soft materials (e.g. muscles) may react on external forces by high forming work and deformation without any risk of fracture. In this aspect maximum deformation potential is of pronounced interest.

To compare different materials concerning their mechanical properties (e.g. stability or ductility) and to exclude geometrical size effects, it is essential to normalise external forces and deformations: A bolt doubled in cross sectional area can carry doubled loads – therefore, it is appropriate to describe the material property strength by impingement forces, defined as stresses: Stress, MPa (N/mm^2) = Force, N per unit of area, $mm^{2;}$ $\sigma = F/A$. Stresses are directed and therefore, act as vectors affecting the planes of the loaded structure (Figure 1a). Axial stresses react normal to the surface while shear stresses are acting as tangential paired forces (Figure 1b).

Equivalent to the forces appropriately normalized by the corresponding area, deformations are referred to the initial dimensions, e.g. the elongation Δl to the initial length l_0. The obtained relation for the deformation is non-dimensional: Strain = Elongation, mm per initial length, mm; $\varepsilon = \Delta l/l_0$.

With respect to the direction of the induced forces the following types of mechanical loading are differentiated (and Table 1).

Table 1: Definitions of mechanical loading conditions.

Loading condition		Material response	
Tensile stress	σ	Dilatation	$\varepsilon = \Delta l/l_0$
Compressive stress	σ	Compression	$\varepsilon = -\Delta l/l_0$
Shear stress	τ	Shear strain	$\gamma = \Delta\gamma/h_0$
Torsional moment	M_T	Torque	$\Phi = \Delta u/r_0$
Bending moment	M_B	Deflection	$\Delta h/l_0$
Hydrostatic pressure	p	Compaction	$-\Delta V/V_0$

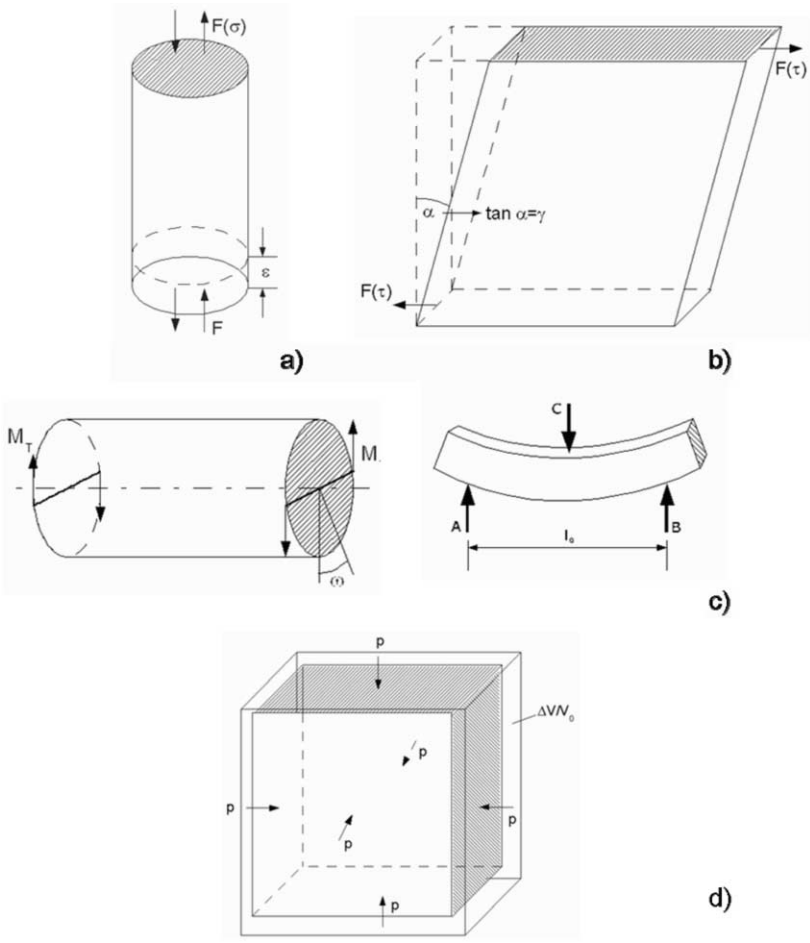


Figure 1: Systematic of mechanical loading classified according to the acting load vectors: axial stresses, a); shear stresses, b); bending and torsion loading, c); hydrostatic pressure, d).

2.1. Elastic Deformation

Under small loads, each solid body at first deforms in a purely elastic manner. This implicates that for subsequent unloading the deformation is completely reversible. In most technical cases a linear relationship between the acting stress and the corresponding deformation is observed. This constant of proportionality between stress and strain is defined as Young´s modulus (Equation 1), for planar shear stress and the corresponding shear strain as shear modulus (Equation 2) and for hydrostatic pressure and the resulting reduction in volume as compression modulus (Equation 3). As strain, shear and volume reduction have no dimension, the corresponding moduli have the dimensions of a stress and, analogous to Hooke´s law can be considered as the spring constant of the material.

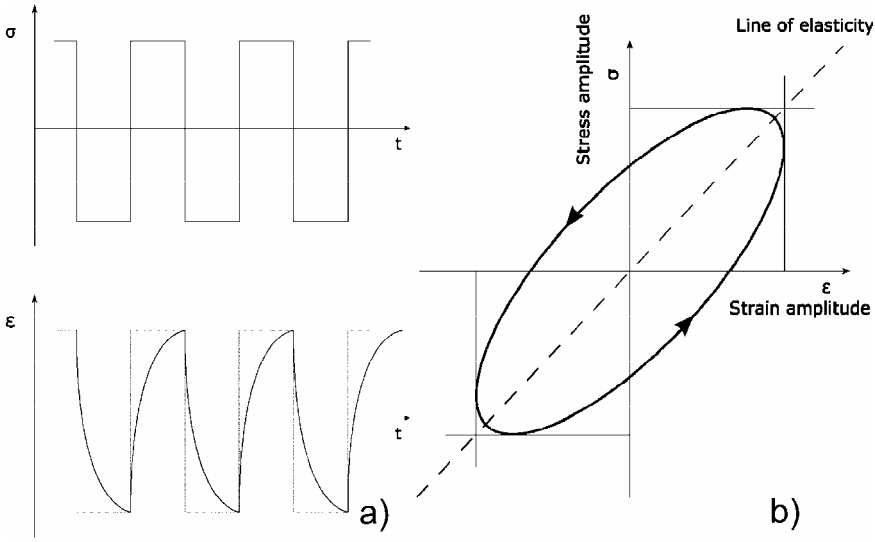


Figure 2: Inelastic deformation: rectangular stress application (load signal) and delayed deformation response ε, a) and time compensation of σ(t), ε (t) in terms of a stress-strain diagram, b).

$$\frac{\Delta\sigma}{\Delta\varepsilon_e} = E \tag{1}$$

$$\frac{\Delta\tau}{\Delta\gamma_e} = G \tag{2}$$

$$\frac{\Delta p}{-\Delta V / V_0} = K \tag{3}$$

In general, for crystalline materials, elastic elongation is directly related to a reduction in cross-section. (Just think of pulling on a rubber band.) This effect is defined as transverse contraction and expressed by the Poisson´s ratio $\nu = \varepsilon_{trans} / \varepsilon_{long}$. Totally, among the elastic constants, (E, G, ν) only two are independent, i.e. if two variables are known the third can be calculated, $E = 2(1+\nu)\,G$.

2.2. Viscoelasticity

Elastic materials which are deformed to a certain strain $\varepsilon = \sigma/E$ (comp. Equation 1) by a stress σ immediately recedes to $\varepsilon = 0$ after unloading. In many cases and also for biomaterials, deformation under small applied stresses can be regarded as elastic in

terms of reversibility, but is associated with a phase shift (i.e. retarded reaction) between loading and unloading. These cases where elastic deformation is retarded to a measurable extent are defined as viscoelastic deformations (Figure 2a). This delay in material response $\varepsilon(t)$ with respect to loading $\sigma(t)$ is also acting if the loading function $\sigma(t)$ is not rectangular but has a sinusoidal shape with the latter being of pronounced technical importance (as in- or decreases in load never happen infinitely fast). This phase shift between loading and deformation can be expressed in time-compensated stress-strain hysteresis (Figure 2b). For strictly elastic behaviour each stress value would corresponds to exactly one equivalent strain. In terms of viscoelastic material behaviour, for each stress, two strain values (one for loading and one for unloading) are equivalent. The inclination of this mechanical hysteresis corresponds to E and the width describes the phase shift between stress and strain.

2.3. Plastic Deformation and Fracture

Each deformation induced by external forces acting is in a first step basically elastic. With further increase in loading, changes in the material response occur. In this case, two basic phenomena are of pronounced importance: plastic flow and fracture.

a) Above a certain threshold stress σ_y the material deforms by plastic flow (yield stress, related strain $\varepsilon_y = \sigma_y/E$). This strain contribution is irreversible and characteristic for ductile materials, e.g. most biomaterials, most metals and some polymers (Figure 3a).
b) Exceeding a specific threshold stress σ_r, failure of the material occurs without any previous plastic deformation (rupture stress, corresponding strain $\varepsilon_r = \sigma_r/E$). This brittle behaviour is observed especially for glasses, ceramic materials and hardened metals (Figure 3a).

To characterise plastic behaviour experimentally the monotonic stress-strain test is suitable (Figure 1). For the analysis of extremely brittle materials (glasses, ceramic materials) bending experiments are more appropriate.

Plastic deformation can be generally expressed in terms of a complex function (Equation 4), which describes the material behaviour as a relation of five parameters: applied stress σ, deformation ε, time t, microstructure S and temperature T, respectively (*please do not worry about the complexity and the mathematical solution of this "cactus" equation! Things are easier than expected*):

$$\Psi = \Psi(\sigma, \varepsilon, t, S, T) \tag{4}$$

Unfortunately, a closed mathematical solution of this function has not been developed yet, and therefore, a differentiated procedure in experimental analysis is performed. The solution of Ψ is obtained in a way that two of the functional parameters are experimentally measured while the others are kept constant.

Mostly, the mechanical variables σ, ε and the time t in the sense of a deformation or strain rate $d\varepsilon/dt$ are of pronounced interest. Therefore, three basic static experiments are of superior interest:

In the monotonic stress-strain test, the stress is continuously increased and registered as a function of the resulting strain while the parameters time t in the sense of strain rate $d\varepsilon/dt$, S and T remain constant (Figure 3a). Measuring the deformation

ε versus time t under constant stress or load is defined as creep experiment (Figure 3b), and finally, constant deformation ε leads to a functional relation between stress σ and time t which is generally defined as stress relaxation test (Figure 3c).

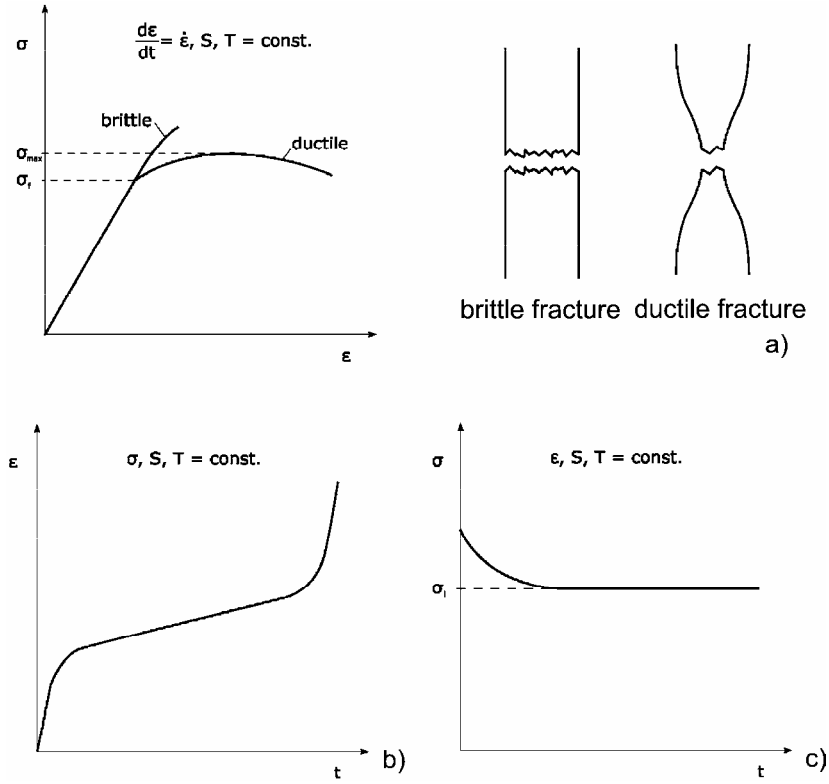


Figure 3: Basic types of results describing the mechanical behaviour under monotonic loading: Stress-strain curves for ductile and brittle materials and corresponding fracture morphology (schematic), a); creep experiment, b); stress relaxation test with time dependent stress decrease until a constant relaxed stress is reached (internal stress), c).

2.4. Summary

Elastic Material Behaviour

Purely elastic material behaviour can be generally defined by three basic characteristics:

- complete reversibility after load reduction
- linear relation between applied load and resulting deformation
- limitation to small deformations (i.e. ≤ 1 % for crystalline materials)

Viscoelastic Behaviour

The material response to loading or unloading is basically reversible, but deformation is retarded with respect to the actual load value.

- characterizes time dependent elastic deformation
- results in damping of externally induced oscillations
- strength and modulus depend on the strain rate at which the material is deformed

Plastic Deformation

- occurs after a material dependent threshold stress is exceeded
- is always combined with the irreversibility of deformation
- no linear relation between stress and deformation, characteristic line for the plasticity of materials is the stress-strain curve
- characteristics of the stress-strain curve are:
    - work hardening and thus more stress is required for further deformation
    - maximum defining the tensile strength
    - decrease until fracture

Ductile/Brittle Behaviour

- ductile materials exhibit measurable (visible) plastic deformations
- brittle material behaviour is defined for fracture without (practical) plastic deformation
- fracture occurs directly in the elastic region

Creep

As often (falsely) assumed, creep does not necessarily implicate deformation for very long times, but only basically describes plastic deformation under constant stresses.

- time dependent plastic deformation, usually under constant load or stress

3. Experimental Methods

Additional to the in the following reviewed standard experimental tests, other techniques have been developed to address specific research questions. For example, indentation tests are widely used to analyse surfaces or layers e.g. [2] and various multi-axial testing methods are used to characterise material properties e.g. [18]. But these methods are not within the scope of this paper.

3.1. Stress-Strain-Experiments

The information obtained from the standardized stress-strain experiment is of basic superior interest for the characterisation of materials under monotonic loading. Figure 4 shows a typical test rig for mechanical testing.

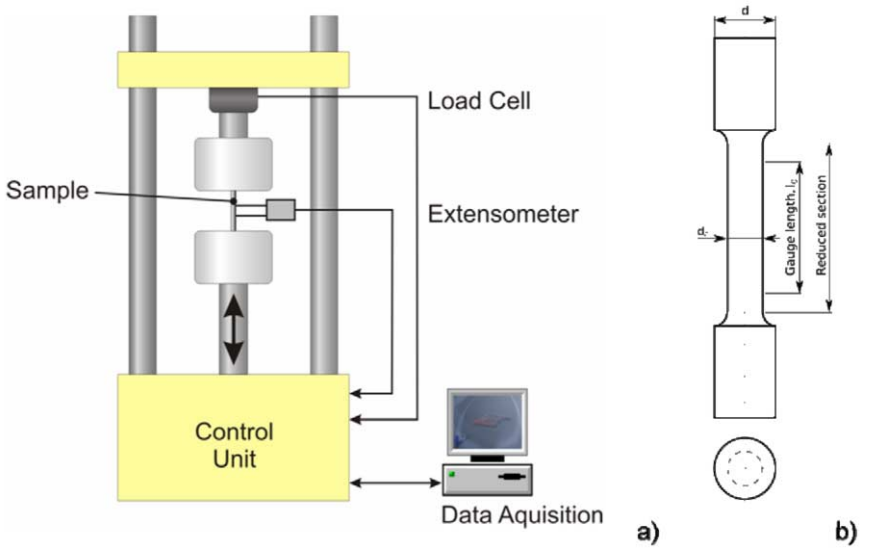


Figure 4: Schematic test rig for analysing mechanical properties, a) and standardized specimen geometry for tensile experiments, b).

In most cases tensile experiments are performed. With respect to the use of different specimen geometries and dimensions, the ratio of cross section to gauge length is specified to: $l_0 = 5.65 \sqrt{A_0}$ (Figure 4b). *(This value is used in traditional material science. Even so, it may not be possible to create arbitrary specimen geometries with biomaterial it is of high importance to use standardised sizes.)* The specimen is deformed under a constant extension rate $d\varepsilon/dt$ while the applied load F and the corresponding deformation Δl are continuously recorded and plotted in force vs. displacement curves. Normalizing the data by the initial specimen cross section A_0 and the initial length l_0, respectively, gives the geometry independent stress-strain curve $\sigma = \sigma(\varepsilon)$, e.g. Figure 3. This curve represents the characteristic deformation of the specific material. It depends not only on the chemical composition of the material, but also on microstructure (and processing conditions).

How to interpret stress-strain diagrams ? Starting from the origin, the first section of the curve corresponds to the linear increase in stress (slope E). The transition from elastic to plastic material behaviour is not clearly obvious in many cases, but is more or less continuous. Due to this, it is often advantageous to define an additional material parameter: The flow stress $R_{p0.2}$ (Yield Strength YS) is defined as the stress under which a plastic deformation $\varepsilon = 0.2$ % is obtained after complete unloading. Strain hardening is characterized by an increase in stress for further deformation (strain). This can be interpreted that for additional strain, the internal resistance against this deformation increases (work hardening, characteristic for the deformation of metals). Consequently, this hardening reaches a peak value R_m (Ultimate Tensile Strength, UTS). In the following, the force needed for further strain decreases and signalizes "that the end is near". The material loses strength until fracture occurs. Strictly spoken, it is impossible to load a material with loads exceeding R_m. It should also be mentioned that in the case of work hardening effects, the elastic regime of the material is also

prolonged. In a second load cycle the maximum stress/strain of the first load cycle may be reached.

The area below the stress-strain curve provides another important parameter. Mathematically, the area corresponds to the integral and is equivalent to the work required for the material deformation and thus, can be regarded as energy density, usually measured in J/m^2:

$$W = \int_0^{\varepsilon_f} \sigma \, d\varepsilon = \frac{1}{A_0 \cdot l_0} \int_{l_0}^{l} F \, dl \tag{5}$$

The practical sense of the integral, W, is the work required for a deformation process, e.g. in terms of the energy absorption, i.e. compensation of kinetic energy due to impact loading. Further characteristic parameters which can be obtained from the stress-strain curve are: uniform deformation, describing homogeneous elongation of a specimen under simultaneous reduction in cross section until R_m. After R_m necking is initiated, leading to a local reduction in cross section and all subsequent deformation is confined at this neck. Finally, fracture occurs at the neck.

Testing of certain biomaterials may not result in as nicely shaped curves. E.g. the tensile testing of cartilage results in a stress-strain curve which exhibits a smooth transition from initial deformation to the elastic regime with rather large deformations [11]. This transition zone is often referred to as the *toe region* and is mostly affected due to the un-crimping of collagen fibres and elasticity of elastin.

3.2. Creep

Creep is generally defined as time controlled plastic deformation at constant stress. The initial deformation rate $d\varepsilon/dt$ continuously decreases, although, the deformation is not stopped (it does not approach zero) and the material still deforms (creeps) with slow deformation rates. For constant stress, deformation ε follows a function of time t and for describing the material behaviour the stress-strain curve is replaced by a strain vs. time curve $\varepsilon(t)$ (creep curve, Figure 5).

Generally, creep can be characterized by three classic regions: Immediately after mechanical loading and elastic deformation equivalent to the magnitude of the applied stress $\varepsilon_{elsat} = \sigma/E$, a transition to plastic deformation is observed (primary creep). During the second stage a constant, stationary creep rate is acting. In the third stage creep rates increase again and failure is initiated (Figure 5). Secondary creep is characterized by a constant microscopic structure of the material. This steady state can be regarded as dynamic equilibrium: While the material reacts on the applied load by continuous work hardening (i.e. $(d\sigma/d\varepsilon)d\varepsilon > 0$) competitive softening effects are acting $(d\sigma/d\varepsilon)d\varepsilon < 0$).

Steady state creep is therefore observed, if the increase in deformation resistance is compensated by softening effects of the material (relaxation). After sufficient deformation (tertiary creep), an acceleration in creep rate is observed until fracture occurs. This increase in creep is substantially driven by rate controlling damaging mechanisms which already take place during secondary creep but become more pronounced in the tertiary creep regime, indicating that the end is near. Creep fracture occurs at sufficient high strains which only slightly depend on loading and temperature.

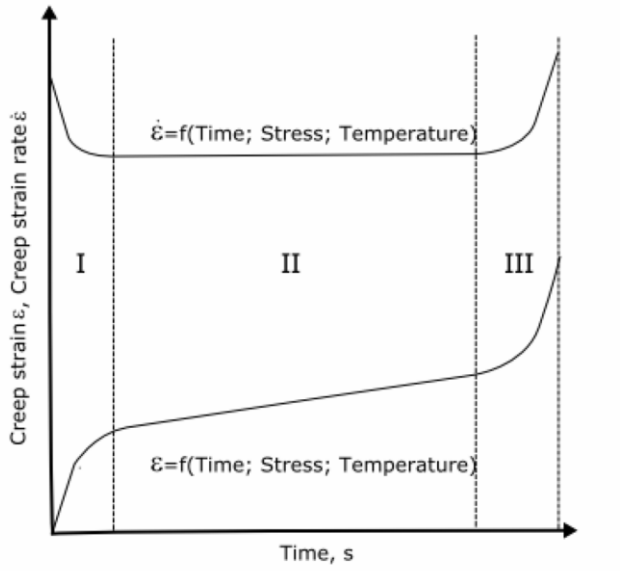


Figure 5: Typical creep curve for loading under constant stress: Strain vs. time, a) and differentiated form strain rate vs. strain.

3.3. Fatigue Testing

Strength is generally defined as deformation resistance against monotonic (static) loads. A further and practically more relevant load condition is the material behaviour under loads varying with time (cyclic deformation). This effect of fatigue is of pronounced importance as many materials or structures exhibit significant reductions in load bearing capability for stresses far below the monotonic strength when exposed to cyclic loading. Fatigue failure is always brittle-like in nature even for ductile materials. It is catastrophic and insidious, occurring very suddenly and without warning. The process of fatigue failure is driven by the initiation and propagation of short cracks and ordinarily the fracture surface is perpendicular to the direction of the loading axis.

Cyclic Stresses

The applied stress may be axial (tension-compression), flexural (bending) or torsional (twisting) in nature. The material behaviour is analysed in most cases for sinusoidal loading.

This is done either in force or in deformation control while the number of load cycles until fracture is counted. In addition to the oscillating stress amplitude, a constant static stress can be superimposed. With respect to this mean stress, different forms of fatigue loading are defined. For example tensile fatigue loading is characterised by the peak and valley loads being both tensile. In the range of stresses where the maximum load is tensile while the minimum load is compressive, the ratio between peak and valley load is defined by the stress ratio $R = \sigma_u/\sigma_o$.

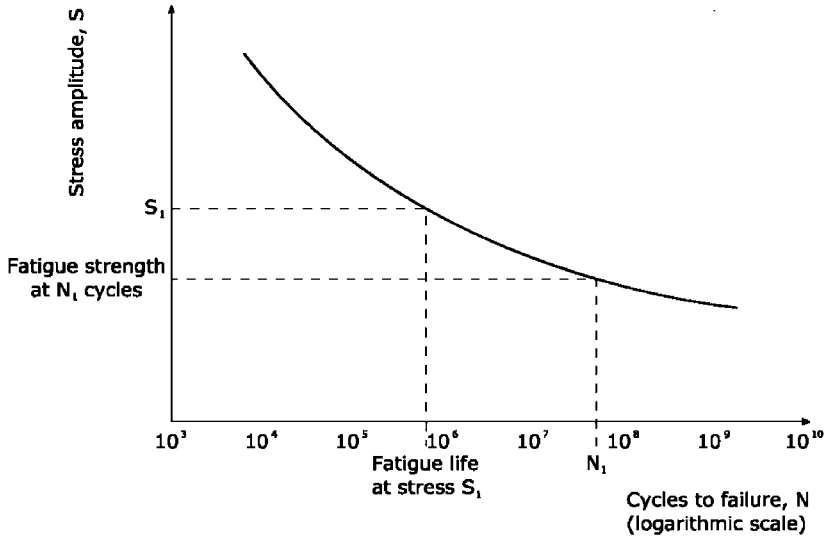


Figure 6: Schematic S-N curve and determination of fatigue life and fatigue strength.

The S-N Curve

For specifically defined variations in the load amplitudes ($\Delta\sigma$) applied in experiments the number of cycles (*N)* is counted until fracture occurs and the results are typically plotted in a semi-logarithmic *S-N* curve (Figure 6). If stresses are sufficiently low, a horizontal section is observed for which the number of cycles is no longer a function of the applied stress. This threshold is defined as endurance limit S_e. This means that for stress amplitudes below S_e no fracture will occur. It should be noted that the *S-N* curve is a purely statistical average indicating that 50 % of the experiments exceed the number of cycles, but 50 % fail at lower cycle numbers, if not specified differently. Generally, two regions of the *S-N* curve are defined: In cases of small stresses and high cycle numbers (usually in the order of 10^6 or more) high cycle fatigue (HCF) is dominating, whereas for stresses in the order of the Yield Strength and load cycles between 10^2 to 10^5 , low cycle fatigue (LCF) is present.

Failure is not always as well defined as in tension fatigue test where the specimen finally ruptures. For example in compression fatigue tests, a failure criterion has to be defined. This criterion could be, for example, a certain reduction in the material stiffness compared to the initial, unharmed stiffness.

Crack initiation and Propagation

Fatigue failure is characterized by three distinct phases: Crack initiation (phase I), wherein microscopically small cracks form (usually at the surface) at points of high stress concentration; crack propagation (phase II) where small cracks grow incrementally with each applied stress cycle (normally preferred during the tensile phase of the load cycle) and failure (phase III), initiated very rapidly as soon as one crack has reached a critical size. It should be emphasised that all geometries which cause stress concentrations can be regarded as sites of crack nucleation, i.e. threads, sharp corners, drill holes, surface roughness or scratches and (micro-) porosity.

4. In-Vitro Characterisation

In order to transfer the experimental procedure for the characterisation of materials to the in vitro characterisation of biomaterials, specific emphasis has to be focussed on the related boundary conditions. This aspect is of utmost importance in order to derive reliable and comparable results in the sense of a clear scientific interpretation and comparison to data in the literature.

4.1. Boundary Conditions

In Table 2, a rough summary of relevant boundary conditions which can significantly influence the results of in vitro experiments is demonstrated. For practical experimental analysis and with respect to the variety of influencing parameters, well defined planning is essential for obtaining clear reliable results, from which further precise conclusions can be made, concerning the transfer of the experimental data to in vivo tissue / material behaviour or interaction. Referring to Equation 4, this implicates that the experiment should be kept as simple as possible and as few parameters as necessary should be varied. Furthermore, the experimental set up must be carefully assembled and optimized in terms of the required accuracy. Thus, danger of the occurrence of artefacts and a misinterpretation of the results can be minimized. Although, in many cases some parameters of interest, for example, the specimen structure, can only be determined with a large experimental effort, it is always better to follow this way than to let parameters unclear and interpret uncertain results by means of statistical methods. Strictly speaking, for experimental in vitro analysis, all required parameters should be defined and controlled during the tests. This guarantees precise and transferable conclusions and finally will lead to minimized deviations in the concluding statistical view which is indispensable in any case.

4.2. Characterisation of the microstructure

Almost all biomaterials exhibit a highly inhomogeneous and anisotropic microstructure. Therefore, the material behaviour will be highly dependent on the orientation of the microstructure with respect to the acting load. This implies the necessity of examining the structure and careful alignment of the specimen axis and the load axis [8]. The analysis of the microstructure is often accomplished with image based procedures. Depending on the scale level, these methods could be light microscopy, x-ray quantitative computed tomography, micro-magnetic resonance imaging or high resolution micro-CT. Numerous researchers studied the structure of materials with these methods and derived measures for quantification e.g. [22], [1], [3].

4.3. Example of In Vitro Behaviour

In order to provide an example of the requirements described above, in vitro experiments on cancellous bone will be demonstrated and discussed in the following section. As physiological loading of cancellous bone structures is mostly compressive, the experiments reported are also performed in compression.

The response of a bovine cancellous bone specimen exposed to monotonic compressive loading is shown in Figure 7. Typical for a cellular solid [14], the monotonic loaded cancellous bone specimens' behaviour was found to be linear (in

stress-strain space) at the beginning of the compression. The linear regime is followed by a decreasing slope, indicating failure and plastic collapse, and an increase (densification) at higher strain values. While the magnitude of the maximum compressive stress varies up to tenfold between the groups, the characteristics of the deformation behaviour are similar. Even if the initial deformation behaviour appears to be rather linear, non-linearities can be found by analysing the deformations with a higher local resolution. For this reason optical surface deformation measurements are applied. The surface deformations of the specimen reveal these inhomogeneous components. Localised strain concentrations can be found already at a small percentage of the maximum stress. These strain concentrations do further localise and increase in value until failure occurs in the same region. Macroscopic damage appears in form of a slip line across the whole specimen cross section.

Table 2: Important boundary conditions to be considered for in vitro testing.

Type of Boundary Condition			
Physiological	**Biological**	**Mechanical**	**Analytical / Preparation**
Temperature	Biological scatter: (Micro-) structure	Preload	Definition of experiment and limitations
Isotonic / wet / dry	Human / animal	Strain rate	Specimen fixing
Strain rate	Male / female	Load application	Integral / differential strain measurement
Loading state	Age	Transient effects / artefacts	Sample rate
	Cell vitality	Fixing quality	Preparation/pre-damage
	Orientation	Specimen dimension	Pre-conditioning
	Test with / without surrounding tissue	Set up stiffness	Initial state characterisation (density, morphology, image resolution ..)
		Stress state	Signal drift
		Specimen alignment	
		Friction effects	

The early appearance of these strain localisations indicates that structural damage is already induced at rather low stress levels, even at the beginning of the elastic regime.

Figure 7 shows also data for other types of cancellous bone, human vertebral bone cored in the main physiological axis and human vertebral bone cored perpendicular to the main axis. A remarkably different behaviour is obvious. An explanation of these differences can be found by looking at the underlying structure. Firstly, there is a large difference in density between bovine and human specimens. Furthermore, the bovine specimens contain a combination of plate and rod like trabeculae (Figure 8), whereas the human bone is strictly composed of rods (Figure 9). A rather large influence of the orientation of the main material axis with respect to the applied load vector is also

observed. Stiffness and strength decrease enormously by changing the load axis from 0 degree (perfectly aligned with the main material axis) to 90 degree (perpendicular). This difference can be explained with the high anisotropy of the trabecular network. Furthermore, the influence of non-optimal boundary conditions is shown in Figure 7. The dotted line shows a distinctive toe region, but in contrast to tests on cartilage where this behaviour is a result of the underlying microstructure, here it is an experimental artefact due to early trabecula failure on the boundaries and small misalignments. In all the other shown curves, the specimens were embedded on both sides in order to achieve a homogeneous load transfer from the test rig to the specimen.

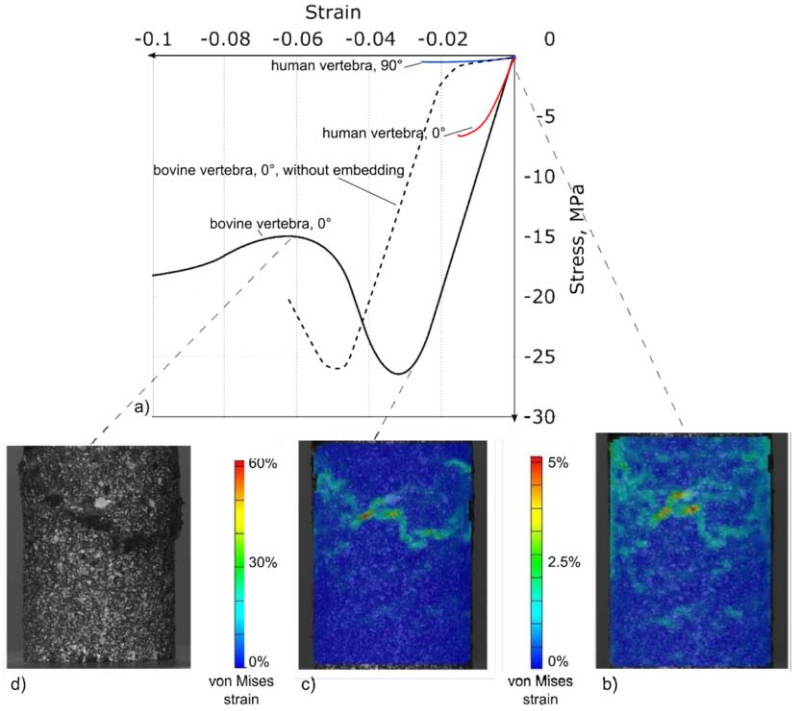


Figure 7: Stress-strain curves for different types of cancellous bone, a). The dotted line shows a specimen without embedding at the ends and therefore experimental artifacts in the beginning of the test. Von-Mises surface strains (analyzed with an optical deformation measurement system) are shown at different stages during the experiment, b)-d). The angle value refers to the deviation between load vector and main physiological axis. [7]

Only taking into account all the experimental and specific material-structure properties results in comparable, reproducible and valuable data.

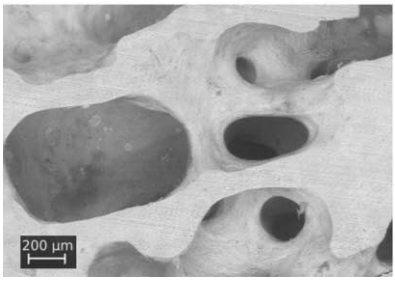


Figure 8: Bovine vertebral cancellous bone, main physiological axis in vertical direction. [7]

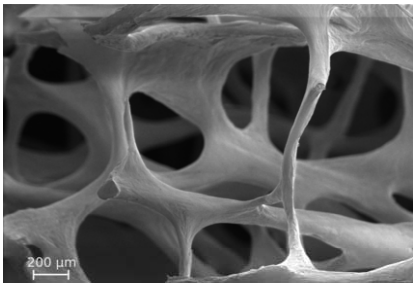


Figure 9: Human vertebral cancellous bone, main physiological axis in vertical direction. [7]

5. State of the art and future directions

Nowadays, image based methods are increasingly used to characterise the materials microstructure. For example, micro-CT based methods set a new standard in the quantitative evaluation of the structure. These techniques provide high resolution images of the structures and image processing methods can create a 3D computational model from the data. An active research field is the derivation of quantitative measures for both, the description of the structure but also for the linkage between structure and mechanical properties [25], [5]. Furthermore, the damage and deformation behaviour of biomaterials can be studied using high resolution scans which results in a much deeper understanding of the ongoing processes. Imaging in combination with numerical methods like Finite Element Analysis and additional data from mechanical testing has proven to be a very promising way forward. [21], [17].

With ongoing research on the microstructures of the materials a basis can be set to improve the relationships, originally derived for engineering materials to the specific need and challenges of biomaterials [23]. Recently, novel methods in musculoskeletal simulations allow insight in the in-vivo forces acting on many biomaterials [6]. This information can be used to adjust experimental conditions to approach the relevant in-vivo scenario.

Besides the technological advances, a standardisation of load cases and experimental conditions will be a challenge for the future. The need for an improved comparability of data from various sources as well as the increasingly complex man-made biomaterials has to be drawn to standards, comparable to engineering design codes.

Acknowledgements

The authors would like to thank the staff at the Laboratory for Materials Science at the University of Applied Sciences Regensburg. Special thanks are due to R. Mai for technical and experimental support, to N. Fischer for preparing the images and to H.-P. Bräu for providing the electron microscopic images.

References

[1] O. Beuf, D. C. Newitt, L. Mosekilde, and S. Majumdar. Trabecular structure assessment in lumbar vertebrae specimens using quantitative magnetic resonance imaging and relationship with mechanical competence. *J Bone Miner Res*, 16(8):1511–9, 2001.

[2] C. Yanping, M. Duancheng, and D. Raabe. The use of flat punch indentation to determine the viscoelastic properties in the time and frequency domains of a soft layer bonded to a rigid substrate. *Acta Biomater*, 5(1):240–248, Jan 2009.

[3] D. Chappard, P. Guggenbuhl, E. Legrand, M. F. Basle, and M. Audran. Texture analysis of X-ray radiographs is correlated with bone histomorphometry. *J Bone Miner Metab*, 23(1):24–9, 2005.

[4] S.C. Cowin. Bone mechanics handbook. *CRC Press Inc., Boca Raton*, 2001.

[5] S.C. Cowin. The relationship between the elasticity tensor and the fabric tensor. *Mechanics of Materials*, 4(2):137–147, 1985.

[6] M. Damsgaard, J. Rasmussen, S. Tørholm, E. Surma, and M. deZee. Analysis of musculoskeletal systems in the anybody modeling system. *Simulation Modelling Practice and Theory.*, 14:1100–1111, 2006.

[7] S. Dendorfer. Cyclic deformation and fatigue behaviour in cancellous bone. *PhD thesis,* University Paderborn, 2008.

[8] S. Dendorfer, H. J. Maier, D. Taylor, and J. Hammer. Anisotropy of the fatigue behaviour of cancellous bone. *J Biomech*, 41(3):636–641, 2008.

[9] S. Dendorfer, H. J Maier, and J. Hammer. How do anisotropy and age affect fatigue and damage in cancellous bone? *Stud Health Technol Inform*, 133:68–74, 2008.

[10] [edited by] Mow, Van C. *Basic orthopaedic biomechanics*. Lippincott Williams and Wilkins, Philadelphia, 2nd edition, 1997.

[11] J. Fierlbeck, J. Hammer, C. Englert, and R. L. Reuben. Biomechanical properties of articular cartilage as a standard for biologically integrated interfaces. *Technol Health Care*, 14(6):541–547, 2006.

[12] Y. C. Fung. *Biomechanics: motion, flow, stress, and growth*. Springer-Verlag, New-York, 1990.

[13] Y. C. Fung and Pin Tong. *Classical and computational solid mechanics.* World Scientific, 2001.

[14] L. J. Gibson and M. F. Ashby. Cellular solids: structure & properties. *Cambridge University Press, Cambridge*, 2nd, 1997.

[15] R.W. Hertzberg. Deformation and fracture mechanics of engineering materials. *John Wiley and Sons, Inc.*, 1995.

[16] R. B. Martin, D. B. Burr, and N. A. Sharkey. *Skeletal Tissue Mechanics*. Springer Verlag, New York, 1998.

[17] A. Nazarian, J. Muller, D. Zurakowski, R. Mueller, and B. D. Snyder. Densitometric, morphometric and mechanical distributions in the human proximal femur. *J Biomech*, Jan 2007.

[18] G. L. Niebur, M. J. Feldstein, and T. M. Keaveny. Biaxial failure behavior of bovine tibial trabecular bone. *J Biomech Eng*, 124(6):699–705, 2002.

[19] M. Nordin, Oe. Nihat. Fundamentals of Biomechanics. *Springer-Verlag New York, Inc.*, 1999.

[20] S. Suresh. *Fatigue of Materials*. University Press, Cambridge, 2 edition, 1998.

[21] P. J. Thurner, P. Wyss, R. Voide, M. Stauber, M. Stampanoni, U. Sennhauser, and R. Mueller. Time-lapsed investigation of three-dimensional failure and damage accumulation in trabecular bone using synchrotron light. *Bone*, 39(2):289–299, Aug 2006.

[22] W. J. Whitehouse. The quantitative morphology of anisotropic trabecular bone. *Journal of Microscopy*, 101(2):153–168, 1974.

[23] G.M Williams, K. R Gratz, and R. L Sah. Asymmetrical strain distributions and neutral axis location of cartilage in flexure. *J Biomech*, Dec 2008.

[24] H. Yamada. *Strength of biological material*. Williams & Wilkins, Baltimore, 1970.

[25] P. K. Zysset. A review of morphology-elasticity relationships in human trabecular bone: theories and experiments. *J Biomech*, 36(10):1469–1485, 2003.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-247

Introduction to Chapter V: Medical Imaging

Peter NIEDERER and T. Clive LEE (eds.)

According to a German proverb, a picture says more than a thousand words. This statement is certainly true in clinical diagnostics. With the advent of the X-ray projection technique, somewhat more that 100 years ago, it became for the first time feasible to obtain insight into the human body without opening it. Since then, various methods of medical imaging have developed into major tools in clinical diagnosis which cover an enormously wide field of pathologic situations. The various modalities are still in partly spectacular development.

Ultrasound, X-ray and magnetic spin resonance are the most often applied modalities in medical imaging performed on the intact human body. The basic physics underlying the procedures derived thereof are well known and established, and they are outlined in the next three sections. The aim is such that the reader is in a position to understand and appreciate present and novel developments and clinical applications for which there exists a vast literature.

Mathematicl methods, in particular Fourier analysis, along with elementary physics are essential prerequisites in the understanding of the methods under consideration in the following. An attempt has nevertheless been made in agreement with the goals of this book such that the basics can be understood without a thorough mathematical background. The relevant mathematics are given in appendices.

There are a number of further imaging modalities with an increasing importance in medical imaging, viz, Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT) and methods based on the application of infrared, to mention the most important. In part, similar (in particular mathematical) methods are applied in these technologies. Since this book is limited to the treatment of the basic elements of the most important imaging methods, these procedures are not included.

A further important area in medical imaging is devoted to image analysis with the aid of the computer. At least some numerical treatment is necessary to present the results of measurements made with ultrasound, X-ray, or magnetic spin resonance on a computer-driven screen in a clinically useful form. However, medical image analysis, regardless whether it is performed automatically or with human intervention, is a wide field in its own right and is not approached here.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-249

V.1. Ultrasound Imaging and Doppler Flow Velocity Measurement

Peter F. NIEDERER
Institute for Biomedical Engineering
ETH Zurich and University of Zurich
Gloriastrasse 35
CH-8092 Zurich, Switzerland

Abstract. High frequency ultrasound (2 – 8 MHz typically) has established itself as a major medical imaging method associated with a wide range of clinical applications. Advantages include real-time applicability, lower cost compared with other medical imaging technologies, possibility of measuring blood flow velocities and desk-top instrumentation. Disadvantage is associated with lower image quality than is obtained with x-ray or magnetic resonance methods.

Keywords. Ultrasound, medical imaging, Doppler flow measurement, color Doppler

Introduction

Clinical applications of ultrasound include diagnostic as well as therapeutic procedures. From a technical point of view, the difference between diagnostic and therapeutic applications firstly consists of the amount of energy thereby delivered to the tissue: In case of diagnostic ultrasound, the power density at the body surface is limited to 100mW/cm^2, except for the eye where a lower limit of 20 mW/cm^2 is set. Below these limits, any effect on the tissue can be excluded with sufficient certainty. Therapeutic ultrasound, in contrast, is intended to induce a tissue reaction, mostly in the form of heating and increase of blood perfusion, therefore irradiated energies are higher. Secondly, ultrasound therapy is essentially limited to the delivery of energy, whereas in diagnostic applications the analysis of the backscattered signal ("echo") is to the fore.

Ultrasound frequency bands used extend from kHz (mostly therapeutic ultrasound) to MHz (imaging and Doppler technique) and GHz (ultrasound microscopy). The wavelengths accordingly are in the cm to sub-µm range (corresponding to the conditions prevailing in water; wave speed about 1500 m/sec).

The first technical applications of ultrasound were made during the first World War where SONAR methods (Sound Navigation and Ranging) were developed in order to detect submarines (the French physicist Paul Langevin built the first ultrasound oscillator in 1914). A breakthrough was however achieved only after electrical short-pulse techniques became available. During the twenties, material testing methods based on ultrasound were introduced (S. Sokolov, F. Firestone and others).

The first use of ultrasound for medical imaging was made by the Austrian neurologist Karl Dussek together with his brother Friedrich Dussek who presented their

results in 1938 (Figure 1). Doppler methods for the measurement of blood flow velocities were finally developed by the Japanese physicist S. Satomura in 1957.

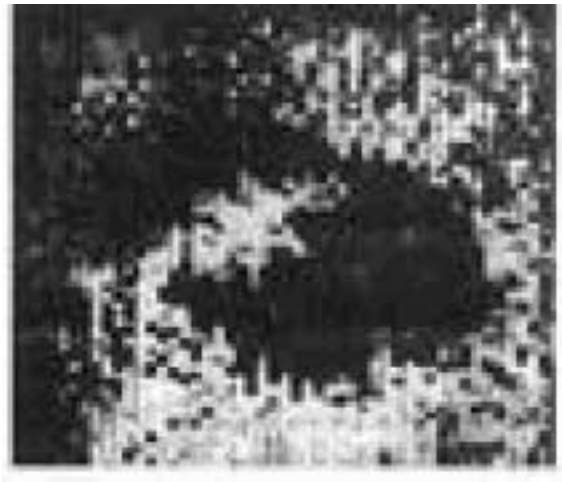


Figure 1. "Hyperphonogram" of the brain, Dussek 1938

This primer is devoted to ultrasound imaging and Doppler methods. In both techniques an electromechanical transducer is positioned on the body surface from which short ultrasound pulses in the form of a wave package with a defined center frequency are propagated into the tissue. The backscattered echo signal is then recorded by the same transducer in order to synthesize a cross section of the body part under consideration and/or to investigate the local perfusion of the tissue by way of an analysis of the Doppler-shifted frequency components of the backscattered signal. As soon as all the echoes generated by the pulse in the tissue and backscattered within the predetermined range of the instrument have reached the transducer, the next pulse is emitted, etc.

In comparison with other medical imaging modalities such as computed tomography or magnetic resonance imaging, ultrasound has three major advantages, viz.,

- *Real-time operation*
- *Inexpensive*
- *Availability of small and light-weight instruments which can easily be moved around in a clinic*

A disadvantage derives from the generally limited image quality, however.

Ultrasound waves as used in medical imaging consist in essence of compression waves; other types of waves such as shear or surface waves are of minor importance and can be neglected. The propagation of such waves in an inhomogeneous biological medium is however highly complex. Concurrent coherent and incoherent scattering processes along with strong damping cause the generation of complicated echo signals from which it is difficult to extract the signal components of interest. Mirror-like

reflections occurring at layered structures or speckle formation are therefore among the major difficulties.

Typical pressure amplitudes associated with diagnostic ultrasound applications in the MHz range are of the order of a few MPa. Such pressures are sufficiently high to induce nonlinear scattering processes involving, e.g., mode conversion. This phenomenon can be used for image enhancement which is usually denoted as 2^{nd} harmonic imaging.

Bony or air-filled structures are, for practical purposes, non-transparent for diagnostic ultrasound waves because of almost complete reflection (an exception is the transcranial Doppler method). Accordingly, one cannot "see" behind bones, the lung or air bubbles in the intestine. Due to the inhomogeneous composition of typical biological materials, furthermore, the propagation of ultrasound waves is not directed along exactly straight lines (such as x-rays). For this reason, amongst others, all attempts to develop ultrasound CT imaging have not been successful to date.

According to a rule of thumb, diagnostic ultrasound in the MHz range is damped by about 1 db per MHz and cm penetration depth (Figure 2). In applications such as cardiology where distances up to some 25 cm should be covered to create a useful image, frequencies higher than about 8 MHz cannot be used because of signal/noise limitation. The resolution, in turn, should at least reach around 1 mm for practical applications such that frequencies below 2 MHz are not useful because the resolution is essentially given by the wavelength which increases with decreasing frequency. In applications requiring a short penetration distance only (eye, skin), frequencies up to 20 MHz can be used. Even higher frequencies are applied in ultrasound microscopy, where GHz waves with a wavelength below 1 μm are utilized. Penetration depth is however limited to a few tens of μm because of damping.

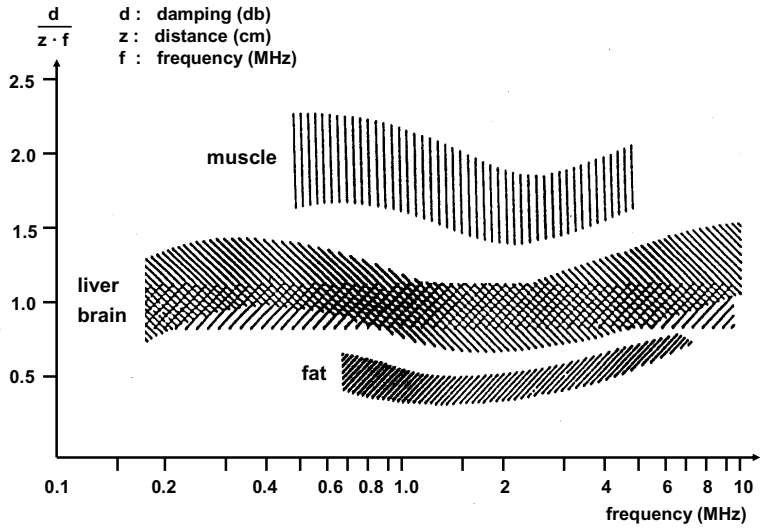


Figure 2. Damping of ultrasound waves in selected biological tissues

From the point of view of signal generation and analysis, medical ultrasound technology has much in common with RADAR (Radio Detection and Ranging) used in general aviation; major differences derive from the fact that RADAR waves are electromagnetic waves and operate in the GHz range.

1. Ultrasound Transducers

Ultrasound waves are mechanical by nature whereas the preparation of suitable pulses, as well as the signal analysis and representation of measured results, is entirely based on electronics. Accordingly, electromechanical transducers are necessary for the generation of ultrasound waves and for the recording of echoes. Piezoelectricity provides the most useful tool for this purpose: When certain crystals are exposed to an electric tension, they deform because of rearrangement of electric charges in the interior; in turn, upon deformation, charges are produced at the surface. This effect is denoted as piezoelectricity, and it can be applied because of its reciprocity for emitting as well as for receiving ultrasound waves.

A piezoelectric ultrasound transducer consists of an arrangement where one or more small bars or plates made of piezoelectric material are placed on a rigid, electrically inactive backing material mounted in a hand-held application device. Both surfaces of the piezo-elements are electrically connected, much as an electric condenser, to an oscillator for transmission and switched to a receiver for recording echoes. While the oscillator sets the transducer into vibration such that ultrasound waves (mostly in the form of short pulses) are emitted, the returning echoes deform the piezo-material passively whereby electric signals are generated that can be recorded. Typical materials used for medical ultrasound transducers are based on sintered lead zirconate titanate (PZT) where the piezoelectric effect is particularly strong.

1.1. Generation of Ultrasound Waves, Beam Characteristics of Transducers, Phased Array Technology

In order to analyse the characteristics of wave fields produced by oscillating transducer surfaces theoretically, Huygens' principle (see Appendix 2) is most useful[1]. According to this principle, a wave field can be synthesized from a continuous distribution of a dense set of vibrating point sources located on a surface enclosing the area of interest. Each point is thereby the origin of a spherical wave.

As an example, we consider the wave field produced by a circular transducer of radius a which is set into vibration in a thickness mode at a given frequency (Figure 3). The transducer is assumed to be located in a uniform fluid-filled cartesian space (coordinates x, y, z).

Only the field along the axis of symmetry, x, can be calculated analytically; off-axis values have to be determined numerically (see later). As outlined in the introduction and worked out in Appendix 1, compression waves in the fluid are only considered. For the solution of such wave propagation problems, it is advantageous to

[1] The principle is valid for linear problems only. The nonlinearities mentioned in the introduction, however, are of importance only in the interaction of the waves with scattering centers and not in the generation and propagation of the wave field itself.

introduce a scalar velocity potential, $\Psi(x,y,z,t)$, from which the velocity field $\vec{v}(x,y,z,t)$ to be calculated is obtained as the gradient (t denotes the time)

$$\vec{v}(x,y,z,t) = -\vec{\nabla}\Psi(x,y,z,t) \tag{1}$$

.

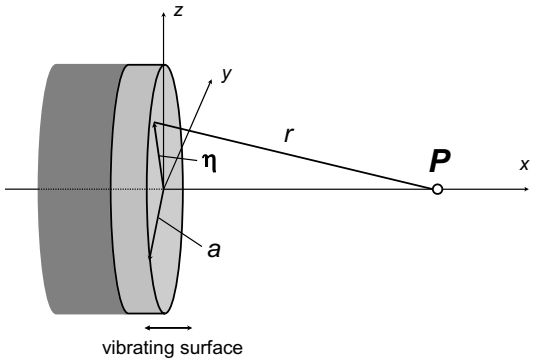


Figure 3. *Vibrating disk on a rigid backing support. The frontal surface is assumed to vibrate uniformly with a constant amplitude while the one in the back is at rest.*

According to Huygens' principle, the velocity potential has to be prescribed over a closed surface containing the space of interest. This surface includes the transducer surface along with the y - z plane and an enclosure at infinity over the right half space. Since only the transducer surface vibrates, $\Psi = 0$ outside. The transducer furthermore vibrates sinusoidally in $x-$ direction with a constant amplitude such that only one component of the gradient, $\vec{\nabla}\Psi$, namely $\partial\Psi/\partial x$ is $\neq 0$. We therefore have as boundary condition

$$\begin{aligned} \frac{\partial\Psi}{\partial x} &= v_0\, e^{i\omega t} \qquad for \quad \eta \le a \\ &= 0 \qquad\qquad outside \end{aligned} \tag{2}$$

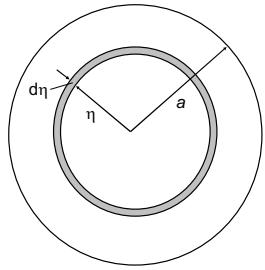
whereby v_0 denotes the (constant) amplitude of the vibration,
$\omega = 2\pi f$ the circular frequency (frequency f),
η the radial coordinate on the transducer surface.
(Since linearity is assumed, a complex notation can be used).

The velocity potential at an arbitrary point ***P*** located on the $x-$ axis is now obtained according to Huygens' principle as the integral of a continuous distribution of point sources over the transducer surface, all of them producing the same spherical wave $w(r,t) = \frac{v_0}{r} e^{-i(kr-\omega t)}$ with r denoting the distance between the point

source and **P** ($r^2 = \eta^2 + x^2$) and k the wave number, $k = \omega/c = 2\pi/\lambda$ (wave speed c, wave length λ). (The amplitude of a spherical wave is proportional to the square-root of its energy per unit area, therefore, it is proportional to $1/r$.)

The integral extends over the surface with radius a (single integration due to symmetry with the circular surface element $2\pi\eta d\eta$ and the factor 2π according to Appendix 2)

$$\Psi_P = \frac{v_0\, e^{i\omega t}}{2\pi} \int_0^a \left[\frac{1}{r} e^{-ikr}\right] 2\pi\eta\, d\eta \qquad (3)$$



After integration the relative intensity (energy Density I at point **P** can be calculated from the real part of Ψ_P squared. One arrives at

$$I \propto sin^2\left[\frac{\pi}{\lambda}\left(\sqrt{a^2 + x^2} - x\right)\right] \qquad (4)$$

Figure 4 exhibits this result qualitatively. It is important to note the typical near-field and far-field characteristics; from a practical point of view, the useful area of such a transducer is at a distance of a^2/λ. This aspect will further be discussed below.

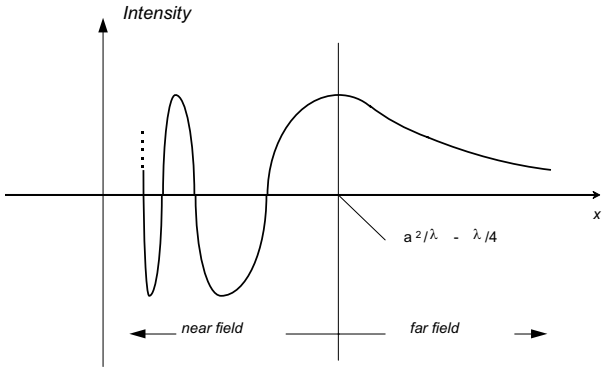


Figure 4. Near field and far field characteristics of a circular ultrasound transducer. The intensity varies rapidly close to the transducer.

The off-axis intensity can be determined by numerical integration (Figure 5; also analytical approximations have been presented in the literature). It is found that besides a main lobe, several side lobes exist. The larger the diameter of the transducer is in relation to the ultrasound wavelength, the narrower is the beam, furthermore, the number of side lobes increases.

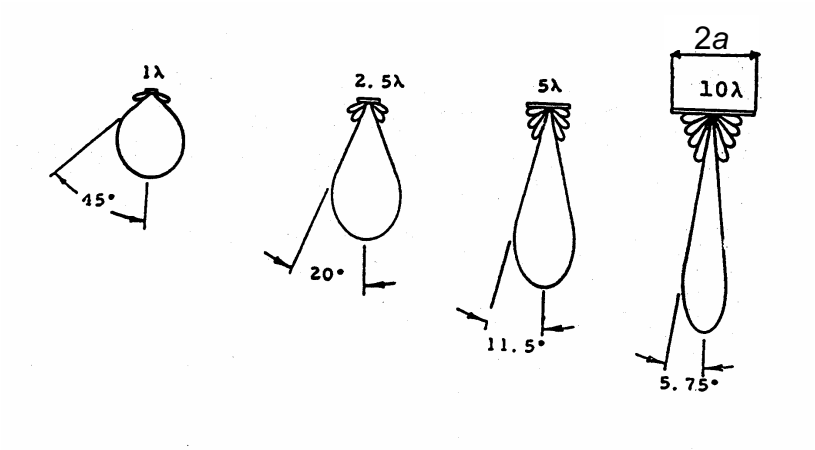


Figure 5. Main and side lobes of a circular ultrasound transducer. Note the influence of the a/λ ratio on the beam characteristics.

Note: Since linearity is assumed, these results hold for ultrasound propagation in continuous wave or pulsed mode. As we shall see, most medical ultrasound applications use pulsed mode techniques however; the similarity with RADAR has been mentioned before. A typical ultrasound pulse as used in medical applications is shown in Figure 6. (Schlieren imaging method).

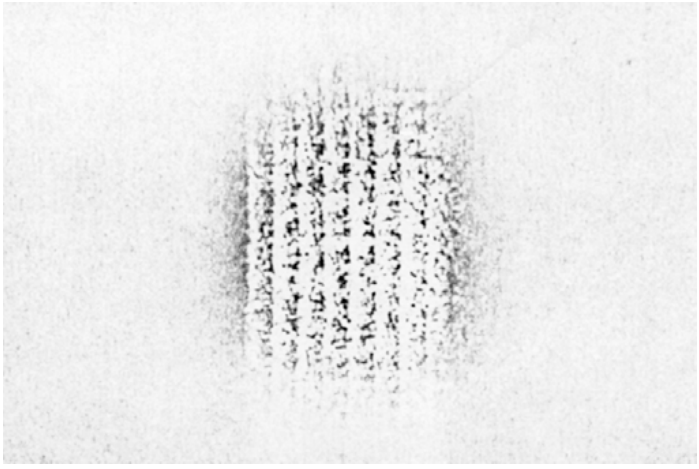


Figure 6. Schlieren image of a 5 MHz ultrasound pulse consisting of a package of 8 sinusoidal waves, wavelength 0.3 mm.

One of the methods adopted from RADAR is the "phased array" technique which is in widespread usage in pulsed ultrasound devices. Thereby, the Huygens' principle is approximated by an array of small transducers which produce each a quasi-spherical wave package having the envelope of a short pulse. The timing of each pulse along with the amplitude and phase of the center frequency are controlled such that, upon superposition, a combined wave forming a main lobe with a defined spatial orientation results (Figure 7). The problem how the amplitudes and phases of the individual transducer elements of a phased array have to be controlled in order to obtain a predefined far field is addressed in Appendix 3.

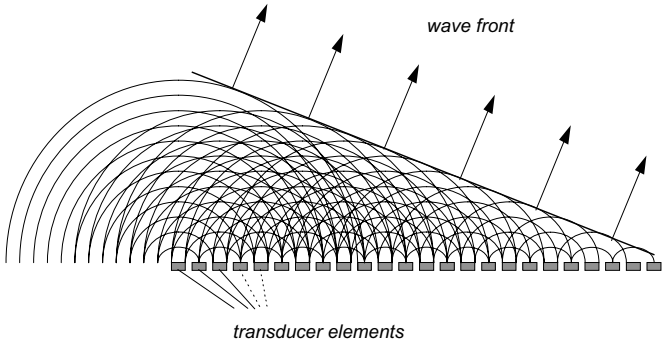


Figure 7. Phased array technique. Each individual transducer element is excited independently with a pulse whose timing is such that by superposition a beam with a defined direction of propagation is formed. By dynamic variation with time, e.g., an oscillating beam can be produced which scans an area of interest much like a windshield wiper. This technique is used for producing real-time imaging of selected cross-sections of the body. Most phased arrays presently in use are linear, one-dimensional arrangements; two-dimensional arrays which allow scanning of a volume can also be made, however.

The more transducer elements are used, the better a desired wave form can be obtained. In turn, transducer elements have to be made smaller when more elements are implemented which may cause mechanical sensitivity and electrical impedance matching problems. Furthermore, individual transducer elements cannot be positioned too close to each other because of mechanical crosstalk. If too far away, however, the Huygens' principle my not be properly approximated in that the distance of adjacent elements has to be smaller than $\lambda/2$, otherwise, the sampling theorem is violated (not worked out here).

In the analysis of propagating ultrasound waves considered so far, we have assumed a vibrating surface consisting theoretically of an infinite number of point sources according to Huygens' principle. Yet, the transducer is a mechanical oscillator in its own right associated with its particular mechanical characteristics, viz., mass, elasticity and damping. In the simplest approximation, a transducer can be modeled as a linear oscillator. Its eigenfrequency is given by the mass and deformation parameters, while the width of the resonance curve is determined by the damping. Transducers are usually operated at their basic eigenfrequency and the bandwidth of the response corresponds to the resonance curve. Without going into mathematical detail, one finds

that the resonance curve becomes higher and narrower as the damping is smaller. Such a transducer exhibits on the one hand a high sensitivity; on the other, undesired transient vibration behavior ("ringing") renders a transducer with insufficient damping unsuitable. The higher the damping, in turn, the larger the bandwidth becomes and the better a pulse form is reproduced mechanically, this at the cost of sensitivity. In conclusion, the fabrication of ultrasound transducers requires a high amount of experience, precision work and skill and is a true industrial specialty.

1.2. Impedance Matching

The acoustic impedance $\rho \cdot c$ (density of the material times speed of sound, see Appendix 1) of transducers is typically a few orders of magnitude higher than the one of biological tissues (the density of water is 1 g/cm^3, of PZT around 10 g/cm^3, the speed of sound in water is 1500 m/sec, in PZT ten times higher). At interfaces between different materials, waves are reflected and refracted much as in optics. The amount of reflexion and the angle of refraction thereby depends on the ratio of the impedances of the adjacent materials (corresponding to the indices of refraction in optics). The solution of the wave equation (Appendix 1) shows that in case of large impedance jumps, waves are mostly reflected. Accordingly, ultrasound waves cannot be propagated through the skin from a transducer without appropriate impedance matching measures; in particular, if there is air (impedance four orders of magnitude smaller than that of water) in between. It has to be kept in mind, thereby, that in medical ultrasound echo techniques each impedance jump occurs twice, first during emission, second during backscatter.

Impedance matching is achieved by positioning suitable interface materials between transducer and skin in the form of surface layers on the transducer and coupling gels. The rule according to which the properties of the ideal matching material are determined can be found on the basis of a simplified model (Figure 8). We consider three layers, viz., the transducer, the matching layer (thickness D) and the medium into which the sound wave is to be propagated:

Z_1, Z_2, Z_3	acoustic impedances ($\rho_i c_i$, see Appendix 1), $Z_1 >> Z_3$
Ψ_1	velocity potential in the transducer, primary wave
Ψ_1'	velocity potential in the transducer, reflected wave
Ψ_2, Ψ_2'	same for the matching layer
Ψ_3	medium, outgoing wave

At each interface, there are two physical conditions to be met:

1st There is no cavitation, i.e., the velocities, $-\vec{\nabla}\Psi$, on both sides are the same.

2nd The interface itself is massless, implying that the pressures, $\rho\left(\partial\Psi/\partial t\right)$ (see appendix 1), on both sides are the same.

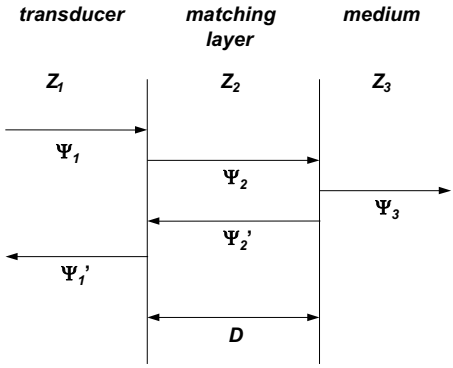


Figure 8. Simplified model of sound propagation through an interface layer of thickness D

The matching is ideal, when $\Psi_1' = 0$ (no reflexion into the transducer, all of the energy is transmitted into the medium). Upon solving the set of linear equations associated with these conditions, one finds that the thickness of the interface layer, D, has to be equal to

$$D = \lambda / 4 \text{ while the impedance } Z_2 = \sqrt{(Z_1 \cdot Z_3)}\,. \tag{5}$$

The first result ($D = \lambda/4$) can be obtained without calculation: Taking into account the phase jump of π when the wave is reflected at an interface with higher impedance, considering positive and negative interference of the forward and backward wave will lead to the desired result.

The refraction characteristics of acoustic waves can be used to design ultrasound lenses for focusing purposes (Figure 9). If the speed of sound in the lens material is higher than in biological tissue, a concave lens will produce focusing (in contrast to optical lenses). With phased arrays, focusing can be achieved by suitable timing of the individual pulses.

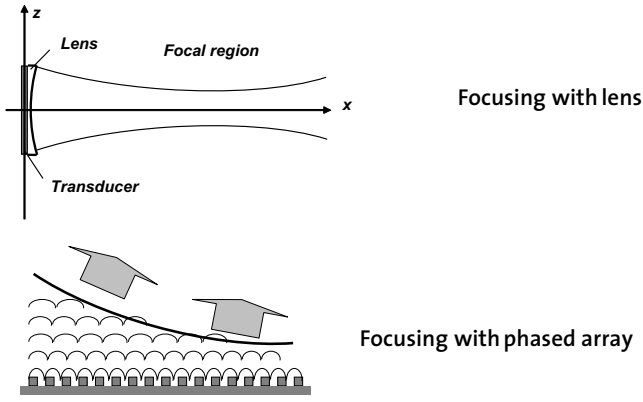


Figure 9. Focusing of ultrasound beams

2. Medical Echo technique

In clinical ultrasound applications, ultrasound pulses in the form of wave packages consisting of 4 to 16 sinusoidal waves are typically utilized (Figure 6). At a frequency f of 4 – 8 MHz this corresponds to pulses with a duration of 0.5 μsec – 4 μsec (as mentioned in the introduction, higher frequencies cannot be used because of damping, at lower frequencies, the spatial resolution would be insufficient). The length of such pulses is 0.75 mm - 6 mm while the width is determined by the transducer characteristics and focusing methods. In all of these numbers, it is assumed that the pulse rise and decay times are short in comparison with the pulse duration. It was discussed earlier that these properties, i.e., the transient behavior of the transducer is strongly related to the damping characteristics; the transducer is usually operated at its lowest resonance frequency.

The pulses are emitted repetitively with the pulse repetition frequency, f_{PR}. Since all echoes of interest have to return to the transducer before the next pulse is emitted, the pulse repetition frequency is determined by the operating range of the instrument. If, e.g, a maximal distance from the transducer of 25 cm has to be reached, the pulse repetition frequency is limited to 3 kHz corresponding to the return time at a speed of sound of 1500 m/sec. In case of shorter distances, f_{PR} can be increased which may be of importance for Doppler application as will be seen later.

With the aid of a phased array, a cross section of the human body can be obtained without moving the transducer. To this end, the beam is deflected in real time like a windshield wiper (Figure 10). If 25 frames/sec are made at a pulse repetition frequency of 3 kHz, say, each frame consists of 190 individual rays. Parallel beam technique can also be used if the anatomical location allows the application of an extended transducer array.

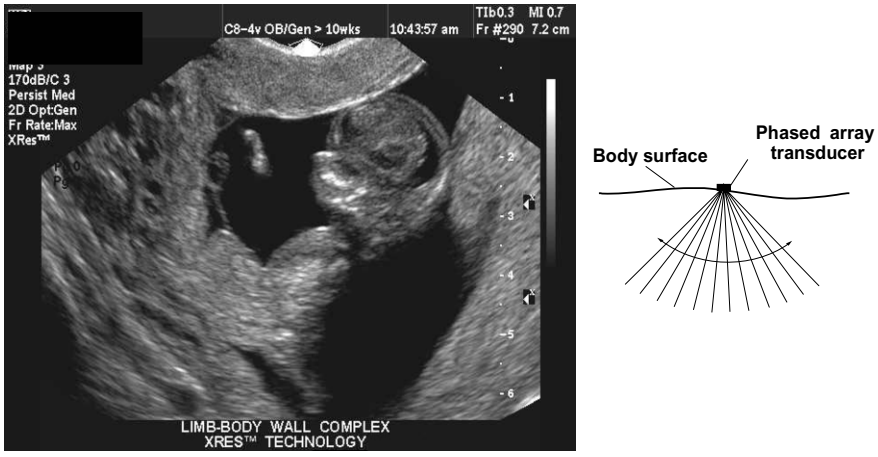


Figure 10. Scanning of a 10 week old foetus in a typical obstetrical application.

An echo image exhibits in essence all impedance jumps which a beam encounters as it propagates through the tissue of interest. If the propagation speed of the sound package can assumed to be constant (1500 m/sec such as in water) the conversion of the echo sequence into an image is straightforward. Damping is compensated by exponential amplification.

As mentioned earlier, backscatter may be noncoherent (advantageous for imaging) or coherent (image deterioration due to mirroring and speckle formation which is due to interferences). Nonlinear effects during scattering cause the formation of higher harmonics of the basic ultrasound frequency. The evaluation of the 2nd harmonic can be used to reduce speckle noise ("2nd harmonic imaging").

All body parts which are not covered by bone or air volumes where the large impedance jump usually prevents sufficient wave transmission can be imaged. Primary applications are in cardiology (heart) and obstetrics (foetus). Physiological fluids containing minute air bubbles (μm diameter) are sometimes injected systemically to serve as contrast media to enhance image contrast.

3. Medical Doppler technique

We consider again a pulsed ultrasound method where pulses are emitted with a pulse repetition frequency f_{PR}. Pulse characteristics (frequency, length, amplitude) are largely the same as those implemented in echo (imaging) applications.

If ultrasound waves are scattered by moving targets, Doppler shifts are observed which depend on the speed of the target and the direction of incidence and observation. Accordingly, this effect can be used to measure blood flow velocities since clouds of red blood cells act as scattering objects. (The diameter of red blood cells, 8 μm, is very small with respect to the ultrasound wavelength of around 0.5 mm. The wave backscattered from one single cell is therefore much too weak to be determined. Blood cells are inhomogeneously distributed within a blood vessel such that a net effect remains from the myriads of backscattered waves which would cancel themselves by interference if the scattering centers would be arranged homogeneously. Nevertheless, this net effect that makes up the Doppler signal, is usually small.)

Upon transmission of a pulse, the echoes return continuously to the transducer which, after completion of the transmission cycle, is immediately switched to receiving. Each echo is processed sequentially whereby it is subdivided into contiguous constant time steps of duration $\delta = \Delta t$, numbered as $m = 1,2,...,m_{max}$. Accordingly, each individual step corresponds to a fixed distance from the transducer (Figure 11).

We consider a pulse, n_i *(i = 1, …)* from the sequence of pulses regularly emitted by the transducer with the pulse repetition frequency f_{PR}, which is emitted at time $t_i = t_{i=1} + \frac{i}{f_{PR}}$. $t_{i=1}$ thereby denotes the instant in time when the measurement begins. The echo associated with this pulse is processed in consecutive channels *m* according to the sampling function (Figure 12)

$$\Omega_n^m(t) = 1 \quad for \quad t \in \left[t_{i=1} + \frac{n}{f_{PR}} + m\cdot\delta - \frac{\delta}{2}, \quad t_{i=1} + \frac{n}{f_{PR}} + m\cdot\delta + \frac{\delta}{2} \right] \qquad (6)$$
$$= 0 \quad otherwise$$

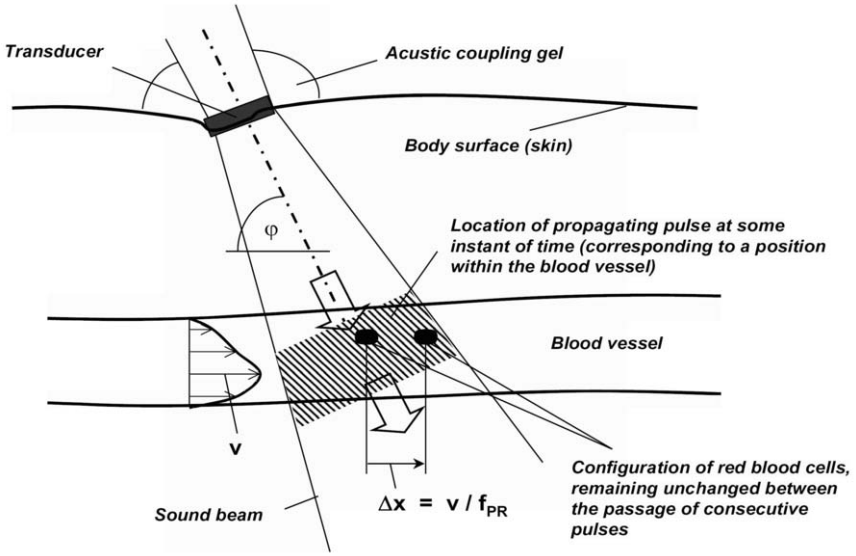


Figure 11. Principle of pulsed Doppler measurement

v	blood flow velocity
$\Delta x = v / f_{PR}$	distance traveled by a typical erythrocyte (red blood cell) configuration between the passage of two consecutive pulses
φ	angle of the beam axis with respect to the axis of the blood vessel, i.e., the direction of flow

As mentioned in Figure 11, a particular blood cell cloud travels a distance $\Delta x = \frac{v}{f_{PR}}$ between the passage of two consecutive pulses. We assume that this configuration remains constant during the time that it passes through the volume examined by the beam, i.e., each pulse "looks" at the same cloud such that differences between echoes which are due to small, flow-induced changes of the configuration can be neglected. Between each two consecutive pulses, the emitted pulse as well as the backscattered echo from this cloud have to propagate by an additional distance $\Delta x \cdot \cos(\varphi)$ on their way from and to the transducer. The associated time delay is

$$\Delta\tau = 2\frac{\Delta x}{c}\cos(\varphi) = 2\frac{v}{f_{PR}}\cdot\frac{1}{c}\cos(\varphi) \tag{7}$$

causing a phase shift denoted as Doppler shift of

$$\Delta\alpha = 2\frac{v}{f_{PR}}\cdot\frac{1}{c}\,\omega_0\cos(\varphi) \tag{8}$$

between consecutive echoes arriving at the transducer (ω_0 denotes the frequency of the ultrasound, c the speed of sound).

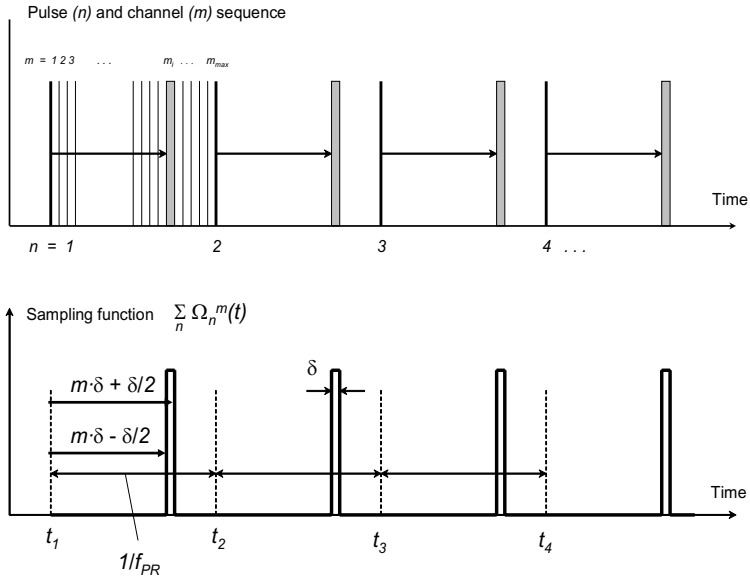


Figure 12. Echo processing sequence and sampling function. δ denotes the duration of each channel, *m*

Starting with pulse $n = 1$, the cloud under consideration is now assumed to be within the volume covered by the channel m_k of the beam, located within the vessel lumen. The signal produced by the transducer in channel m_k is

$$S_{m_k} = \sum_n \Omega_n^{m_k}(t) \left[A_1 \cos(\omega_0 t) + A_2 \cos(\omega_0 t + n \cdot \Delta\alpha) \right] \tag{9}$$

$A_1 \cos(\omega_0 t)$ echo from stationary objects, e.g., the vessel wall within the channel volume m,

$A_2 \cos(\omega_0 t + n \cdot \Delta\alpha)$ Doppler-shifted signals,

i.e., the signal consists of stationary, non-shifted contributions as well as of Doppler-shifted parts. In practical applications, it turns out that for the amplitudes $A_1 >> A_2$ usually holds so that signal/noise considerations are particularly important in Doppler methods.

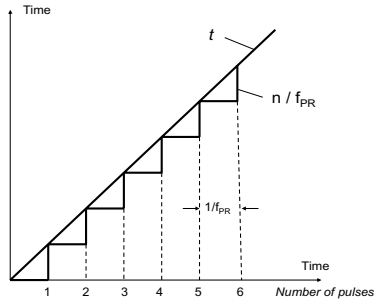
In order to obtain the flow velocity "hidden" in S_{m_k}, the signal first has to be demodulated by multiplication with $\cos(\omega_0 t)$. Upon application of the theorem

$2 \cos\gamma \cos\delta = \cos(\gamma + \delta) + \cos(\gamma - \delta)$

and low-pass filtering (contributions with the frequency $2\,\omega_0 t$ can easily be removed) one obtains

$$S'_{m_k} = \sum_n \Omega_n^{m_k}(t) \left[A_1 + A_2 \cos\left(2 \frac{n}{f_{PR}} \cdot \frac{v}{c} \omega_0 \cos(\varphi) \right) \right] \quad (10)$$

The step function $\frac{n}{f_{PR}}$ can be approximated by the time t and the expression $\omega_D = 2 \frac{v}{c} \omega_0 \cos(\varphi)$ corresponds to the usual Doppler formula. The (mostly strong) stationary echo A_1 has to be eliminated by high-pass filtering. This is a nontrivial task because, as can be seen from the Doppler formula, the Doppler shift goes to zero with the flow velocity. If a filter is used that does not exhibit an extremely steep fall-off above zero, important contributions to the flow signal are eliminated along with the stationary echo. In turn, radial vessel wall motions which may reach up to 1/10 of the average flow velocity under pulsatile flow conditions (arteries) and produce a strong echo may adversely contaminate the desired signal from the blood and may have to be suppressed by choosing appropriate filter characteristics.



An important consequence is associated with the fact that the Signal S_m is sampled with the sample frequency $1/f_{PR}$ according to the sample function Ω. Assuming, say, $\omega_0 = 4$ MHz, $v = 1$ m/sec and $\cos\varphi = 0.5$, one finds $\omega_D = 2.7$ kHz. If the pulse repetition frequency is 3 kHz in order to reach a desired distance, the Nyquist limit (sampling theorem) is already close. Higher flow velocities which are often reached in arteries will induce aliasing. Although there are several methods to extend the Nyquist limit, the problem remains. It can be added that in continuous-wave Doppler instruments (not described here) this problem is absent. If the channels are sufficiently small and the blood vessel sufficiently large, finally, the pulsed Doppler method allows flow profiles to be determined.

4. Duplex Scanner, Color Scanner

A combination of echo-imaging and Doppler analysis is of interest in many medical applications, in particular in cardiology. In duplex scanners both modalities are integrated independently while in color scanners the blood flow information obtained from the Doppler signals are overlaid in color over a black and white echo image. These modalities are especially attractive because they operate in real time. Since the imaging process involves a continuous scanning of the entire region to be imaged the area covered by a blood vessel or a heart chamber from where the Doppler information is desired is “seen” only during short fractions of the scanning sequence (Figure 13). As has been demonstrated above, the phase shift between consecutive pulses is needed in order to evaluate the Doppler signal quantitatively, however. In general, therefore, duplex or color scanners exhibit qualitative flow information only, based on limited measurements. (The imaging frame rate, typically 25 Hz, is too small, by orders of

magnitude, to allow for a Doppler evaluation.) More sophisticated scanning schemes have nevertheless been devised (the imaging beam is e.g. delayed at the area of interest for a more precise Doppler measurement) which allow for more detailed analyses. Figures 14 and 15 exhibit typical examples.

5. Further Developments

It has been mentioned previously that nonlinear effects associated with the propagation and scattering of ultrasound involve „mode conversion“, i.e. the generation of higher harmonics. This effect can be used mainly to reduce the influence of speckle noise (2^{nd} harmonic imaging or tissue harmonic imaging).

Intravasal ultrasound imaging using circular transducers emitting beams in a radial direction have been developed. This method allows us to document, e.g., pathologic vessel wall conditions such as intimal thickening, fatty streaks or stenoses. Catheterization is of course necessary. Full 3D noninvasive flow imaging is presently only possible with the aid of expensive and partially non-real time MR methods.

Two dimensional phased arrays allow the real-time imaging in volumes as, with the aid of such transducers, a volume can be scanned without moving the transducer. Figure 16 shows an application in cardiology.

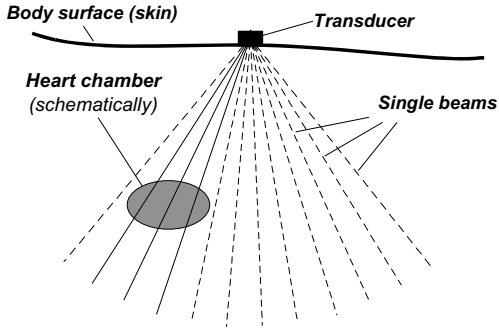


Figure 13. Principle of duplex scanning (schematically). The area of interest for a Doppler flow measurement (heart chamber) is covered only by three beams during each imaging cycle. Accordingly, the Doppler signal has to be obtained from two phase measurements only unless a more sophisticated scanning scheme is applied.

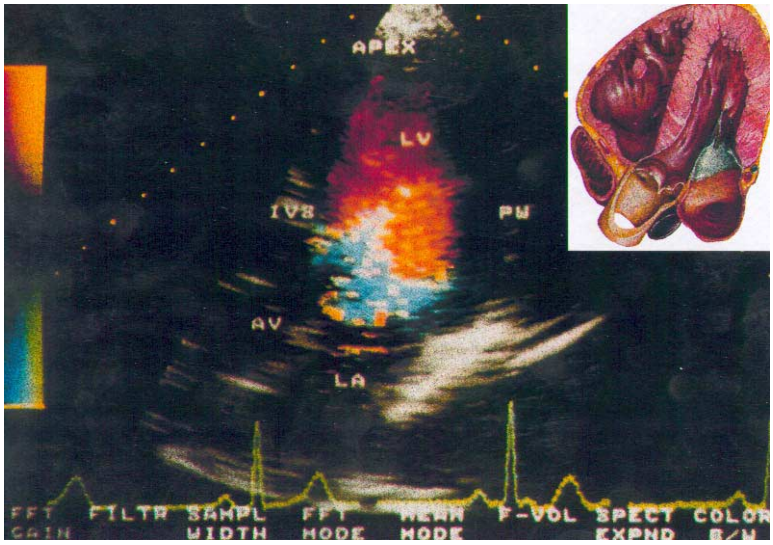


Figure 14. Imaging of the heart with overlaid Doppler information in color

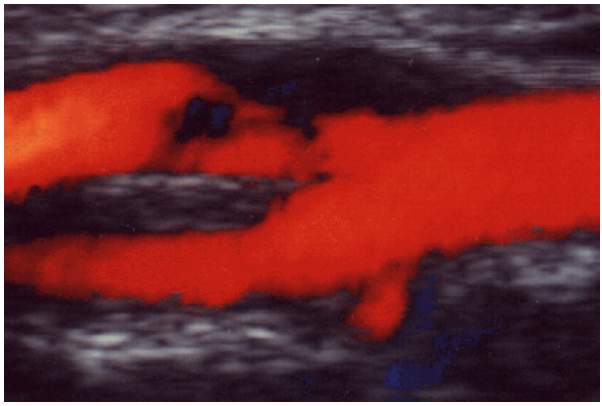


Figure 15. Imaging of a carotic artery bifurcation with stenosis. A recirculation zone behind the stenosis manifests itself.

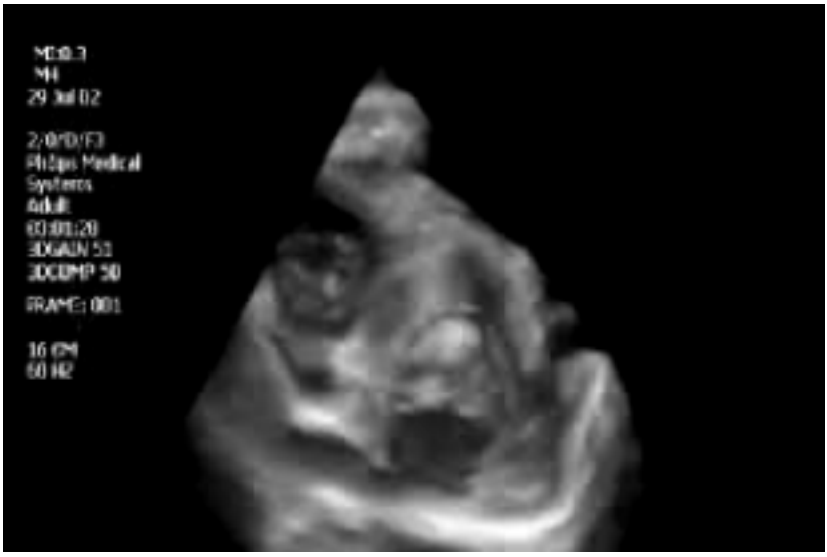


Figure 16. Three-dimensional real-time imaging of the heart valves by utilizing a two dimensional phased array transducer

Further Reading

[1] P.N.T. Wells, Ultrasound Imaging, *Phys Med Biol* **51** (2006), R83-R98
[2] M.F. Hamilton, D.T. Blackstock, *Nonlinear Acoustics*, Academic Press, New York, 1998
[3] F.A. Duck, Nonlinear Acoustics in Diagnostic Ultrasound, *Ultrasound Med Biol*, **28** (2002), 1-18
[4] A. Kurjak, S. Kupesic, *Clinical Application of 3D Sonography*, Parthenon Publ., New York, 2000

Appendix 1: Wave Propagation

We consider linear wave propagation in an ideal fluid (simplified acoustic model) which can serve as a sufficiently accurate approximation for most conditions prevailing in medical ultrasound modalities. The two physical quantities of importance within the framework of this approximation are the density of the material where the sound propagates $\rho(\vec{r},t)$ and the velocity field therein $\vec{v}(\vec{r},t)$ as function of space $(\vec{r})$ and time (t). The two equations that are needed to describe this model are the

- continuity equation $\vec{\nabla}\cdot(\rho\vec{v}) = -\frac{\partial\rho}{\partial t} = -\rho_p\frac{\partial p}{\partial t}$ and the (1)

- momentum equation $\rho \frac{\partial \vec{v}}{\partial t} = -\vec{\nabla} p$ (2)

p thereby denotes the (hydrostatic) pressure and $\rho_p = \frac{d\rho}{dp}$ the compressibility of the fluid. The second equation represents a linear approximation of the Navier-Stokes equation for an ideal (frictionless) fluid.

Upon introducing a scalar velocity potential $\Psi(\vec{r},t)$ (irrotational flow field, i.e., no eddies and no turbulence) such that the velocity field $\vec{v}(\vec{r},t) = -\vec{\nabla}\psi(\vec{r},t)$ one finds $p = \rho \frac{\partial \Psi}{\partial t}$, furthermore the wave equation

$$\Delta \Psi = \rho_p \frac{\partial^2 \Psi}{\partial t^2} \quad \text{with the wave speed} \quad c^2 = \frac{1}{\rho_p} \tag{3}$$

In this derivation, a term containing $\vec{\nabla} p$ has been neglected which is in agreement with the acoustic approximation.

We now introduce a Cartesian coordinate system such that the velocity field $\vec{v}(\vec{r},t)$ can be represented as $\vec{v} = (v_x, v_y, v_z)$ (the independent variables $\vec{r}$ and t are thereby omitted for brevity). Forward and backward running plane waves propagating in direction z are solutions of equation (3)

$$\begin{aligned} v_x &= v_y = 0 \\ v_z &= v_0 \, e^{i(k z \pm \omega t)} \\ \Psi &= -\frac{v_0}{i k} e^{i(k z \pm \omega t)} \end{aligned} \tag{4}$$

with the frequency ω and the wave number $k = 2\pi/\lambda$ (wavelength λ).

In analogy with electric systems, the acoustic impedance of a medium is defined as the ratio between the pressure (analogous to electric tension) and the flow (analogous to electric current)

$$I = \frac{p}{v} = \rho c \tag{5}$$

At locations where the impedance changes, scattering, reflexion and refraction set in as can be found from general solutions of the above equations. If the impedance change occurs over a well-defined surface (e.g., a blood vessel wall) mirror effects and coherent scattering are also observed. The latter leads to interference phenomena which are associated with speckle formation. Further effects (not included in the foregoing equations) are damping and nonlinear contributions.

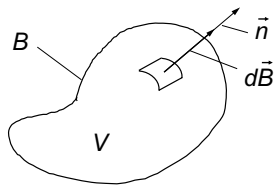
Appendix 2: Huygen's Principle

We consider an arbitrarily shaped surface of limited extent which is set into motion, e.g., by the piezoelectric effect. We intend to calculate the wave field in the adjoining space. The boundary of this space contains the vibrating surface and is closed by extending this surface to enclose a virtual volume. Mathematically, the problem consists of the wave equation according to Appendix 1, to be solved for the virtual space whereby the motion of the boundary (enclosing surface) is prescribed.

Green's formula is useful to approach this problem

$$\int_V \left(u \cdot \Delta w \; - \; w \cdot \Delta u \right) dV \;\; = \;\; \oint_B \left(w \frac{\partial u}{\partial \vec{n}} \; - \; u \frac{\partial w}{\partial \vec{n}} \right) d\vec{B} \tag{1}$$

u, w denote arbitrary scalar differentiable functions,
V an arbitrary volume with closed boundary B
$\vec{n}$ the outer normal
($d\vec{B}$ has the direction of the outer normal and $\frac{\partial}{\partial \vec{n}}$ denotes the derivative in the direction of the outer normal.)



With the aid of this formula, a volume integral is transferred into an integral over its (closed) surface. It is applied here such that u is identified with the solution of the wave equation to be determined, ψ, and for w a special function (Green's function) is inserted, namely a spherical wave (distance from the center r, wave number k, frequency ω, time t)

$$w \;\; = \;\; \frac{1}{r} e^{-i(kr - \omega t)} \tag{2}$$

(The choice of the Green's function is crucial and problem-dependent; for a related theory see Courant and Hilbert, *Methods of Mathematical Physics*). Since ψ and w are supposed to satisfy the same wave equation (harmonic waves with frequency ω and wave number k, the left side of equation (1) vanishes.

After this specialization, equation (1) reads

$$\oint_B \frac{1}{r} e^{-ikr} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} \;\; - \;\; \oint_B \Psi \frac{\partial \left[(1/r) e^{-ikr} \right]}{\partial \vec{n}} d\vec{B} \;\; = \;\; 0 \tag{3}$$

(The term $e^{i\omega t}$ is omitted for brevity and only included at the end again.) In order to determine the function ψ in point $\boldsymbol{P}$ (Figure 1), we proceed as follows: For simplicity (without loss of generality), we assume that the vibrating surface (transducer) is plane, bounded, part of the boundary B and located in the $y - z$ plane according to Figure 1. The virtual space with boundary B is then extended to include the entire right half space.

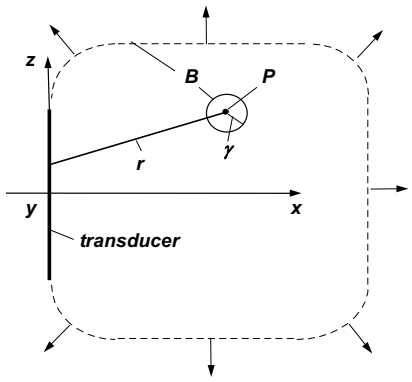


Figure 1. Derivation of Huygens' principle. The closed surface ***B*** includes the vibrating surface (transducer) and is closed such as to include the entire right half space (the broken line is extended to infinity.

Next, the point ***P*** is excluded from the volume by a small sphere (radius γ). The surface ***B*** now includes also the surface of this sphere. Later, we let $\gamma \to 0$. (That this procedure leads to the desired result, is in essence due to the particular properties of the Green's function.) We can assume (since we are seeking a physically realistic solution) that ψ is finite and almost constant (in particular, for $\gamma \to 0$) within the small sphere such that the term $\frac{\partial \Psi}{\partial \vec{n}} d\vec{B}$ equals $\left[\frac{\partial \Psi}{\partial r}\right]_\gamma \gamma^2 \cos(\varphi)\, d\varphi\, d\vartheta$

($\gamma^2 \cos(\varphi)\, d\varphi\, d\vartheta$ is the surface element on the sphere with polar angles φ and ϑ).

Upon evaluation of equation (3) for the sphere, we obtain, for $\gamma \to 0$,

$$\oint_B \frac{1}{\gamma} e^{-ik\gamma} \left[\frac{\partial \Psi}{\partial r}\right]_\gamma \gamma^2 \cos(\varphi)\, d\varphi\, d\vartheta - (\Psi)_\gamma \left[\frac{\partial}{\partial r}\left(\frac{1}{r} e^{-ikr}\right)\right]_\gamma 4\pi\gamma = 0 - 4\pi\Psi_P \quad (4)$$

Accordingly,

$$4\pi\Psi_P = \oint_B \frac{1}{r} e^{-ikr} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} - \oint_B \Psi \frac{\partial \left[(1/r) e^{-ikr}\right]}{\partial \vec{n}} d\vec{B} \quad (5)$$

whereby the integral now extends over the outer boundary only.

$\frac{\partial \Psi}{\partial \vec{n}} = \vec{v}_{\vec{n}}$ denotes the component of the velocity to be prescribed along the surface ***B*** which is perpendicular to the surface. Since only the surface of the transducer is in vibratory motion, edge effects and components of the velocity which are directed along the surface are neglected and Ψ is assumed to go to zero at infinity, $\frac{\partial \Psi}{\partial \vec{n}}$ is different

from zero only on the transducer surface. The integral (5) therefore extends over the transducer surface only,

$$4\pi\Psi_P = \int_{Tr} \frac{1}{r} e^{-ikr} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} - \int_{Tr} \Psi \frac{\partial \left[(1/r) e^{-ikr} \right]}{\partial \vec{n}} d\vec{B} \qquad (6)$$

The integral (6) can further be simplified by making the special assumption that the gradient $\frac{\partial \Psi}{\partial \vec{n}} = \vec{v}_{\vec{n}}$ is symmetric with respect to the surface, i.e., both sides of the transducer (the surface facing the $+x$ direction and the one oriented towards $-x$, Figure 1) vibrate in phase but in opposite directions. The same procedure can now be applied to the negative half space except for the small sphere which is absent since the point ***P*** of interest is on the right side. We therefore obtain a similar expression (6) for the left half space, but the left hand side is zero,

$$0 = \int_{Tr} \frac{1}{r} e^{-ikr} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} - \int_{Tr} \Psi \frac{\partial \left[(1/r) e^{-ikr} \right]}{\partial \vec{n}} d\vec{B} \qquad (7)$$

Again, the integral extends only over the transducer surface. However, comparing the two sides of the transducer, we observe that

$$\frac{\partial}{\partial \vec{n}} \rightarrow -\frac{\partial}{\partial \vec{n}} \quad \text{and} \quad d\vec{B} \rightarrow -d\vec{B}$$

while we assume that $\frac{\partial \Psi}{\partial \vec{n}} \rightarrow \frac{\partial \Psi}{\partial \vec{n}}$ since both sides exhibit the same vibration component. Accordingly, the relative sign in equation (7) is different from the one in equation (6). Both integrals are therefore equal, and we obtain the usual formulation of Huygens' principle,

$$\Psi_P = \frac{1}{2\pi} \oint_B \frac{1}{r} e^{-i(kr - \omega t)} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} \qquad (8)$$

The wave field can be represented as a superposition of spherical waves emanating from the transducer surface. From a numerical point of view, this formulation is far more efficient for calculating sound fields that the integration of the wave equation. An application is shown in Figure 2.

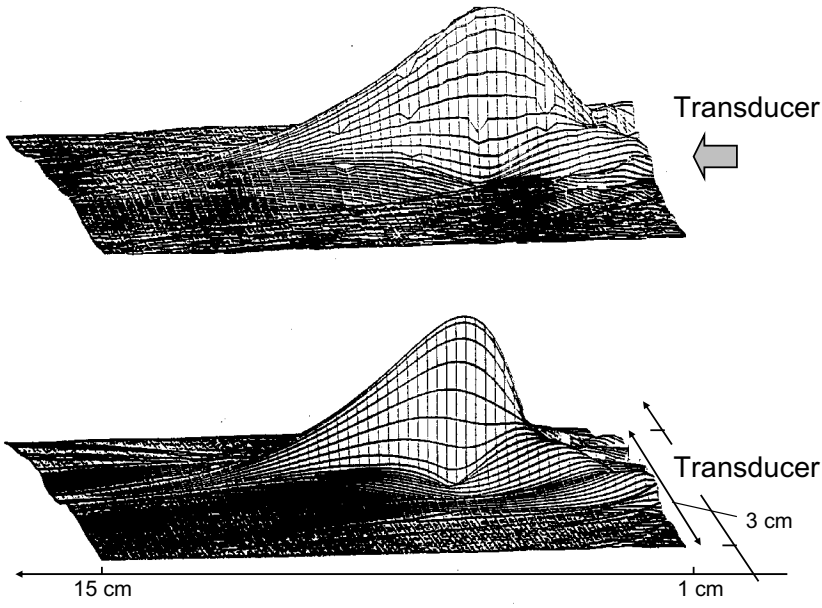


Figure 2. Intensity distribution of a circular transducer measured (top) with the aid of a hydrophone and calculated using Huygens' principle (bottom).

Appendix 3: Beam Formation

Phased array transducers are used to produce ultrasound beams according to a pattern in space and time that is shaped for a particular application. To this end, amplitude and phase of each individual transducer element have to be chosen appropriately. In the following, a strategy is derived according to which these characteristics can be determined when a particular sound field is prescribed.

We confine ourselves to a one-dimensional array (Figure 1) for simplicity and apply Huygens' principle as shown in Appendix 2

$$\Psi_P = \frac{1}{2\pi} \oint_B \frac{1}{r} e^{-i(kr-\omega t)} \frac{\partial \psi}{\partial \vec{n}} d\vec{B} \qquad (1)$$

The array consists of elements with width b and infinitesimal height $d\eta$. (The latter is assumed for ease of calculation. This corresponds to an unrealistic amount of infinitely many transducer elements. The question with respect to the number of elements necessary to approximate Huygens' principle with sufficient accuracy is addressed later.)

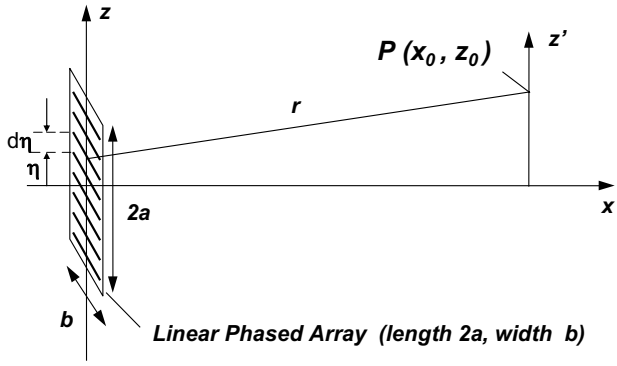


Figure 1. Linear phased array. The elements (width *b* and height dη) are aligned along the coordinate *z*.

In the problem to be solved here, Ψ_P is prescribed while in turn

$$\left[\frac{\partial \psi}{\partial x}\right]_{x=0} = v_0(\eta)\, e^{i\varphi(\eta)} \tag{2}$$

is to be determined. Thereby, $v_0(\eta)$ denotes the amplitude and $\exp[i\varphi(\eta)]$ the phase function. Outside $z = \pm a$ we have $\psi = 0$, and within $z = \pm a$ there is only the derivative with respect to x different from zero. η denotes a coordinate along the array in direction z.

The field ψ_P in point $\boldsymbol{P}(x_0, z_0)$ (Figure 1) is therefore given by

$$\Psi(x_0, z_0, t) = \frac{b\, e^{i\omega t}}{2\pi} \int_{-a}^{a} v_0(\eta)\, e^{i\varphi(\eta)} \left(\frac{e^{-ikr}}{r}\right) d\eta \tag{3}$$

with $r = \sqrt{x_0^2 + \left(z_0 - \eta^2\right)}$

Upon expansion of the square root $\left(\eta, z_0 << r\right)$ one arrives at Fresnel's approximation,

$$\Psi(x_0, z_0, t) = \frac{b \cdot exp\left[-i \left(k\, x_0 + \frac{k\, z_0^2}{2\, x_0} - \omega t \right) \right]}{2 \pi x_0} \bullet \int_{-a}^{a} v_0(\eta)\, e^{i \varphi(\eta)}\, exp \left(\frac{i k \left[z_0\, \eta - \frac{\eta^2}{2} \right]}{x_0} \right) d\eta \tag{4}$$

In addition, in the far-field, we have $x >> \eta$, such that a further step can be performed to arrive finally at the Fraunhofer approximation, viz.,

$$\Psi(x_0, z_0, t) = \frac{b \cdot e^{-i(k\, x_0 - \omega t)}}{2 \pi x_0} \cdot \int_{-a}^{a} v_0(\eta)\, e^{i \varphi(\eta)}\, e^{\frac{i k z_0 \eta}{x_0}}\, d\eta \tag{5}$$

Accordingly, the farfield and the distribution of the amplitude and phase on the transducer are connected by a Fourier-transform, i.e., the excitation $v_0(\eta) e^{i \varphi(\eta)}$ to be determined can be obtained from a Fourier-transform of the desired field.

In a practical application, the integral in equation (5) is replaced by a sum according to the size, shape and number of individual transducer elements making up the phased array. The smaller the elements, the more elements can be accommodated within a given transducer surface and the better the integral can be approximated, yet, mechanical and electrical impedance problems along with manufacturing procedures prevent the implementation of too small elements. In order to avoid crosstalk, furthermore, the individual elements have to be separated by a certain distance which has to be determined empirically. It can be shown (not worked out here, but easily understandable qualitatively) that the distance between adjacent transducer elements must not be larger than half the wavelength of the desired ultrasound wave (Nyquist limit). The larger the transducer elements (however observing the Nyquist limit), the less elements can be accommodated; it turns out that the undesired side lobe activity increases at the same time. Manufacturing ultrasound transducers is a great specialty, scientific considerations and practical manufacturing aspects are both important.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-274

V.2. X-Ray-Based Medical Imaging:

X-Ray Projection Technique
Image Subtraction Method
Direct Digital X-Ray Imaging
Computed Tomography (CT)

Peter F. NIEDERER
Institute for Biomedical Engineering
ETH Zurich and University of Zurich
Gloriastrasse 35
CH-8092 Zurich, Switzerland

Abstract. X-ray projection imaging, introduced in the early 20th century, was for a long time the only and still is a major routine diagnostic procedure that can be performed on the intact human body. With the advent of Computed Tomography in 1973, for the first time, a true 3D imaging of anatomical structures became feasible. Besides, digital recording and processing allowed further developments such as subtraction angiography in real-time or direct x-ray imaging without wet photographic methods.

Keywords. X-ray, Computed Tomography, digital imaging

Introduction

In 1896 W.C. Röntgen announced the discovery of the potential of an at that time yet unknown type of radiation to penetrate materials and to cause opacification of photographic plates (after development). It turned out that various biological tissues attenuated this radiation differently such that the generation of projections of the human body onto a photographic plate was possible (Figure 1). As the physical nature of the radiation was unknown, it was said to consist of „X-rays“ – a designation which is still in use today.

A further milestone was the introduction of computed x-ray tomography (CT) by G.N. Hounsfield in 1973 which for the first time allowed to produce cross sections of the human body for diagnostic purposes *in vivo* and noninvasively apart from the application of a radiation dose (Figure 2).

The x-rays used for diagnostic purposes in medicine are electromagnetic waves with a photon energy between 30 - 200 KeV typically. “X-ray” was until a few years ago the most often applied medical imaging technology. Since then, methods on the basis of ultrasound (in contrast to x-ray non-ionizing) have become more often used. In addition, MRI (Magnetic Resonance Imaging) is of rapidly increasing significance in medical imaging, in particular as with the combination MRI/MRS (MR Imaging and

Spectroscopy) also functional examinations can be made. The combination of x-ray CT or MRI with PET (Positron Emission Tomography), in turn, has in recent years fostered significant and novel clinical applications for functional investigations.

Figure 1. X-ray of the hand of W.C. Röntgen's wife

The radiation dose which is delivered to the patient is always of concern; it should be as small as possible. An enormous progress has been achieved through the years such that today a "simple" x-ray image can be produced with the aid of a radiation dose which can be considered harmless (Figure 3).

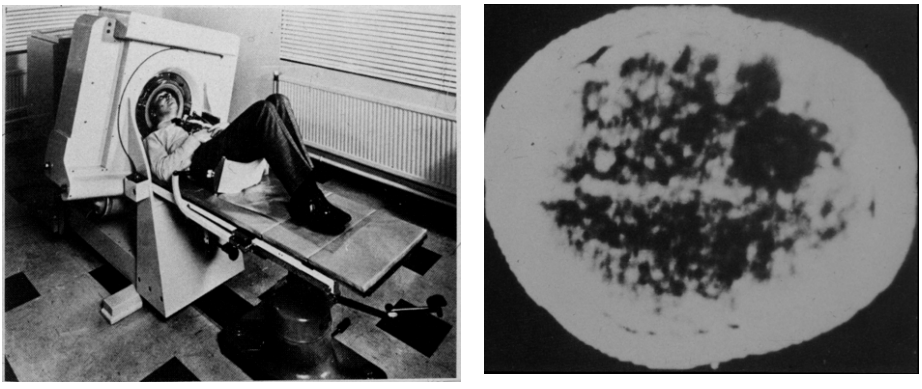


Figure 2. First CT scanner (1973) and first image of a brain tumor.

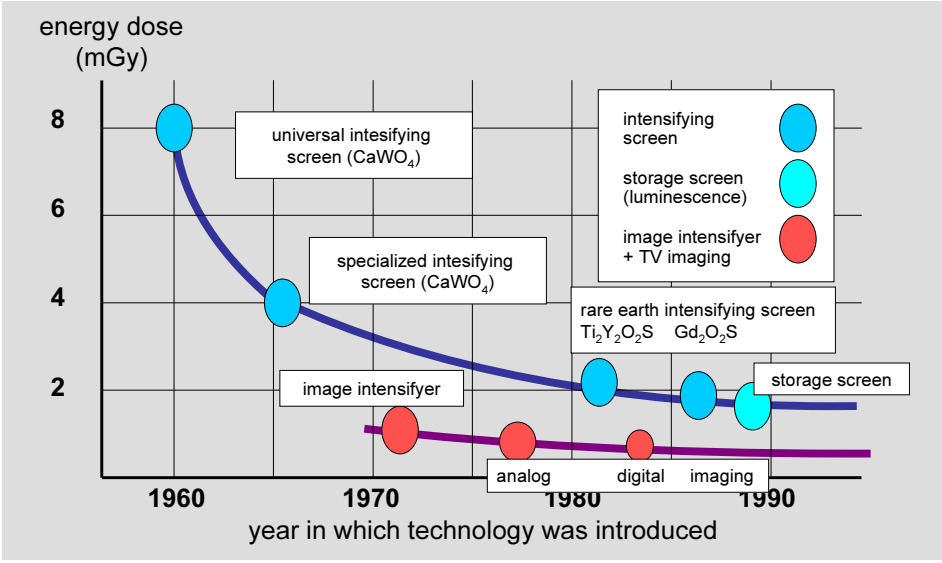


Figure 3. Typical radiation dose delivered for a x-ray projection image

1. Generation of X-Rays

X-rays for diagnostic purposes in clinical medicine are usually generated with the aid of a x-ray tube (in order to produce the much higher-energy therapeutic radiation for tumor treatment in the MeV-range, linear accelerators are mostly used). In such a tube whose interior is highly evacuated (the pressure is typically 10^{-6} Torr ~ 10^{-4} Pa), electrons are „evaporated“ from a wolfram cathode by heating (the melting temperature of wolfram is 3400° C) and exposed to an anode voltage, V, of some 30’000 to 200’000 Volt for acceleration towards the anode. The electrons thereby obtain a kinetic energy of 30 to 200 keV (electron charge, e, times potential difference, V). The geometric arrangement of the cathode along with the spatial characteristics of the accelerating field are designed such that the electron beam hits the target (anode) in a carefully outlined distribution (focal spot, see later).

When the accelerated electrons are being absorbed by the target material, x-ray is generated due to two effects, viz.,

- *bremsstrahlung*
- *characteristic radiation*

Besides that, most of the kinetic energy of the electrons is converted into heat (see below).

Bremsstrahlung is emitted, when fast electrons interact with the strong (positive) electric field of atomic nuclei. The electrons are deflected in this field and radiate electromagnetic waves in the form of photons. The electrons thereby loose kinetic energy, *i.e.*, are braked down (therefore the German expression „bremsstrahlung"). The spectrum of this radiation is continuous because the effect depends essentially on the distance at which the electron passes the nucleus and each distance has the same probability. The spectrum therefore extends from (theoretically) zero up to the maximal possible photon energy, E_{max}, which occurs when the electron looses its entire kinetic energy in one step.

The spectrum as function of energy associated with a thin target is constant, because most electrons, while traversing the thin layer, are involved in only one scattering process, thereby loosing part of their energy such that each portion of energy from zero to the maximum has the same probability. In a thick target, in turn, the spectrum is linearly decreasing with $E \rightarrow E_{max}$, because except for a few electrons which are backscattered at the surface layer all electrons always loose their entire kinetic energy in various single steps of different intensity (there are more small steps possible than large ones). According to

$$E_{max} = e\,V = h\,\nu_{max} \qquad (1)$$

(h: Planck's constant, ν: photon frequency, e: electron charge), the minimal wavelength, λ_{min}, which is associated with the maximal photon energy E_{max}, is (speed of light $c = \lambda\nu$)

$$\lambda_{min} = h\,c / e\,V \qquad (2)$$

with $h = 6.63 \cdot 10^{-34}$ W sec^2
$c = 3 \cdot 10^{8}$ m/sec
$V = 100$ kV
$e = 1.6 \cdot 10^{-19}$ Clb

one obtains $\lambda_{min} = 12.4 \cdot 10^{-2}$ Å $= 1.24$ nm. Typical spectra are shown in Figure 4. The decrease at low energies is on the one hand due to self-absorption in the target material, and on the other to prefiltering (low energy photons are removed purposely because they do not contribute to image formation but just increase the radiation dose imparted to the patient, see below and 2. Absorption of X-Rays).

The electrons hitting and penetrating the target not only interact with the nuclei of the target material, but also with the electrons thereof. Electrons may thereby be ejected from their orbit, and radiation quanta are emitted during subsequent recombination. Since the orbits have defined energy levels, the quanta have energies corresponding to the various orbits (K, L, ..., Figure 4) and the spectrum consists accordingly of lines (characteristic x-ray).

Most of the kinetic energy is however converted into heat (~99%) in the target material (electromagnetic waves with long wavelengths, largely due to interactions with free and weakly bound electrons of the target material). As mentioned above, the low-energy part of the spectrum is usually removed with the aid of prefilters (Al or Cu sheet metal plates) in order to reduce the radiation dose.

The radiation beam is strictly speaking not continuous as it is composed of a large number of incoherent electromagnetic wave packets (photons; the scattering processes in the target material occur independently from one another). The number of photons is

so large, however, that quantum noise is not of importance (exception: image subtraction, see later).

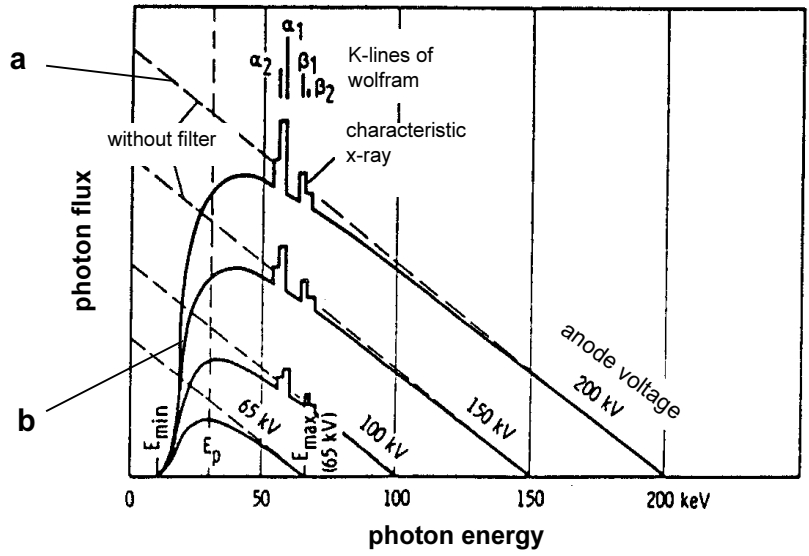


Figure 4. Typical spectrum of an x-ray tube (photon flux as function of energy)
a: Bremsstrahlung (theoretical, without prefilter)
b: Spectrum at different anode voltages; wolfram target,
1 mm Al prefilter (bremsstrahlung + superimposed characteristic x-ray)

There is a great variety of x-ray tubes available for different medical applications; yet, the basic physical effects are always the same. Material problems are prominent (hot cathode, extreme local heating of the anode). At a typical anode current (= electron current) of I = 100 mA and a voltage U = 100 kV the power deposited in the anode is

$$E = I \cdot U = 10 \text{ kW} \quad (3)$$

A projection image with a good resolution is only obtained if the focal spot (Figure 5) is small (another possibility is to locate the x-ray tube as far away from the object as possible; this is however limited due to practical reasons). X-rays propagate along straight paths, there exist no x-ray „lenses“ for practical medical imaging purposes[1]. The electrons are focused on the focal spot by the suitably shaped accelerating field. Since all the heat is produced in the focal spot, a small spot is difficult to achieve. Therefore, the anode is usually inclined with respect to the electron beam such that the focal spot can be extended (Figure 6) and the heat production is distributed accordingly. In the direction of the projection, the focal area still has a localized aspect (for special

[1] New developments in x-ray technology include the usage of diffraction (phase contrast imaging, much as in optical microscopy); likewise, Bragg reflection can be used for special purposes.

purposes, also tubes with a line focus are used). In addition, in high power tubes the anode rotates (5000 min^{-1} typ.) such that the focal spot is not static.

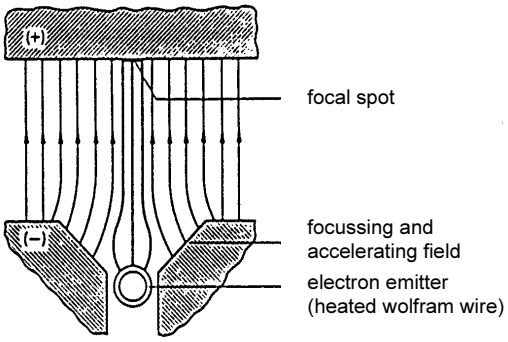


Figure 5. Focusing of electrons, originating from the cathode (-) onto the focal spot of the anode (+)

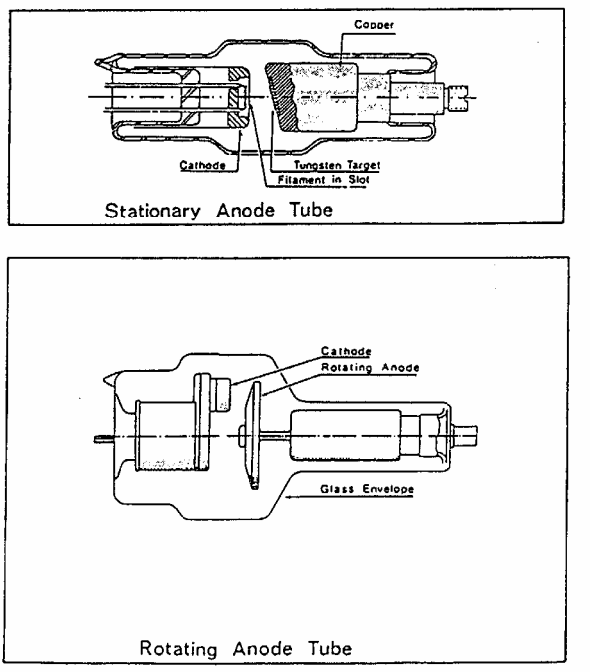


Figure 6. Schematic view of static (top) and rotating (bottom) anode tub

The heat is disposed of by diffusion through the anode as well as by radiation. The latter obeys the Stefan-Boltzmann law according to which a black body radiates the energy per unit area, W, as

$$W = \sigma T^4 \tag{4}$$

(σ = Stefan-Boltzmann constant, $5.7 \cdot 10^{-12}$ W/cm^{2}K^4, T = temperature)

At $T = 2500°$ K and an emission efficiency of 0.7 (graphite; the emission efficiency of a black body is 1) the radiated power per unit area is $W \sim 200$ W/cm^2. Even under normal operating conditions, the anode may be extremely loaded thermally (Fig.ure 7).

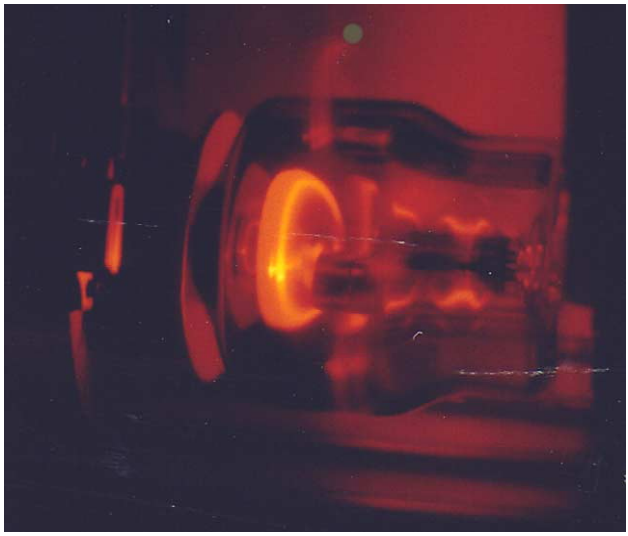


Figure 7. Red hot rotating anode under typical working conditions (courtesy: Comet AG, Bern)

The further away the tube is positioned from the object, the sharper the images become since the focal spot has always a certain size. In contrast, the recording medium, *e.g.*, a x-ray film cassette should be located as closely as possible behind the object.

For certain applications, in particular computed tomography (see later), the high voltage used in the tube for the acceleration of the electrons has to be extremely stable in order that the x-ray spectrum is constant in time. For the same reason, the stability of the electron beam hitting the target is an important aspect.

2. Absorption of X-Rays

In the energy range which is used for diagnostic medical imaging (ca. 30 - 200 KeV) two absorption effects occur when the photons interact with material, *viz.,*

- *photo effect* (Figure 8, a)
- *Compton effect* (Figure 8, b).

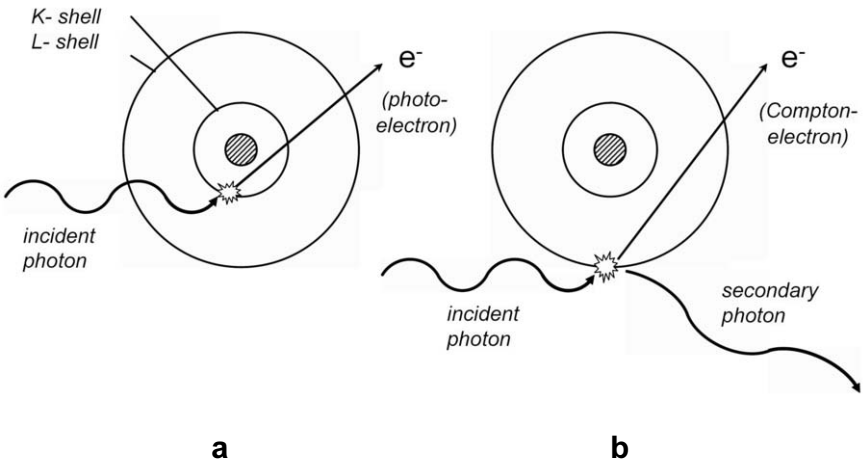


Figure 8. Photo effect (**a**), Compton effect (**b**)

The photo effect is a process where a photon is absorbed by an electron which is ejected from its shell to a higher energy level within the atom or completely removed from the atom (ionisation). The photon energy must at least correspond to the energy difference between the two electron states in question, *i.e.*, to the difference between the initial and final state of the excitation process (multiphoton processes are not observed with the photon fluxes used in medical imaging). When the absorption spectrum of a material which consists only of one type of atoms is considered, "edges" are therefore observed which correspond to the energy levels associated with the various electron shells (K-, L-, M, etc.). As the K-electrons exhibit the highest binding energy, the K-edge is encountered first on the energy scale when approached from above (see Fig. 9, a). This property is being used in the energy subtraction method (see later).

With the Compton effect (Figure 9, b), occurring at higher energies, a photon is likewise absorbed by an electron which is thereby excited, but there is excess energy which is emitted at the same time in the form of a lower energy secondary photon. These secondary photons represent scattered radiation since their direction does in general not coincide with the one of the original photon. They have a blurring effect and may lead to such a high noise level in an image that countermeasures have to be taken (collimators, anti scatter grids, see below).

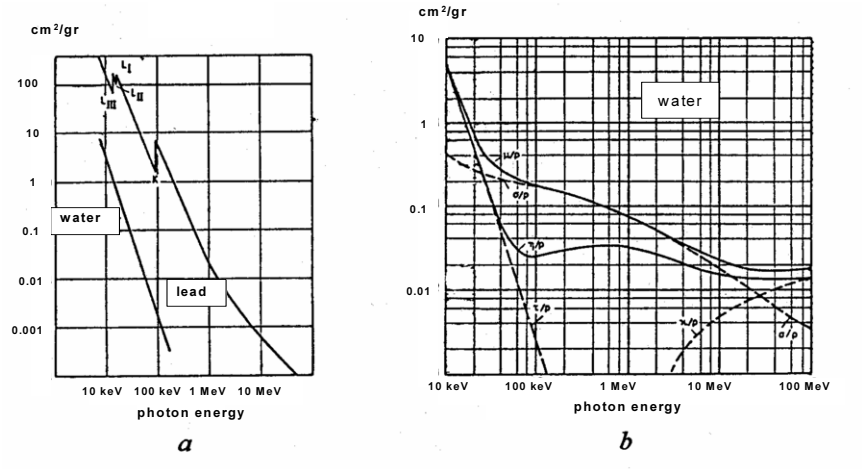


Figure 9. ***a***: Mass absorption coefficient (photoeffect alone) for lead and water.
b: Total mass absorption coefficient $\mu' = \mu/\rho$ (the definition of μ is given below, Beer-Lambert's law) as function of the photon energy, E . It is composed of the contribution of the photoeffect (τ/ρ), of the Comptoneffect (σ/ρ) as well as the one of pair production (κ/ρ). The latter effect is not considered here because for the production of an electron-positron pair an energy of at least $2\, m_0 c^2 = 1.022$ MeV is necessary (m_0 = electron mass, $9.11 \cdot 10^{-31}$ kg). Since this energy is much higher than what is used in medical x-ray imaging, this effect is not of importance here (the additional curve, η/ρ , in the graph, is related to the energy conversion).
In total, the relation holds $\mu' = \mu/\rho = \tau/\rho + \sigma/\rho + \kappa/\rho$

Since the Compton effect requires on the average a higher amount of photon energy, at lower energies the photoeffect prevails while at higher energies the Comptoneffect sets in.

The cross section of an atom with respect to a certain interaction process is defined as the (theoretical) area σ (in cm^2) perpendicular to the incoming radiation which – much like the target in a shooting range – is available for the initiation of the process in question. If the photon hits this area, the effect is triggered, if not, no interaction takes place. In order to obtain the specific absorption capacity (per unit of mass) of a material, the cross section has to be multiplied by the number of atoms per unit of mass. Accordingly, the mass absorption coefficient is defined as the total cross section per gram, *i.e.*,

$$\mu' = \sigma N_A / A \qquad (N_A \text{ Avogadro's number, } A \text{ atomic weight}) \tag{5}$$

Since N_A indicates the number of atoms per A grams of a substance ($6.023 \cdot 10^{23}$), the quotient N_A / A corresponds to the number of atoms per gram, *i.e.*, μ' has the dimension cm^2/g.

With regard to medical imaging, the dependence of the mass absorption coefficient on the atomic number, Z (= number of protons in the nucleus), is decisive. Because of this dependence, the various biological tissues in the body exhibit different x-ray absorption characteristics such that their appearance in the projection image (grey level) varies according to their chemical composition. In case of the photo effect, σ is

about proportional to Z^5, while for the Compton effect it is about proportional to Z. For both effects, however, the cross section decreases with increasing energy, *i.e.*, the penetrating power of the x-rays increases (Figure 9). According to the object to be imaged, therefore, the voltage of the x-ray tube is set.

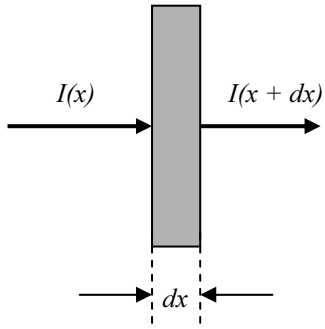
Because of the stochastic nature of the occurrence of the various effects, the attenuation of an x-ray beam when penetrating a thin sheet (thickness dx) is proportional to dx,

$$I(x + dx) - I(x) = -\mu\, dx$$

which yields

$$I(x) = I_0 \exp(-\mu x) \qquad (6)$$

(Beer-Lambert's law).
The linear attenuation coefficient, μ, is determined from the mass absorption coefficient introduced above according to $\mu = \mu' \rho$ (ρ = density), because the entire mass of the material in the layer dx has to be taken into account. In case of composite materials or mixtures (which is always given in biological tissues) the total mass attenuation coefficient μ'_{tot} can be determined from a weighted average (according to the density) from the individual coefficients of the various components.



I(x) : Number of particles per area and sec

As mentioned above, biological materials are always composed for various constituents in different concentration such that they can be distinguished according to their x-ray absorption properties (Figure 10). These effects are systematic and reproducible.

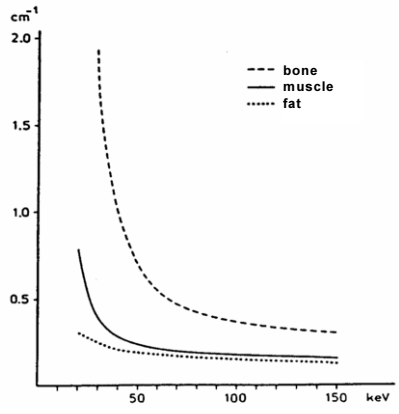


Figure 10. Dependence of the linear attenuation coefficient on the energy for various tissues

3. X-Ray Projection Imaging

3.1. Resolution of Imaging Procedures

From a technical point of view the resolution provided by an entire x-ray imaging chain (x-ray tube → image presentation) has to be considered. One has thereby to distinguish between

a) *spatial*
b) *contrast*
c) *temporal resolution.*

Only linear mappings are considered here.

a) *Spatial Resolution*: The *spatial* resolution of an imaging chain is often given in terms of the normalized intensity amplitude as a function of line pairs/mm (lp/mm), denoted as Modular Transfer Function (MTF). The practical procedure to measure the MTF for x-ray systems consists of imaging lead grids of variable spacing (Figure 11). The closer the lines are together, the smaller is the amplitude of the image and the average grey level increases until the lines cannot be discerned any more. The maximal amplitude (arbitrary unit 1) is defined as the maximal intensity step that can be represented. Typical MTF curves are shown in Figure 12.

Instead of the MTF, the line spread or point spread function (LSF, or PSF), respectively, are sometimes given. These functions indicate onto what image a line or point is mapped, in particular, how a point is blurred (Figure 13).

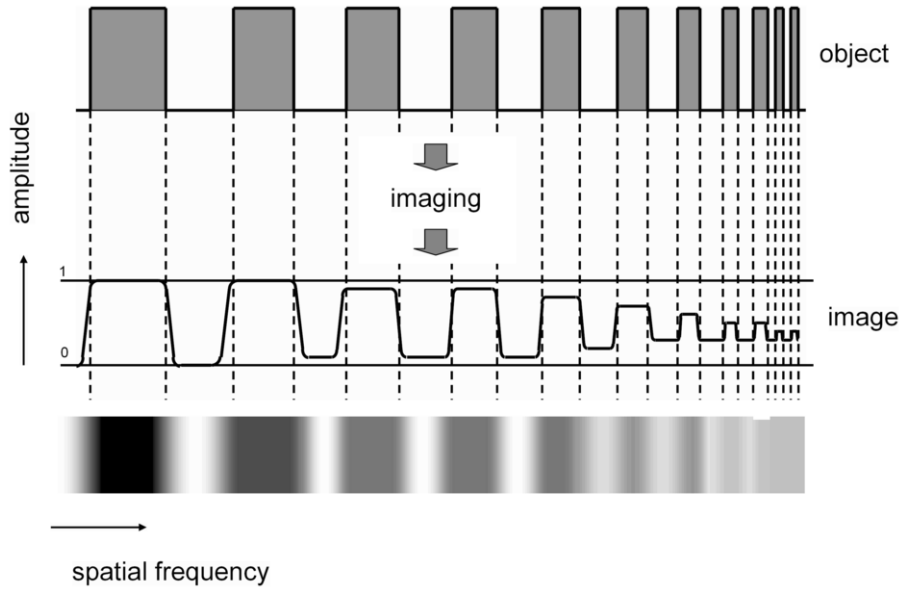


Figure 11. Imaging of a lead grid. With increasing spatial frequency (lp/mm) the amplitude of the image (grey level) decreases.

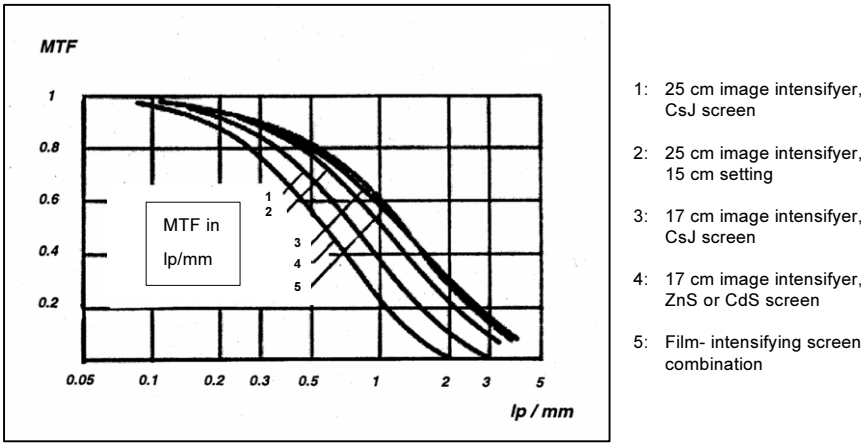


Figure 12. MTF of various x-ray systems (system definition see later)

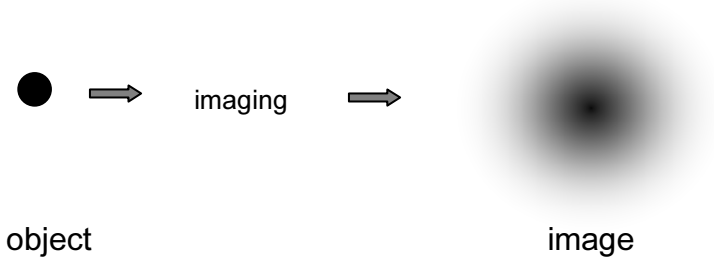


Figure 13. Principle of the point spread function (PSF)

In the Appendix 1, a mathematical derivation of the MTF is given. It can be shown that the MTF and LSF or PSF (in two dimensions) are related by a Fourier transform. Since every object (one or two dimensional) can be thought to be composed of lines (one-dimensional image) or points (two-dimensional image), in case of a linear mapping it is sufficient to know the LSF or PSF, respectively, to determine images. This procedure is often used in medical imaging.

b) *Contrast Resolution*: In order to assess the *contrast* resolution of x-ray projections, the contrast in the radiation behind an object has to be considered. As an example, we determine the intensity of an x-ray beam after passing through $d = 20\ cm$ of typical biological tissue:

$$I = I_0 \exp(-\mu d) , \quad \text{or} \quad I / I_0 = \exp(-\mu d) \tag{7}$$

With $\mu = 0.35\ cm^{-1}$ (average for biological tissue devoid of bone and air) we obtain a contrast of exp (-7), *i.e.*, about 1/1000. In the presence of bone and air (lung, intestine) the contrast is even considerably higher. Neither with photographic film nor with conventional TV such a contrast range can be recorded (film, as shown below, has a linear gray scale range of about 2.5 - 3, corresponding to a contrast range of about 1/300 to 1/1000, while conventional TV has less). Depending on whether, e.g., the lung, other soft tissues or bones are to be imaged, the radiologist has to choose the parameters anode voltage, anode current, exposure time in order to keep the desired contrast within the available range.

Digital imaging techniques allow to record images with an extended contrast range (12 – 16 bit resolution). Typical TV monitors, however, visualize only a limited grey level range of about 6 bit. Since images are stored in digital form, the visible contrast range can be selected and altered up and down in real time („windowing“).

c) *Temporal Resolution*: The *temporal* resolution is of importance whenever dynamic processes are to be investigated (phlebogram, ventriculogram, stomach-intestine passage, etc.). Cassette changers (see later) reach a frequency of up to about 6 images/sec which is sufficient, e.g., for phlebograms. In order to resolve a heart cycle adequately, up to 50 images/sec are necessary.

3.2. Photographic X-Ray Film

Diagnostic x-ray cause precipitation of silver bromide in a photographic emulsion such that x-ray intensity is reflected as grey level after development. A sharp image is only obtained, if the photographic layer is thin because of the mostly oblique incidence of x-rays in a projection arrangement. Besides, there is always undesired scattered radiation. Since the absorption in the photographic emulsion increases exponentially with the thickness (Beer-Lambert's law), in a typical x-ray film only about 2 % of the incident radiation is absorbed. If not highest resolution is necessary (e.g., mammography or dental applications) double-sided film is used in order to increase the thickness. Accordingly, a typical photographic x-ray film consists of a middle polyester carrier sheet (thickness 0.15 mm) to which on both sides a silver emulsion layer (thickness 7 – 20 μm) is attached by way of a thin (1μm) adhesive coating. An outer covering layer on both sides of the film serves as protection..

The low efficiency (requiring a correspondingly high radiation dose delivered to the patient) can be increased by adding intensifying screens on both sides of the film which is then called a cassette. This increases the absorption but decreases the resolution. The intensifying screen contains *e.g.* $CaWO_4$, a substance which exhibits blue fluorescence under the incidence of x-ray which increases film exposure (other fluorescent substances include Y_2O_2S, Gd_2O_2S, BaFCl). The efficiency can so be increased up to a factor of about 5.

The opacity $1/T$ (inverse transmittance, dimensionless) of a (processed) photographic sheet is defined as

$$1/T = (I_0 / I) \tag{8}$$

(I_0: incident, I: behind the sheet densitometrically measured intensity or brightness of a normalized white light source, unit Lumen [Lm])

The density S [= log ($1/T$)] as a function of the exposure (characterizing the quantity of light delivered to the emulsion) of photographic film follows the diagram shown in Figure 14.

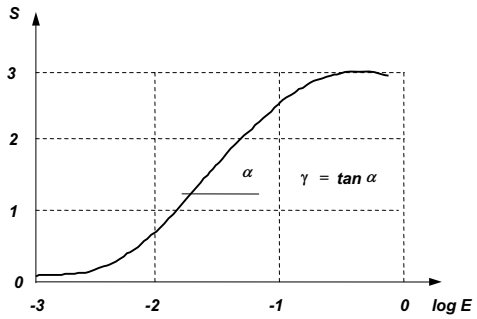


Figure 14. Typ. characteristic curve of photographic film [**S** : log(opacity), **E** : exposure (in meter-candle-sec)]

The γ value of a film is a measure of the maximum change of film density for a certain change of exposure. This corresponds to the part of the characteristic curve with the steepest slope, *i.e.*, at the point where the maximum derivative is found (= tangent of the inclination angle α). Typical photographic film has a linear range of about $\Delta S = 2.5$. The "steepness" γ is also used for other imaging modalities.

To eliminate scattered radiation, anti-scatter grids (Figure 15) are used. Such grids are characterized by the distance and thickness of the lamellae (typ. 25 – 50 lamellae/cm) as well as the grid ratio (height of the lamellae/distance, typ. 5 - 15).

X-ray films are mostly used for single-shot static exposures. If slow processes are to be examined (e.g., in case of a phlebogram), cassette changers are applied which allow to take up to about 6 images/sec. The advantage of this procedure consists of the high image quality which corresponds to the one of x-ray film. If higher frame rates are required (e.g., for imaging of the heart), other procedures have to be applied, in particular in order to avoid excessive radiation doses (see later, image intensifier).

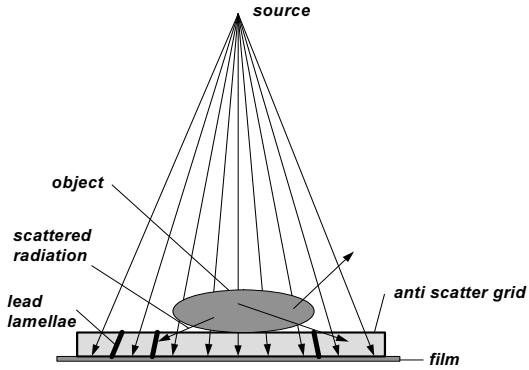


Figure 15. Anti-scatter grid arrangement

3.3. *Direct Digital X-Ray → Image Conversion*

Photographic film is an analog storage medium and preparation for computer analysis and archiving including networking necessitates off-line digitization. In addition, film development requires chemistry (cost and waste). A number of methods have therefore been developed which allow for a direct digital image acquisition and which are mostly in use today.

Digital Luminiscence Radiography (DLR), Storage Screen Radiography, Computer Radiography (CR, Figure 16): On a screen made of BaFBr:Eu^{2+} (Europium-doted barium halegonide) a latent image is created by irradiation (excited electron states) which is converted off-line into visible fluorescence by laser scanning. One of the advantages of this technology consists of its large dynamic range.

Selenium-based detectors: Prior to the exposure to x-ray, the surface of the detector is charged electrically. The photons are converted into electrical signals in that they create charges in the interior of the selenium layer which neutralize the surface charges. Again, a latent image is formed which is scanned off-line.

Flat screen detectors on the basis of amorphous silicium: On a silicium wafer a pixel matrix containing the necessary electrical components (transistors) is implemented which allows for a direct digital conversion of the charges created by the incident x-rays Because the absorption of x-ray by silicium is only weak, additional layers containing heavy atoms have to be overlaid on the silicium screen.

Scintillator-fiberoptic-CCD chip arrangement (in the future also the much simpler and less expensive CMOS technology will be used instead of CCD): For particular applications, (e.g., μCT) a scintillator screen serves for spatially resolved photon acquisition which is imaged on a CCD chip by fiber optic connection. The possibility thereby exists for image enlargement by cutting the fiber optic bundle obliquely.

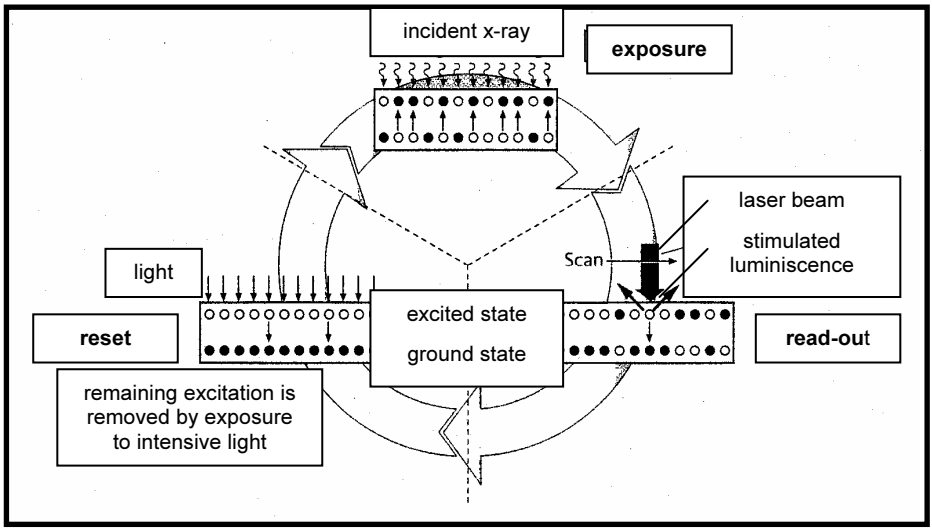


Figure 16. Principle of digital luminescence radiography

3.4. Image Intensifier

In a x-ray image intensifyer (Figure 17) the radiation image is converted into an equivalent electron image in the entrance screen. The conversion occurs in two steps: First, light flashes are created by x-ray photons in thin CsJ needle crystals (scintillator) which are about 0.5 - 1 cm long (good absorption of x-ray) and densely arranged perpendicular to the screen in order to suppress scattered radiation as well as lateral diffusion of light (veiling glare). These light flashes cause the creation of free electrons in the photocathode which is connected to the scintillating screen. In addition, on the entrance side of the scintillator a thin aluminum foil is located which reflects light back onto the photocathode. The electrons whose lateral diffusion in the photocathode adds to the veiling glare and has therefore to be suppressed, are accelerated by 25 kV typically and directed towards the exit screen. Thereby, an electrostatic field acts as an imaging electron-"optics" such that a bright image on the exit screen appears. The intensifying effect is due to the kinetic energy which the electrons obtain during acceleration. The exit screen is recorded digitally; in the past, video or cinéfilm was used (up to 50 images per sec). Thanks to the good absorption properties of the entrance screen, the efficiency of present intensifiers is about 40 - 50 %, i.e., substantially larger than film/intensifying screen combinations. The efficiency is given by the conversion factor G in (cd/m^2) / (mR/sec) whereby cd (candela) denotes the light intensity $dI/d\Omega$ (light intensity or light flux per solid angle), mR (milli-Röntgen), the radiation dose (old unit).

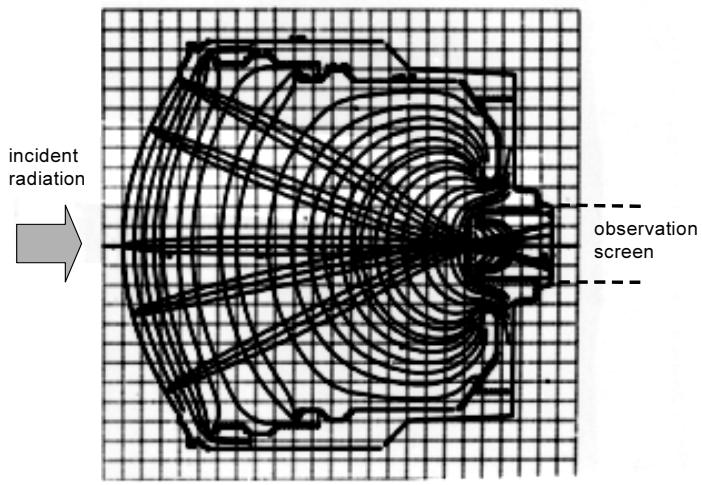


Figure 17. Cross section through image intensifyer showing electron paths and field lines

4. Image Subtraction Techniques, Digital Subtraction Angiography (DSA)

The lumina of blood vessels, stomach, intestine, ventricles of the heart or of the brain are only faintly or not visible on a routine x-ray projection image. In order to obtain a diagnostically useful visualization, the lumen in question has to be filled with an appropriate contrast agent (e.g., bariumsulfate in case of stomach-intestine, or a compound containing iodine for blood vessels or the heart which is well tolerated, i.e., which causes no immune reactions).

To image a section of an arterial vessel (e.g., if an aneurysm, a stenosis or a dissection is suspected) or of a ventricle of the heart it is necessary to administer the contrast agent at the very location under consideration (selective catheterization) in order to obtain sufficient image contrast. This procedure is associated with a considerable expense (catheter laboratory) since an arterial vessel has to be opened, furthermore, there is always the danger of an afterbleeding. In contrast, a venous infusion of a contrast agent is largely without problems. Although part of the contrast agent arrives at the location to be imaged on the arterial side also if a venous administration is made (e.g., through the vena cava), the resulting image contrast is insufficient because of the dilution of the contrast agent after the lung passage necessary to reach the arterial side and the subsequent distribution in the entire vasculature of the body. In addition, not only the location of interest contains therefore contrast agent, but all adjacent, overlying and underlying vessels including capillaries and veins such that the image of the vessel section in question is embedded and covered by other vessels.

With the aid of image subtraction techniques, however, an increase of contrast can be reached such that at least in the case of not too small arterial vessels a sufficiently good representation for diagnostic purposes can be obtained also with venous administration of contrast agent. Since for this technology on-line digital image treatment is applied, the method is called Digital Subtraction Angiography (DSA). (Note: Image subtraction can also be made easily with photographic film in that a negative can be obtained from contact exposure; for medical purposes, this procedure has been in use since about 1935). With DSA, an image called “mask” is first made (a digital projection image before the contrast agent is administered) and stored in the computer. Then, the contrast agent is applied transvenously, mostly through the vena cava. After about 20 sec, the exact time delay thereby depending on the location to be imaged, the contrast agent appears in diluted form at the desired section on the arterial side. Image sequences can now be acquired whereby the mask is subtracted in an on-line fashion (video, Figure 18). Providing that the patient has not moved (among other, breathhold is necessary) between the acquisition of the mask and the later images, only the difference between the images, i.e., the shadow of the contrast agent should be visible.

Overlay is still present, but above all, noise amplification occurs since images characterized by small differences associated with noise are subtracted and subsequently amplified. Since minimal exposure to radiation is desired, the discrete nature of x-ray may become apparent (“quantum noise”). This can partially be compensated by using more contrast agent and/or a higher radiation dose, both of course undesired (on the one hand, the incidence of immune reactions increases with an increasing amount of contrast agent, on the other, a higher radiation dose is undesired anyway). In spite of these drawbacks, the advantage of avoiding an arterial catheterization is by far dominating such that DSA is applied whenever possible.

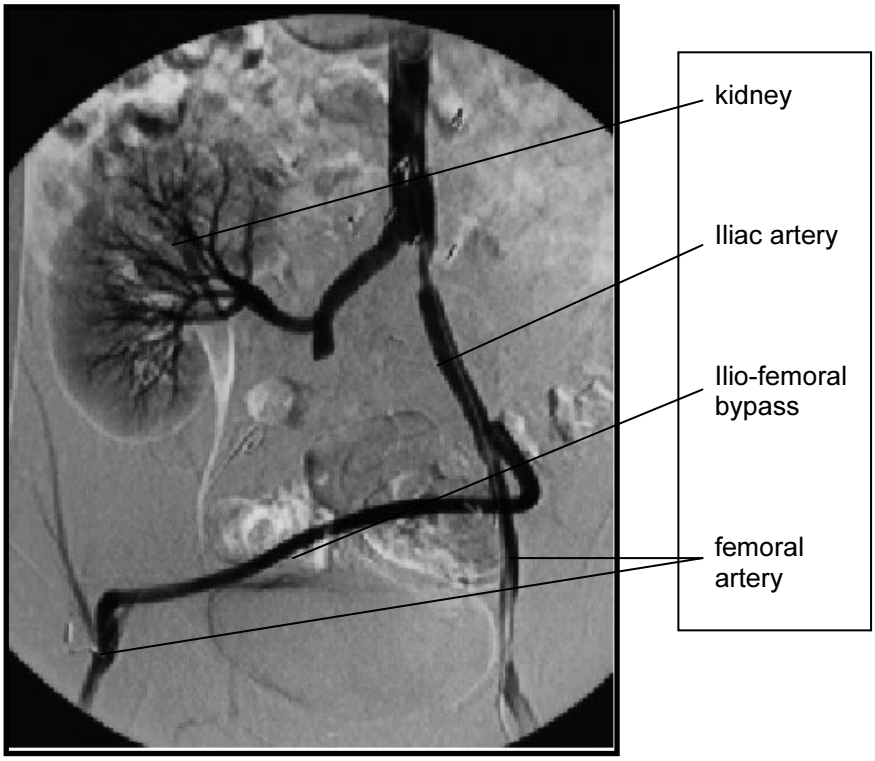


Figure 18. DSA image (real-time video subtraction) of a highly pathologic abdominal vessel situation

5. Computed Tomography (CT)

The „classical“ tomographic procedure in radiology was performed in such a way that the x-ray source (tube) and the photographic film were shifted during the exposure continuously in opposite directions. Only one plane is thereby imaged sharply, structures which are located underneath or above this plane are smeared. The images have therefore a low contrast, however. Computed tomography (derived from greek τομειν, to cut), in turn, was introduced in 1973 by G. N. Hounsfield in England (Figure 2). A. M. Cormack in South Africa had in fact already earlier investigated methods which allowed to reconstruct objects from a series of suitably chosen x-ray projections (Hounsfield made his later work however independently from Cormack's results). Since at the time when Cormack made his investigations there were no reasonably priced computer-assisted imaging methods available, he could not represent the results of his calculations as true tomograms, i.e., images of cross sections, but just in the form of numbers and curves; accordingly, his work had no practical consequences (nevertheless, both researchers received the Nobel prize in 1979). What both researchers did not know, however, was the fact that the mathematical problem of

calculating a *n*-dimensional object from *n-1* dimensional projections had been solved much earlier by the Austrian mathematician Johann Radon (1887 – 1956) already in 1917. This was recognized only after CT had been introduced into clinical routine. Yet, this was not of importance because Radon's formal mathematical solution was found to be unsuitable for practical applications for numerical reasons.

The method of CT consists of the acquisition of a sequence of x-ray projections of a predetermined cross section of the body under different angles. In scanners of the first generation, a single projection was taken by translating a collimated x-ray beam ("pencil beam") over the chosen cross section. This procedure required a synchronous movement of the source (x-ray tube) and the detector. Later, the "fan" beam technology was introduced whereby an entire projection is taken at the same time with the aid of a detector array (consisting, e.g., of 512 detectors) without translation of the x-ray source. Subsequent rotation of the x-ray source and detector array around 180° in equally spaced angle intervals (e.g., 180°/512) further projections are taken (Figure 19). A total set of (one dimensional) projections then allows to reconstruct the interior of the cross section, i.e., the two-dimensional distribution of the attenuation coefficients $\mu(x,y)$ (x, y are Cartesian coordinates in the cross section) The image is represented in the form of a grey level distribution $D(x,y)$. In present scanners, the fan beam – detector array is rotated and advanced in a spiraling motion around the patient such that entire spatial body sections can be imaged within a few seconds (Figure 20).

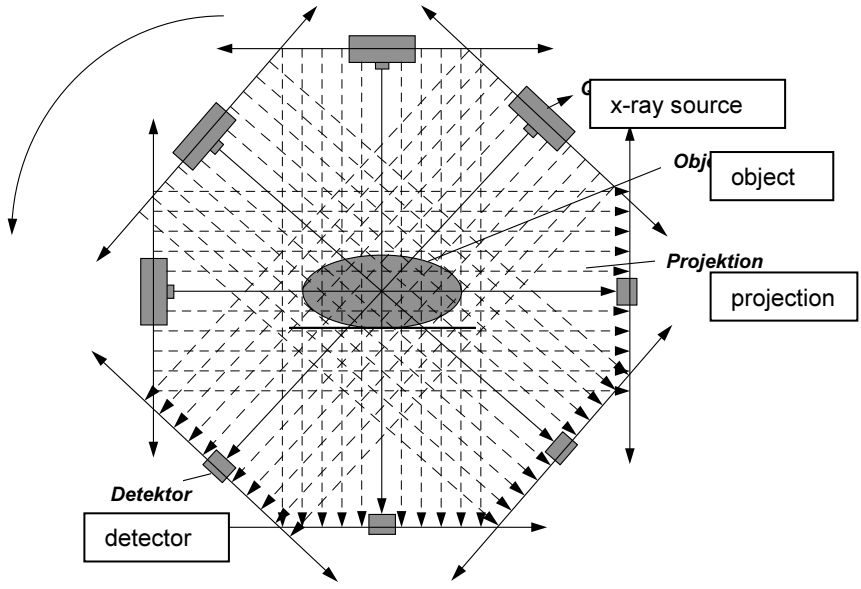


Figure 19. Principle of first generation CT scanner: A narrow x-ray beam (pencil beam) is generated by collimators at the x-ray source as well as at the detector (the latter primarily suppresses scattered radiation). One-dimensional projections are obtained by translating the unit (embedded in a gantry) over the object. By successive rotation of the gantry after each translation by, e.g., $\pi/512$, a complete set of projections is recorded.

For simplicity, we consider the situation of a first-generation translation-rotation scanner (Figure 21). The beam intensity, $I_\varphi(r)$, which is recorded in a direction perpendicular to a fixed angle φ as function of the linear coordinate, r, is obtained as

$$I_\varphi(r) = I_0 \, exp\left[- \int_{-s_0}^{s_0} \mu\left(s\{r,\varphi\}\right) ds \right] \tag{9}$$

where I_0 denotes the intensity of the unattenuated beam (before impinging on the object) and $s[r,\varphi]$ is the line defined by r and φ between source and detector in the x - y plane. The integration is performed along the line s from $-s_0$ to $+s_0$ (distance between source and detector).

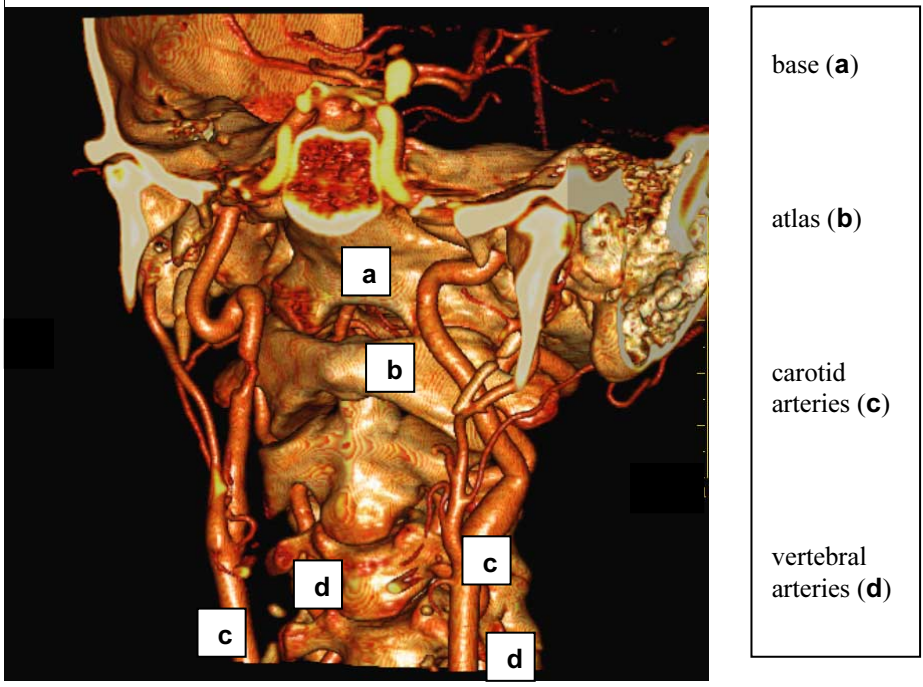


Figure20. Spiral CT image of the base of the skull (after segmentation)

As projection $P_\varphi(r)$ the quantity

$$P_\varphi(r) = log\left[I_\varphi(r)/I_0\right] = - \int_{-s_0}^{s_0} \mu[s(r,\varphi) ds] \tag{10}$$

is defined. The task now consists of the determination of the linear attenuation coefficients, $\mu(x,y)$, from a complete set of projections $P_\varphi(r)$.

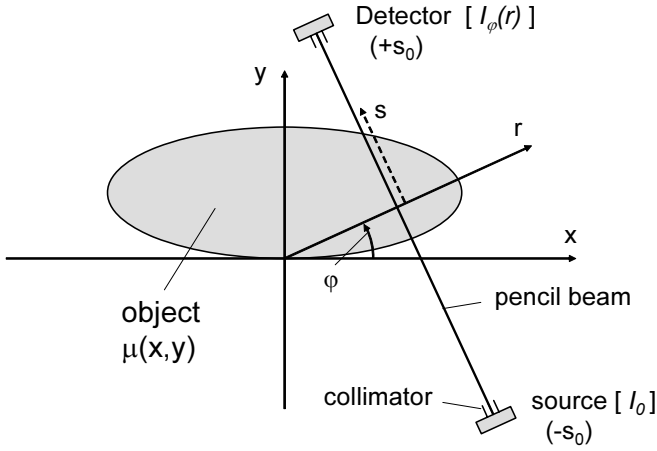


Figure 21. Principle of 1st generation translation-rotation scanner

As mentioned above, Radon found an analytic solution of the problem, which is unsuited for a numerical application, however. Among the various methods which have been evaluated for the calculation of the grey scale image $D(x,y)$ derived from the attenuation coefficients $\mu(x,y)$, the method outlined in the following called convolution - backprojection has proven to be particularly useful.

A simple backprojection without further pre- and postprocessing of the data is performed by distributing the entire value of the integral $P_\varphi(r)$ uniformly along the projection path, s (Figure 22). As an example, we consider the projection of a cylindrical object which is scanned perpendicularly to its long axis. If it is assumed that (*i*) the radius of the object $\rightarrow 0$, (*ii*) the number of projections (angles) $\rightarrow \infty$, and (*iii*) each one-dimensional projection is continuous, i.e., it consists of infinitely many points, theoretically, the image of a point results (point spread function, PSF). The PSF is found to be proportional to *1/r* , since along the perimeter of every concentric circle around the point under consideration the same amount of backprojection is accumulated.

According to chapter 3.1. it suffices to know the PSF associated with the imaging procedure in order to be able to determine the image (*i.e.*, the reconstruction) of any two dimensional object. The reconstruction problem would therefore be solved, if the "corrected" PSF was known, more precisely, if the mathematical procedure to be applied in order that a point appears as a point without artifacts was known. Accordingly, the artifacts present in the simple backprojection procedure, in particular the *1/r* dependence, have to be compensated.

This is achieved by applying a prefilter (mathematical convolution) to the projections. It can be shown (Appendix 2) that a convolution of the form

$$P'_\varphi(r) = \frac{1}{2\pi^2} \int_{-\infty}^{\infty} P^*_\varphi(u)\, e^{iur} |u|\, du \qquad (11)$$

yields the desired result. The function $|u|$ thereby compensates the *1/r* dependence as well as the fact that the Fourier space associated with the scanned area is not uniformly sampled (Appendix 2, this follows from the Fourier-slice theorem): High spatial frequencies are sampled less densely than low frequencies.

Particular problems which have to be given careful attention in x-ray CT machine design are related to:

Beam hardening: X-ray sources as used in clinical CT are not monochromatic (see paragraph 1, Generation of X-Ray). Since x-ray absorption is energy-dependent, the spectrum of the radiation changes continuously when traversing an object. This effect is nonlinear and object dependent and can therefore not be corrected mathematically. Careful, object-dependent calibration procedures are necessary for highest precision.

Partial volume effect: For practical computational reasons, the area of interest has to be subdivided into (usually rectangular) volume elements. Oblique rays cover most such elements only partially, however. For each direction and ray, the partial volumes covered by the rays have to be determined and taken into account in the reconstruction process.

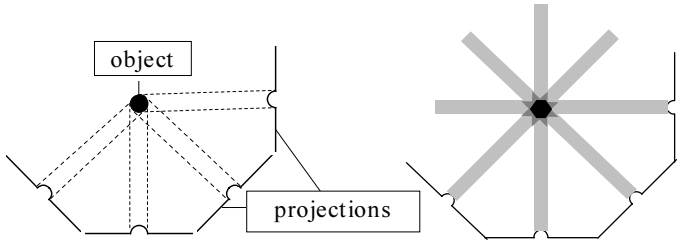


Figure 22. Schematic representation of four projections (top) of a cylindrical, homogeneous object and corresponding simple backprojection (bottom). The measured projection values are thereby distributed homogeneously ("backprojected") over the projection path. The resulting "reconstruction" is characterized by star-like artifacts, moreover, the image of the cylindrical cross section extends over the entire scanned area. If the radius of the cylinder is assumed to approach zero and the number of projections is increased to infinite, the point spread function (PSF) is theoretically obtained.

Scatter: Secondary radiation due to Compton scattering occurs when an object is irradiated. This secondary radiation appears as noise on the detector. Collimators can be used to prevent most scattered rays from reaching the detector because Compton scattering is omnidirectional in contrast to the primary radiation.

First generation scanners are no longer in use, a number of developments were made in order to improve speed and resolution. In particular, various filters (instead of $|u|$) are being implemented depending on the application, e.g., according to Ramachandran and Lakshminarayanan (see the literature listed in Further Reading). Speed was increased, first, by using a fan-beam x-ray geometry, later, cone-beam irradiation and 2D detectors were introduced in order to allow for scanning volumes. With helical (sometimes also referred to as spiral) scanning technology, finally, extended volumes can be scanned within a few seconds. High-precision micro-CT

scanners are furthermore being applied in the analysis of biological samples (Figure 23). It should be noted in this context that there are also significant non-medical applications of CT, in particular for material testing purposes.

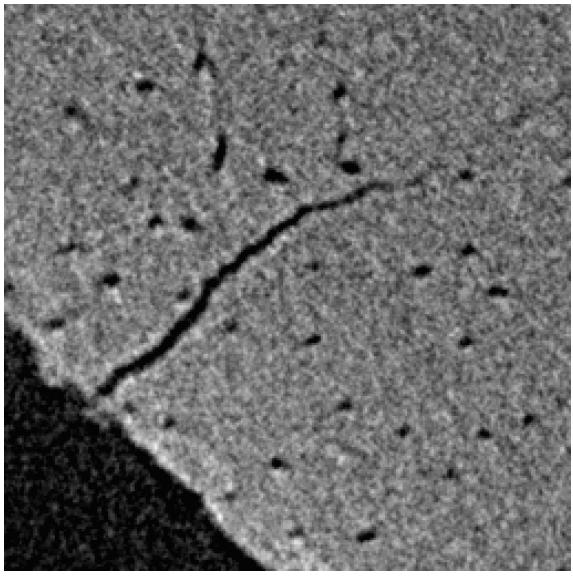


Figure 23. Micro-CT measurement of cortical bone. The long cleft appearing in the middle has a width of 3 μm. The lacunae (openings where the osteocytes are housed) can furthermore be seen.

Further Reading

[1] Elliott K. Fishman and R. Brooke Jeffrey Jr., *Spiral CT: Principles, Techniques and Clinical Applications* (2nd edition), Lippincott-Raven Publishers, Philadelphia, PA (1998)
[2] Marc Kachelriess, *Clinical X-Ray Computed Tomography*, in: Wolfgang Schlegel, Thomas Bortfeld and Anca-Ligia Grosu, New Technologies in Radiation Oncology, Springer, Berlin, 2006
[3] Herman GT, Correction for Beam Hardening in Computed Tomography, *Phys. Med. Biol.* **24** (1979), 81-106
[4] Yu Zou, Emil Y Sidky, Xiaochuan Pan, Partial Volume and Aliasing Artefacts in Helical Cone-Beam CT, *Phys. Med. Biol.* **49** (2004), 2365-2375
[5] Thorsten M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, Springer, Berlin 2008

Appendix 1: The Modular Transfer Function (MTF)

The spatial resolution of imaging systems is often given in the form of the Modular Transfer Function (MTF, i.e., grey-level contrast in the image space as function of the number of line pairs/mm in the object space; we assume for simplicity the one dimensional case, object space coordinate x). This concept can be introduced as follows: Each (one dimensional) object, characterized by a grey-level or intensity function $O(x)$ can be thought to be composed of lines

$$O(x) = \int_{-\infty}^{\infty} O(X)\,\delta(x - X)\,dX \tag{1}$$

by making use of the Dirac function $\delta(x - X)$. We assume now that the Line Spread Function (LSF), $L(x')$, which denotes the mapping of a line at location $x = 0$ in the object space, i.e., of $\delta(x)$, onto the image space with coordinate x', is known. A line at location X in the object space with intensity $O(x)$ will then yield an image

$$O(x)\,\delta(x - X) \;\rightarrow\; O(X)\,L(x' - X) \tag{2}$$

In case of an optical imaging system with lenses, *e.g.*, the LSF will be a diffraction pattern, for x-ray projection images, the line will be blurred[2].

A general object, $O(x)$, can be decomposed into Fourier components (providing that it fulfils the usual mathematical conditions) according to

$$O(x) = \int_{-\infty}^{\infty} \left[a(k)\,sin(kx) + b(k)\,cos(kx)\right] dk \tag{3}$$

with $a(k)$, $b(k)$ amplitudes
$k = 2\pi / \lambda$ wave number, wavelength λ

One component

$$J_k(x) = a(k)\,sin(kx) = a(k) \int_{-\infty}^{\infty} \delta(x - X)\,sin(kX)\,dX \tag{4}$$

is imaged by way of the LSF into the image space as $J_k'(x')$ in the form

$$J_k'(x') = a(k) \int_{-\infty}^{\infty} L(x' - X)\,sin(kX)\,dX \tag{5}$$

Upon performing the transformation $x'' = x' - X$ and making use of the addition formulas for circular function, one arrives at

$$\begin{aligned} J_k'(x') &= a(k)\left[sin(kx') \int_{-\infty}^{\infty} L(x'')\,cos(kx'')\,dx'' \;-\; cos(kx') \int_{-\infty}^{\infty} L(x'')\,sin(kx'')\,dx'' \right] \\ &= a(k)\left[sin(kx')\,a_1 \;-\; cos(kx')\,a_2\right] \end{aligned} \tag{6}$$

with

[2] For two dimensional images, the Point Spread Function (PSF) is used instead of the line spread function, in that a two-dimensional image can be decomposed into points, using the Dirac function in both dimensions.

$$\eta(k) = \left\{ \left[\int_{-\infty}^{\infty} cos(kx'')\, L(x'')\, dx'' \right]^2 + \left[\int_{-\infty}^{\infty} sin(kx'')\, L(x'')\, dx'' \right]^2 \right\}^{1/2} \qquad (7)$$

and using the addition property of circular functions once more, the result can be written as

$$J_k'(x') = a(k)\,\eta(k)\, sin(kx' - \varphi) \qquad (8)$$

The Fourier component $J_k(x) = a(k)\, sin(kx)$ is therefore imaged onto the image space as harmonic function of the same frequency, k, but displaced by the (spatial) phase φ and with the amplitude $a(k)\,\eta(k)$. The relation holds for every k.

The function $\eta(k)$ is denoted as Modular Transfer Function (MTF), and it is seen that the relation holds $0 \le \text{MTF} \le 1$. This function describes the dependence of the amplitude of the original intensity function in the image space (which is always reduced) as function of the spatial frequency k. Since the relation holds for every k, one can conclude that

$$F\,(\text{image}) = \text{MTF} \cdot F\,(\text{original}) \qquad (F: \text{ Fourier transformation}) \qquad (9)$$

respectively, that the image corresponds to the convolution of the original with the MTF according to the convolution theorem.

The relation with the LSF (for two-dimensional images with the PSF) is seen from the definition of the MTF.

Appendix 2: The Convolution / Backprojection Method

It was seen in Chapter 5, that a simple backprojection leads to characteristic artifacts; in particular, if the projections of a point object are backprojected, a *1/r* dependence results. By prefiltering the projections, the artifacts can be compensated (or at least substantially reduced). This aim is best reached in an indirect way, since on the one hand the function *1/r* exhibits a singularity and on the other it has to be noted that that the real scanning procedure (translation - rotation) is made in finite steps which – as will be seen below – leads to an inhomogeneous sampling in the Fourier space, or, in the real space, there are significant deviations from the theoretical *1/r* - dependence.

First, the projection $P_\varphi(r)$ is Fourier transformed

$$F\{P_\varphi(r)\} = P^*{}_\varphi(u) = \int_{-\infty}^{\infty} P_\varphi(r)\, e^{-iur}\, dr = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mu(s[r,\varphi])\, e^{-iur}\, dr\, ds \qquad (1)$$

(As mentioned above, the integration over *s* is extended to infinite for formal reasons, as outside the source - detector area we can assume $\mu = 0$.) Upon transformation of the variables $r = x\, cos\, \varphi + y\, sin\, \varphi$, $s = -x\, sin\, \varphi + y\, cos\, \varphi$ one obtains

$$
\begin{aligned}
-P^*(u) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu(s[r,\varphi])e^{-iur}\,dr\,ds \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu(x,y)e^{-iu[x\cos\varphi + y\sin\varphi]}\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu(x,y)e^{-ix\,[u\cos\varphi]}\,e^{-iy[u\sin\varphi]}\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mu(x,y)e^{-ix\,p}\,e^{-iy\,q}\,dx\,dy\Big|_{p=u\cos\varphi,\,q=u\sin\varphi}
\end{aligned}
\tag{2}
$$

This implies that the Fourier transform of the projection $P_\varphi(r)$ corresponds to the two-dimensional Fourier transform of $\mu(x,y)$ on a straight line which runs (in the two-dimensional Fourier space (p, q) under the angle φ through the origin (Fourier - slice - theorem). The Fourier space is therefore sampled inhomogeneously in a star-like fashion in case of translation-rotation scanning (Figure 24). In particular, the high frequencies (image sharpness !) are covered less densely than the low frequencies. In order to recover the Fourier transform $P^*(p,q)$ of $\mu(x,y)$ with uniform density in the entire Fourier space of interest from the Fourier-transformed projections $P^*_\varphi(u)$, interpolations are necessary.

From the relation (1) can be seen that the reconstruction problem, i.e., the calculation of $\mu(x,y)$ can be solved by a reverse Fourier-transformation (after having determined the Fourier-transformed projections and interpolated $P^*(p,q)$). A two dimensional reverse Fouriertransform has however to be executed numerically for this purpose which is an unsuitable procedure. This can be prevented in the following fashion.

From the formula for the reverse Fourier-transformation

$$
-\mu(x,y) = \frac{1}{2\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}P^*(p,q)e^{ixp}\,e^{iyq} \tag{3}
$$

one obtains after having transformed the variables according to

$$
p = u\cos\varphi,\quad q = u\sin\varphi,\quad dp\,dq = u\,du\,d\varphi \tag{4}
$$

$$
-\mu(x,y) = \frac{1}{2\pi^2}\int_{0}^{2\pi}\int_{0}^{\infty}P^*(u,\varphi)e^{iu[x\cos\varphi + y\sin\varphi]}\,u\,du\,d\varphi \tag{5}
$$

Upon making use of the symmetry of $P^*(u, \varphi)$, viz.,

$$
P^*[u, \varphi + \pi] = P^*[-u, \varphi] \tag{6}
$$

since the same projection is recorded after a rotation of 180°), this expression can be written as

$$
-\mu(x,y) = \frac{1}{2\pi^2}\int_{0}^{\pi}\int_{-\infty}^{\infty}P^*_\varphi(u)e^{iu[r]}\,|u|\,du\,d\varphi \tag{7}
$$

with $x \cos\varphi + y \sin\varphi = r$. By writing

$$P'_\varphi(r) = \frac{1}{2\pi^2} \int_{-\infty}^{\infty} P^*_\varphi(u)\, e^{iur}\, |u|\, du\, d\varphi \tag{8}$$

and

$$-\mu(x,y) = \int_0^\pi P'\varphi(r)\, d\varphi = \int_0^\pi P'\varphi(x \cos\varphi + y \sin\varphi)\, d\varphi \tag{9}$$

one finds that this corresponds to a filtering or a weighting of the Fourier-transformed projections $P^*_\varphi(u)$ with $|u|$ before execution of the reverse transform, which now is restricted to one dimension. The subsequent integration over the angle φ is a summation over all weighted reverse-transformed projections and is equivalent to a backprojection.

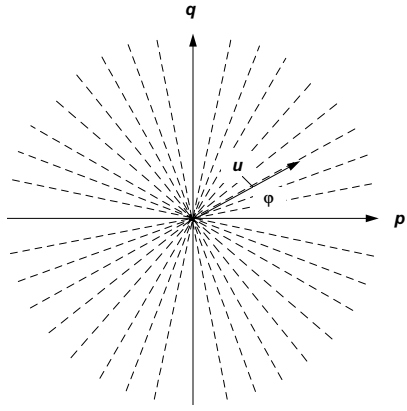


Figure 24. Fourier-slice theorem; sampling of Fourier space

On the basis of (2) and (3) the question relating to the PSF can finally be answered. According to chapter 3.1. the convolution theorem reads (" * " denotes a convolution)

$$\text{Image} = \text{PSF} * \text{Object}, \quad \text{resp.} \quad F\{\text{Image}\} = F\{\text{PSF}\} \cdot F\{\text{Object}\} \tag{10}$$

Here, "Image" denotes the reconstruction which is provided by the PSF and "Object" the true distribution of the linear absorption coefficients $\mu(x,y)$. The (two-dimensional) Fourier-transform of $\mu(x,y)$ is known, if measurements are made under all angles and the entire two-dimensional Fourier-space (p,q) is interpolated (it is however inhomogeneously covered by the one-dimensional Fourier-transforms $P^*_\varphi(u)$). Eq. (2) indicates that the factor $|u|$ compensates the artifacts. The inverse transform of $|u|$ would therefore correspond to the corrected PSF. Upon performing

an inverse transform of the function $|u| \exp(-\varepsilon |u|)$ for $\varepsilon \to 0$ ($|u|$ itself cannot be transformed !), a dependence $\sim 1/r^2$ is found. This can be interpreted such that the $1/r$ – dependence of the PSF (in simple backprojection) is additionally compensated by a further $1/r$ factor which is due to the inhomogeneous coverage of the Fourier space which also is characterized by a $1/r$ dependence.

If therefore for the backprojection (3) not the original projections $P_\varphi(r)$, but the corrected "projections" $P'_\varphi(r)$ are used, the desired results are obtained. In the object space, the procedure involves a convolution such that the method is denoted as convolution-backprojection method. It is the method of choice in computed tomography technology. If further effects which cause image deterioration are taken into account by adaptation of the kernel function $|u|$ still further improvements can be achieved.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-302

V.3. Basic Elements of Nuclear Magnetic Resonance for Use in Medical Diagnostics:

Magnetic Resonance Imaging (MRI)
Magnetic Resonance Spectroscopy (MRS)

Peter F. NIEDERER
Institute for Biomedical Engineering
ETH Zurich and University of Zurich
Gloriastrasse 35
CH-8092 Zurich, Switzerland

Abstract. Magnetic Resonance Imaging (MRI) has established itself as a major imaging modality in life science research and clinical practice. It is characterized by high spatial resolution, high soft tissue contrast, non-invasiveness, and universal applicability in terms of orientation and location of imaging areas. The procedure allows furthermore the investigation of physiological and pathophysiological processes, in particular in combination with magnetic resonance spectroscopy (MRS). MR methodology is not exhausted, new procedures and areas of application develop widely in life science and medicine. This article is limited to basic physical aspects.

Keywords. Magnetic resonance imaging, magnetic resonance spectroscopy, diagnostic imaging

Introduction

Nuclear Magnetic Resonance (NMR) was discovered and documented concurrently by F. Bloch (Stanford University) and E. Purcell (Harvard University, both Nobel laureates in 1952) in the year 1946. Since then, analytic laboratory methods which are based on NMR have become indispensable tools in solid state physics as well as in chemistry. Thanks to the availability of large and extremely strong magnets (super conductivity), powerful spectrum analyzers and computers the NMR-technology could be introduced in clinical medicine during the early seventies. Of particular importance in this context is the pulse-spectrometry developed by R Ernst (Nobel laureate in 1991). Further Nobel laureates were P.C. Lauterbur and P. Mansfield (2003) who made major contributions in the development of medical MR, in particular of imaging methods. The first cross section of a human thorax was however presented by R. Damadian et al. in 1977 (Figure 1).

The principal application of NMR in clinical medicine is imaging (MRI, Magnetic Resonance Imaging). For the purposes of biological and medical research, spectroscopy (MRS, Magnetic Resonance Spectroscopy) in combination with imaging

is also of interest. In comparison with other clinical imaging methods (CT, ultrasound), MRI/MRS has a number of distinct advantages and some disadvantages.

Compared to CT, MRI yields a better contrast of soft tissues, allows the imaging of cross sections which exhibit an arbitrary spatial orientation and there is no ionizing radiation - the strong constant and time-dependent magnetic fields have not been found to cause negative effects in humans. Some disadvantages, however, derive from the high cost of the technology and installation, the access surveillance (no ferromagnetic objects must come close to the magnet, patients with pacemakers, surgical clips and some types of implant cannot be treated or, at least, extreme care has to be exercised). Claustrophobia is sometimes a problem. For the investigation of the skeleton (bone) CT remains the imaging method of choice, since mineralized, hard tissue cannot be imaged directly with MRI. Likewise, in emergency units where rapid examinations are required, CT scanners are advantageous because of speed, accessibility as well as the fact that hemorrhages show a good CT contrast due to the iron content of hemoglobin.

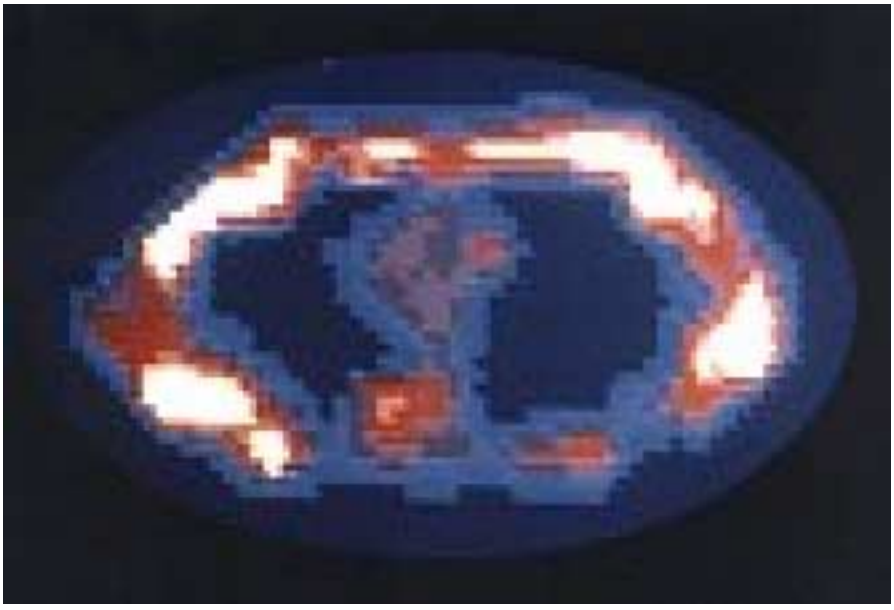


Figure 1. First in vivo human MR image (cross section, thorax) made by Damadian et al. 1977

Ultrasound is even more rapid and real-time imaging can be performed. In addition, ultrasound equipment is considerably less expensive and smaller that MRI or CT installations. Yet, MR image quality is largely superior compared to ultrasound, moreover, ultrasound can only be applied through acoustic windows and bony structures and air-filled cavities are almost intransparent to ultrasound.

Of great future interest is real-time MR imaging which is under development and may become feasible within the next few years. A further interesting development is associated with “open” systems which allow direct patient access and provide the possibility for MRI-guided surgery.

MRI and MRS have established themselves as significant research tools also in a wide range of non-clinical applications relating to human behavior in general and psychology in that brain activity can be monitored under various real and artificial exposure situations. Due to its noninvasive character, MR technology is furthermore widely used in animal research.

At this time, MR technology in the life sciences is by far not exhausted and new applications emerge constantly. Accordingly, the literature on MRI/MRS is vast; this introductory text is limited to basic physical aspects which are presented in a partly simplified form. Quantum mechanics and thermodynamics are only used as far as necessary.

1. Nuclear Spin Resonance

1.1. Intrinsic Angular Momentum (Spin)

Elementary particles (protons, neutrons, electrons, etc.) have an intrinsic angular momentum (spin), whose size is specific for each particle and may in particular be zero. It has no classical analogon, i.e., it cannot be 'explained' on the basis of classical mechanics. In fact, relativistic quantum field theory is necessary to fully appreciate spin systems.

Atomic nuclei consist of protons and neutrons and therefore also possess a spin. This spin depends on the type of nucleus and derives in a nontrivial manner from the internal nuclear structure. Of particular interest for medical applications are the stable nuclei ^{1}H, ^{4}N, ^{19}F, ^{23}Na, ^{31}P, as well as the isotope ^{13}C, which occur naturally in the human body in various concentrations and which all have a spin different from zero.

According to the laws of quantum mechanics the values of the modulus of angular momentum vectors, $\vec{J}$, can in stationary states assume only integer multiples of Planck's constant $h/2\pi = \hbar$ ($h = 6.63 \cdot 10^{-34}$ W·sec^2) with respect to a fixed axis. This includes nuclear spins as well. As a reference axis, one usually chooses the z – component of a Cartesian coordinate system x, y, z, such that for the z - component of the angular momentum, J_z, follows

$$J_z = \hbar m \qquad (m \text{ integer, denoted as magnetic quantum number}) \tag{1}$$

The possible values of $|m|$ are furthermore limited; the relation holds

$$m = I, I-1, I-2, \ldots, -I \tag{2}$$

I represents the spin quantum number. It is specific for each nucleus and can be integer or half-integer. A nucleus can therefore assume $2I + 1$ stationary states with respect to angular momentum. In the following, the nucleus ^{1}H with $I = 1/2$ will mainly be considered (this nucleus is abundant in the human body), such that in this case

$$J_z = \pm \hbar/2 \tag{3}$$

As atomic nuclei are electrically charged, the spin is associated with a magnetic dipole moment. This amounts to

$$\vec{\mu} = \gamma \vec{J} \tag{4}$$

γ is denoted as gyromagnetic ratio and is likewise specific for each type of nucleus. For ^{1}H , $\gamma = 4.26 \cdot 10^7$ (T·sec)$^{-1}$ (the unit T stands for „Tesla“, the unit used for the magnetic $\vec{B}$ - field).

In a magnetic field with flux density $\vec{B}$ (also denoted as magnetic induction) the axis of reference is given in a straightforward way by the direction of the vector-field which is identified with the z – axis. Accordingly,

$$\mu_z = \gamma \hbar m ; \quad \text{or, in case of } ^1\text{H}, \text{ we have } \quad \mu_z = \pm \gamma \hbar / 2 . \tag{5}$$

Following the laws of quantum mechanics, physical quantities assume a defined value only if there is an interaction with the surroundings. Here, this interaction is provided by the magnetic field $\vec{B}$. In the absence of such a field, there is no *a priori* direction defined, the magnetic quantum number undefined and all corresponding states are degenerated, i.e., all states possess the same energy. (This also explains the term “magnetic quantum number” for m : The spin manifests itself only in the presence of a sufficiently strong interacting field. The gravitation of the earth or its magnetic field are far too weak to have a noticeable effect.)

The energy of a magnetic dipole, $\vec{\mu}$, in the field $\vec{B}$ reads (Figure 2)

$$E = -\left(\vec{\mu}, \vec{B}\right) \quad , \text{ resp. } \quad E_m = -\gamma \hbar B m \tag{6}$$

in the states $m = \pm 1/2$, if the potential energy of the system in the field-free space is normalized to zero

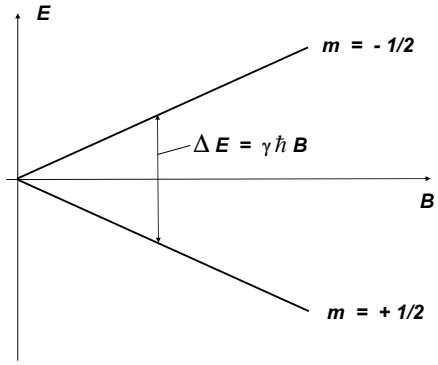


Figure 2. Energy of the states *m = +/- 1/2* as function of the field *B*

In thermal equilibrium, the state associated with lower energy (*m = +1/2,* spin parallel and oriented along the same direction as the field) is more populated than the state with higher energy. The population density thereby obeys the Boltzmann distribution according to the probability

$$W = N\,exp\left[\frac{-E}{k_B\,T}\right] \text{ (general), resp. } W_m = N\,exp\left[\frac{\gamma\,\hbar\,B\,m}{k_B\,T}\right] \tag{7}$$

with W_m probability for the system being in state m,
N normalization factor (sum of all probabilities = 1),
k_B Boltzmann constant ($1.38 \cdot 10^{-23}$ W sec / °),
T absolute temperature

In the case considered here, the system has only two states ($m = \pm 1/2$). Upon calculating the Boltzmann factor, $exp\left[\frac{\gamma\,\hbar\,B\,m}{k_B\,T}\right]$, one finds a value around $exp\left[10^{-6}\cdot B\,m\right]$ (at T = 300°K), i.e., the populations of the two states differ little even in the presence of high fields. The normalization factor, N, can therefore be approximated by 0.5, the population difference Δn can furthermore be determined by linearization as

$$\Delta n = n_{+1/2} - n_{-1/2} \approx n\,\gamma\,\hbar\,B/\left[2\,k_B\,T\right] \tag{8}$$

n thereby denotes the number of magnetic moments per unit volume. The total nuclear magnetization $\vec{M}$ results as the sum of all dipole moments per unit volume

$$\vec{M} = \Sigma\,\vec{\mu}_i \quad , \quad \left|\vec{M}\right| = \Delta n\left|\vec{\mu}_z\right| = n\,\gamma^2\,\hbar^2\,B/\left[4\,k_B\,T\right] = \chi_0(T)B \tag{9}$$

χ_0 is denoted as nuclear magnetic susceptibility.

1.2. Bloch's Equations

Systems of identical charged particles i with angular momentum $\vec{J}_i$, that are exposed to a magnetic field $\vec{B}$, are subjected to a moment according to $\vec{\mu}_i \times \vec{B}$. The macroscopic magnetisation $\vec{M}$ executes a motion which can be described by the classical angular momentum equation,

$$\frac{d\,\Sigma\,\vec{J}_i}{d\,t} = \Sigma\,\vec{\mu}_i \times \vec{B} = \vec{M} \times \vec{B} \tag{10}$$

respectively, $$\gamma\,\Sigma\frac{d\,\vec{J}_i}{d\,t} = \Sigma\frac{d\,\vec{\mu}_i}{d\,t} = \frac{d\,\vec{M}}{d\,t} = \gamma\,\vec{M} \times \vec{B} \tag{11}$$

In the most simple case, i.e., for $\left|\vec{M}\right| = const.$ and $\vec{B} = (0, 0, B_0 = const.)$, one obtains

$$\frac{d\,M_x}{d\,t} = \gamma\,B_0\,M_y\,, \quad \frac{d\,M_y}{d\,t} = -\gamma\,B_0\,M_x\,, \quad \frac{d\,M_z}{d\,t} = 0 \tag{12}$$

The solution of this differential equation, ie.,

$$M_x = m_{x0}\, cos(\omega_0 t), \quad M_y = -m_{y0}\, sin(\omega_0 t), \quad M_z = m_{z0} \tag{13}$$

(the values for m_{i0} are thereby given by the initial conditions) represents a precession (i.e., a rotation around the *z*-axis) of $\vec{M}$ with frequency $\omega_0 = \gamma B_0$. The direction of the vectorial rotation is opposed to $\vec{B}$. ω_0 is denoted as Larmor frequency. According to

$$\Delta E = h\nu = \hbar\omega_0 = \gamma\hbar B_0 \tag{14}$$

This corresponds to the frequency, which is equivalent to the energy difference between the two states with $m = \pm 1/2$.

Typical Larmor frequencies at a field strength of 1 T and for protons (^{1}H, spin 1/2) are 42.6 MHz, for ^{23}Na (spin 3/2) 11.3 MHz, for ^{31}P (spin 1/2) 17.2 MHz, i.e., the intrinsic angular momentum vectors execute a precession with these respective frequencies around the *z*-axis. If one excites the individual atomic spins from their equilibrium states and synchronizes their precession, the macroscopic magnetization $\vec{M}$ will precess with the same frequency and emit a recordable response.

In medical MR applications, the object to be examined is positioned in a stationary, strong magnetic field $\vec{B}_0$, whose orientation coincides with the z – axis of the reference (laboratory) system. A magnetization $\vec{M}_0$ as strong as possible is attempted to be obtained in this fashion. In a typical medical MR scanner today the stationary field strength is 1.5 T or 3 T, respectively. (higher field strengths are used experimentally as well as in small animal scanners). The exciting fields mentioned above which induce the desired effects (see later) are time-dependent and have the components $B_x(t)$ resp. $B_y(t)$, perpendicular to $\vec{B}_0$. The total magnetic field is therefore of the form $\vec{B} = \left[B_x(t), B_y(t), B_0\right]$.

The time-dependent fields cause an excitation of the system from its equilibrium state characterized by $M_x = 0$, $M_y = 0$, $M_z = |\boldsymbol{M_0}| = M_0$; the latter value being given by the population differences as calculated above under the influence of the field B_0. After termination of the time-dependent exciting fields the system relaxes back to its equilibrium state because of interactions with the environment associated with energy transfer. There are in essence two relaxation phenomena (explained in more detail later):

1. Through interaction (i.e., energy exchange) with the surrounding atoms the system relaxes back to its thermodynamic equilibrium state. Since the early NMR experiments were made in solid state (lattice) materials, this effect is denoted as spin-lattice interaction. It is characterized by a time constant T_1.
2. The spins carry a small magnetic field of their own which causes interactions in that the Larmor frequencies of the individual nuclei differ by a minute amount depending on the arrangement of the surrounding spins. This causes the components of the magnetization perpendicular to the *z*-axis, viz. M_x and M_y to disappear because of a gradual loss of synchronization or dephasing (spin-spin-interaction, time constant T_2).

The relaxation phenomena can be modeled with sufficient accuracy as linear first order processes. Accordingly, the angular momentum equations, denoted as Bloch's equations, reads

$$\begin{aligned}
\frac{dM_x}{dt} &= \gamma\left[M_y\, B_0 - M_z\, B_y(t)\right] - \frac{M_x}{T_2} \\
\frac{dM_y}{dt} &= \gamma\left[M_z\, B_x(t) - M_x\, B_0\right] - \frac{M_y}{T_2} \\
\frac{dM_z}{dt} &= \gamma\left[M_x\, B_y(t) - M_y\, B_x(t)\right] - \frac{M_z - M_0}{T_1}
\end{aligned} \tag{15}$$

We calculate the solution of these equations for the time after the excitation is terminated (for simplicity, we assume that at $t = 0$ both fields $B_x(t)$ and $B_y(t)$ are shut down). The equations then are

$$\begin{aligned}
\frac{dM_x}{dt} &= \gamma\, M_y\, B_0 - \frac{M_x}{T_2} \\
\frac{dM_y}{dt} &= -\gamma\, M_x\, B_0 - \frac{M_y}{T_2} \\
\frac{dM_z}{dt} &= -\frac{M_z - M_0}{T_1}
\end{aligned} \tag{16}$$

Upon elimination of M_y using the first two equations, one arrives at

$$\frac{d^2 M_x}{dt^2} + \frac{2}{T_2}\cdot\frac{dM_x}{dt} + \left(\omega_0{}^2 + \frac{1}{T_2{}^2}\right) M_x = 0 \tag{17}$$

The solution of this equation is a damped oscillation with frequency ω_0 and damping constant $1 / T_2$

$$M_x(t) = M_{x0}\, exp\left[\left(i\,\omega_0 - \frac{1}{T_2}\right) t\right] \tag{18}$$

M_{x0} represents the magnetization in x-direction at time $t = 0$ (initial condition, induced by the excitation). For $t \to \infty$ it is seen that M_x goes to zero. A corresponding solution is obtained for M_y.

The solution for M_z, in turn, reads

$$M_z(t) = \left(M_{z0} - M_0\right) exp\left(-\frac{t}{T_1}\right) + M_0 \tag{19}$$

and shows that M_z goes towards M_0 for $t \to \infty$ (M_{z0} is the magnetization in z-direction at time $t = 0$).

1.3. The Free Induction Decay (FID)

Nuclear spin- "resonance" is reached when the components $B_x(t)$ and $B_y(t)$ of the time-dependent excitation field initializing a measurement oscillates harmonically such

that the resulting field $\vec{B}_1 = \left[B_x(t), B_y(t), 0\right]$ rotates with the frequency $\vec{\omega}_0$, i.e. with the Larmor frequency. This can be shown by solving Bloch's equations as follows.

First, we remember that the Larmor frequency $\vec{\omega}_0$ has a sense of rotation which is opposite to the one of the field direction $\vec{B}_0$. In order to avoid unnecessary calculations, we assume that the exciting field with for the time being arbitrary frequency ω_z has a sense of rotation which likewise is antiparallel to this field (otherwise no resonance is possible), i.e. $\vec{\omega}_z = (0, 0, -\omega_z)$. The associated field $\vec{B}_1$ reads

$$\vec{B}_1 = \left[B_x(t), B_y(t), 0\right] = B_1 \left(cos\,\omega_x t, -sin\,\omega_x t, 0\right) \tag{20}$$

This corresponds to a field rotating around the *z*-axis counterclockwise with constant amplitude $B_1 = \left|\vec{B}_1\right|$ which at time $t = 0$ has *x*-direction.

In order to solve Bloch's equation, $\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \left[\vec{B}_0 + \vec{B}_1(t)\right] + \vec{R}$ (the relaxation terms are included for brevity in the vector $\vec{R}$), a transformation is now made into a system rotating with $\vec{\omega}_z$. This procedure will lead to a quite illustrative solution because the time dependence of the field $\vec{B}_1$ disappears: In the rotating system, both axes *x'* and *y'* (primes denote the new axes in the transformed system) rotate counterclockwise around the *z*-axis with frequency - ω_z, while the *z*-axis rotates on itself (*z'* = *z*). Since the amplitude of the field $\left|\vec{B}_1\right|$ is constant, a time-independent set of equations results. Upon application of the general formula for a transformation into a rotating system,

$$\frac{d'}{dt'} = \frac{d}{dt} - \vec{\omega} \times \tag{21}$$

Bloch's equation read

$$\frac{d'\vec{M}}{dt'} = \gamma \vec{M} \times \left[\vec{B}_0 + \vec{B}_1\right] - \vec{\omega}_{z'} \times M + \vec{R}' = \gamma \vec{M} \times \left[\vec{B}_0 + \frac{\vec{\omega}_{z'}}{\gamma} + \vec{B}_1\right] + \vec{R}' \tag{22}$$

The transformed relaxation terms, $\vec{R}'$, are first not considered in a simplified analysis, they are therefore not calculated in their transformed form. Note that the term $\vec{B}_0 + \vec{\omega}_z/\gamma$ has only a component in the *z'*-direction, while the vector $\vec{B}_1$ is constant after the transformation and is perpendicular to *z'* and oriented in the *x'*- direction. The quantity $\vec{B}_{eff} = \vec{B}_0 + \vec{\omega}_{z'}/\gamma + \vec{B}_1$ represents a constant, "effective" field in the rotating system such that the equation can be written as

$$\frac{d'\vec{M}}{dt'} = \gamma \vec{M} \times \vec{B}_{eff} + \vec{R}' \tag{23}$$

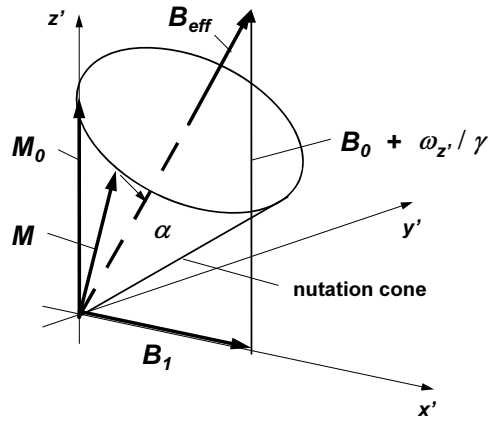


Figure 3. Nutation cone in rotating system

Before the field $\vec{B}_1$ is activated, the magnetization $\vec{M}$ has only a z'-component, M_0, which is constant in time. Once $\vec{B}_1$ is turned on, the equations then valid can easily be solved if the relaxation terms $\vec{R}'$ are disregarded. The solution describes a nutation or secondary precession (i.e., rotation) of the vector $\vec{M}$ around the axis defined by $\vec{B}_{eff}$ with the angular velocity $\omega_1 = \gamma B_{eff}$ (Figure 3). The opening angle of the nutation cone, α, can be changed at fixed $\vec{B}_0$ and $\vec{B}_1$ by variation of $\omega_{z'}$. The surface line M_0 remains thereby unchanged, i.e., the cone always includes the $z = z'$-axis. If one now chooses for $-\omega_{z'}$ the Larmor frequency, $-\omega_{z'} = \gamma B_0$, resonance is obtained. The opening angle α becomes 90°, since $\vec{B}_{eff}$ coincides with $\vec{B}_1$ (note that $\vec{B}_0 + \frac{\vec{\omega}_{z'}}{\gamma} = 0$). The cone becomes a plane lying in the y'-z' – plane; the tip of the magnetization vector $\vec{M}$ follows a circle. Accordingly, at resonance a virtual observer in the rotating system "sees" a constant field $\vec{B}_1 = \vec{B}_{eff}$ in x'-direction because he rotates with the same angular velocity as $\vec{B}_1$. He furthermore can observe the magnetization vector to rotate with constant angular velocity in the y'-z'-plane (Figure 4).

The angular velocity of nutation (or secondary precession) ω_1 at resonance equals γB_1, since B_{eff} is equal to B_1. If the field $\vec{B}_1$ is therefore activated during the time period Δt, the macroscopic magnetization $\vec{M}$ will nutate or precess around the angle $\theta = \Delta t\, \omega_1 = \Delta t\, \gamma\, B_1$. Accordingly, by an appropriate choice of Δt and B_1 it is possible to reach any desired flip angle. In practice, often 90°- or 180°-pulses are applied, which cause a re-orientation of $\vec{M}$ around 90° into the $-y'$ – axis, respectively around 180° into the $-z'$ – axis. The term "pulse" is used in this context because Δt is

typically on the order of a few msec while the Larmor frequency is in the MHz-range. The measurement fields are therefore high frequency (radiofrequency, RF) pulses.

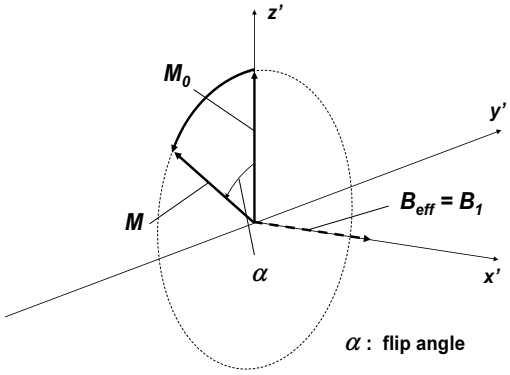


Figure 4. Secondary precession (nutation) at resonance

After termination of the excitation pulse the magnetization vector precesses with constant angular frequency ω_0 around the *z'*-axis. Since this vector is a macroscopic quantity in non-equilibrium conditions, it emits a radio signal with the same frequency which is sufficiently strong that it can be measured in a receiving coil. In order to analyze this signal, the relaxation terms have now to be taken into account.

The spin-spin interaction derives from the fact that each spin represents a magnetic dipole associated with a minute magnetic field which is superimposed over the external magnetic field. In typical biological tissues, the individual spins are inhomogeneously distributed. This causes small local inhomogeneities of the total field implying that the elementary spins have slightly different Larmor frequencies. The spins which are contained in a macroscopic volume element execute initially (i.e., immediately after the excitation) a synchronous precession. Yet, with increasing time, the spins get gradually out of phase because of the slightly different Larmor frequencies such that the component of the magnetization vector which is perpendicular to the *z'*-axis disappears. This can be described with sufficient accuracy as a linear process and the associated relaxation decay is of the first order with a characteristic time constant T_2. This effect is in practice significantly amplified because the external 'homogeneous' and constant magnetic field exhibits small local inhomogeneities itself for technical reasons in spite of correction measures (shim coils), thereby in addition causing locally varying Larmor frequencies. Accordingly, a relaxation time constant, T_2*, is found which is significantly smaller than T_2 and which in particular is machine-dependent. For biological tissues and field strengths around 1-3 T, T_2* is typically on the order of a few tens of msec (see later).

The spin-lattice relaxation, in turn, is due to interactions (collisions) with surrounding atoms whereby energy is exchanged and dissipated. These processes cause the magnetization vector to relax back into its thermodynamic equilibrium orientation in *z'*-axis. This can also be described by a first order relaxation process with an

associated time constant T_1. The latter is usually much larger than T_2^* (a few hundred msec).

Due to the relaxation phenomena and field inhomogeneities, a macroscopic volume induces a signal in a receiving coil, which oscillates almost harmonically, has a frequency ω close to the theoretical Larmor frequency ω_0 and whose amplitude decays exponentially with the characteristic time constant T_2^*. (Since the time constants T_2 and T_1 are significantly larger than T_2^* they manifest themselves only marginally.) This signal is denoted as Free Induction Decay (FID) and has the form, as shown above (only real part),

$$U = U_0 \, e^{-t/T_2^*} \cos \omega t \tag{24}$$

The demodulation of this signal is performed as usual by multiplication with $\cos(\omega_0 t)$,

$$U' = U_0 \; exp\,(-t / T_2^*) \; \cos \omega t \; \cos \omega_0 t\,, \quad \text{after low-pass filtering}$$

$$U'' = (U_0 / 2) \; exp\,(-t / T_2^*) \; \cos\,(\omega_0 - \omega)\, t \tag{25}$$

results. The function describes a sinusoidal curve with exponentially decreasing amplitude (FID, Figure 5, left).

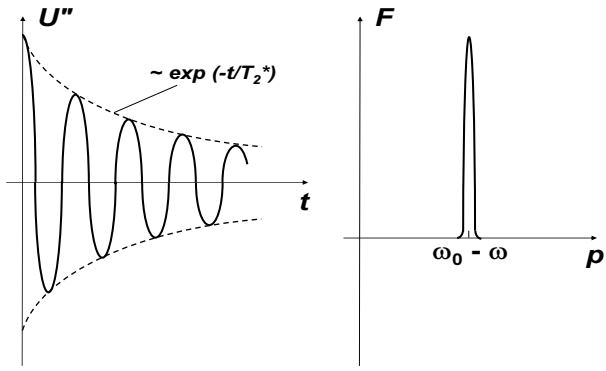


Figure 5. FID signal and Fourier transform

The real part of the Fourier-transform (independent variable p) of this signal reads

$$RT\left[F(U'')\right] \;\propto\; \frac{T_2^*}{1 + T_2^{*2}\left(\omega_p - \omega - p\right)^2} \tag{26}$$

and approaches under optimal conditions (signal-to-noise ratio) a line (Figure 5, right). The determination of such lines in the spectrum (NMR - spectroscopy) is of importance

likewise in medical imaging applications as well as in chemical analysis (see later). In particular, the primary signal strength, characterized by the area under the „line", is proportional to the amount of spins contributing to the signal, i.e. the spin density, ρ. As we shall see, this quantity is one of the main imaging parameters.

1.4. Measurement of T_2, Spin-Echo Procedure

Once the magnetization has been rotated into the *y'*-axis by a 90° pulse (Figure 6***a***), the spins start immediately to de-phase because of the slightly different precession velocities, such that the measured signal – as previously shown – decays exponentially with the time constant T_2^*. In the rotating system this means that the spins slowly fan out in both directions from the *y'*-axis (***b***). If after a time interval T_I a 180° pulse is applied (Figure 7), all spins are rotated in a plane parallel to the *y'-z'*-plane around 180° (***c***).

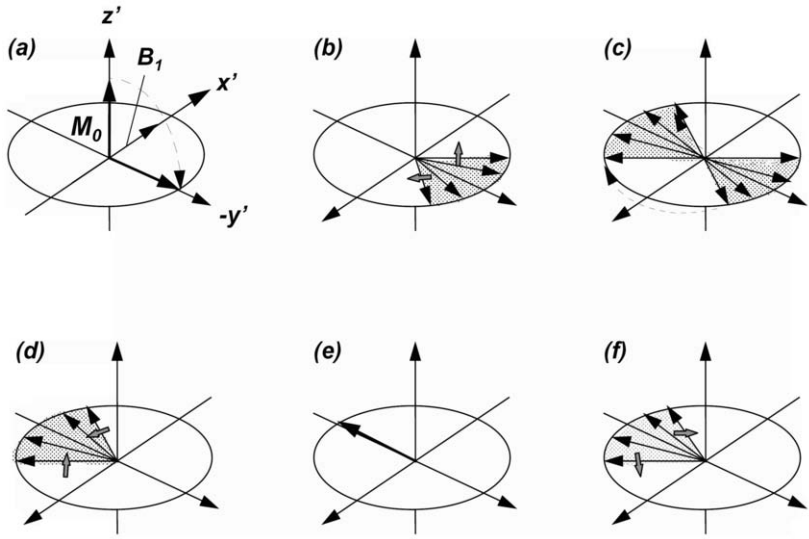


Figure 6. Spin-echo: De-phasing and re-phasing of spins

There are two reasons for the de-phasing of the spins, viz., (1) the spin-spin interaction which is in essence of a stochastic nature, (2) the inhomogeneities of the static magnetic field which are constant and systematic (but machine-dependent). By application of a 180° pulse the second effect can be compensated in that spins that rotate slower or faster due to field inhomogeneities continue to do so after the 180° pulse (***d***). Accordingly, the spins will rephrase after the time period $2T_I$ (***e***). In the receiving coil a signal will therefore build up and subsequently decay (***f***, Figure 7). The stochastic spin-spin interaction is not affected by this maneuver, such that the characteristic signal times are T_2^*. T_2 can e.g. be determined from repeating the procedure (repetition time T_R).

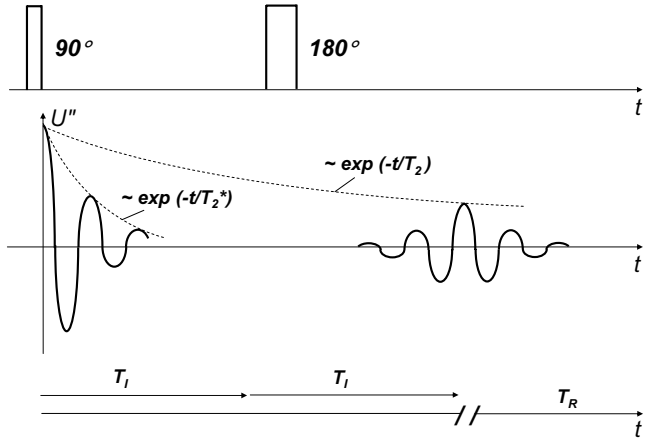


Figure 7. Spin-echo sequence

1.5. Measurement of T_1, Inversion-Recovery-Method

By application of a 180° pulse the magnetization is turned in *-z'* – direction (Figure 8). Due to the spin-lattice interaction it relaxes back to the original orientation. The momentary size of the *z'*- component of the magnetization during the relaxation process can be determined by triggering FID signals that are obtained from 90° pulses. T_1 is determined by fitting the exponential curve.

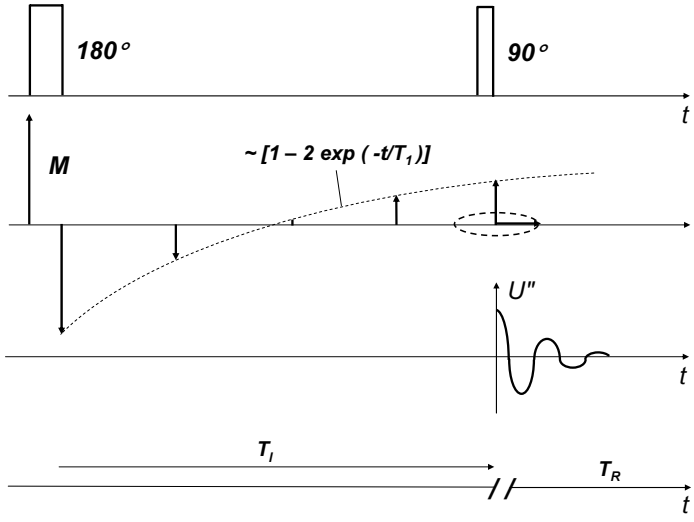


Figure 8. Inversion-recovery method

1.6. Chemical Shift

The spin-spin-interaction has a further important consequence. Depending on the number, type and location of the atoms surrounding a nucleus, the local spin configuration that a nucleus is exposed to is different and, along with this, the local field strength. The surroundings of a nucleus are however determined by the chemical compound in which it is located and by the position it has within this compound. Accordingly, the Larmor frequency of a nucleus is shifted according to its exact location within a chemical compound ("chemical shift"). An analysis of NMR spectra therefore allows us to elucidate the sterical configuration of a compound. This procedure is of particular value in protein analysis (Nobel laureate Kurt Wüthrich 2002).

2. Basics of Medical MR Imaging (MRI)

In MR imaging, the three parameters ρ (spin density), T_1 (spin-lattice relaxation time) and T_2 (spin-spin relaxation time) are measured and imaged in various combinations. These parameters exhibit characteristic differences for different tissues, e.g., for $B_0 = 0.3$ T :

Table 1. Typical relaxation times for biological tissues (B_0 = 0.3 T)

	T_1 (msec)	T_2 (msec)
Brain: Pons	445	75
Brain: Cerebellum	585	90
Bone marrow	320	80

A spatial resolution is achieved by the application of three mutually perpendicular gradient fields. These fields are gated and constant during the duration of application. Their direction is typically oriented along the z – axis such that the stationary field, B_0 has a contribution depending on the spatial location. With the linear gradients $\partial B/\partial i = G_i \quad (i = x, y, z)$ the magnetic field inducing magnetization is

$$\vec{B}_0 = \left(0, 0, B_0 + G_x \cdot x + G_y \cdot y + G_z \cdot z\right) \tag{27}$$

Due to the gradient fields, the Larmor frequency is dependent on the spatial location. The gradients can furthermore be combined arbitrarily, such that arbitrary cross sections through the object can be examined

In the following, the most often applied Fourier-imaging method is considered in a basic, simplified form. For further simplicity, a cross section perpendicular to the z – axis is chosen. The imaging procedure consists of three steps (Figure 9).

In a first step, the gradient G_z is activated during the duration T_P which is necessary to induce a 90° flip angle (excitation field B_1, frequency ω_P). The stationary field during this period of time is

$$\vec{B}_0 = \left(0, 0, B_0 + G_z \cdot z\right) \tag{28}$$

The resonance condition is only fulfilled at that location z, where

$$\gamma\left(B_0 + G_z\, z\right) = \omega_P \tag{29}$$

This implies that only a layer $z = z_0 = const.$ is excited. The thickness of the layer, Δz, depends on the gradient strength and the bandwidth of the applied pulse T_P. Outside this layer the resonance condition is not fulfilled such that associated volume elements will not contribute to the signals to be recorded later.

After the interval T_P, the excitation B_1 and gradient G_z fields are shut down concurrently. The magnetization in the chosen layer Δz which has been flipped around 90° starts to precess in the x - y – plane around the z – axis. Now, in a second step, the gradient G_y is activated. Due to this gradient, the local Larmor frequencies depend on the coordinate y. For a given y_i, however, the Larmor frequency does not depend on the coordinate x, i.e., in a slab extending in x - direction at location z_0, y_i, the precession frequencies are the same. The size of the slab perpendicular to the x – axis, Δy Δz, depends on the gradient strength and the bandwidth of the read-out system active in the third step.

The local precession frequency is

$$\omega\,(y_i) = \gamma\,(B_0 + G_z\, z + G_y\, y_i) = \omega_P + \gamma\, G_y\, y_i \tag{30}$$

After the time interval T_y the orientation of the local magnetisation vector has an orientation which exhibits a relative phase δ depending linearly on y_i:

$$\delta\,(y_i) = [\omega\,(y_i) - \omega_P]\; T_y = \gamma\; G_y\; y_i\; T_y \tag{31}$$

The y – coordinate is therefore „packed“ into a phase; accordingly, the gradient G_y is often denoted as phase-encoding gradient.

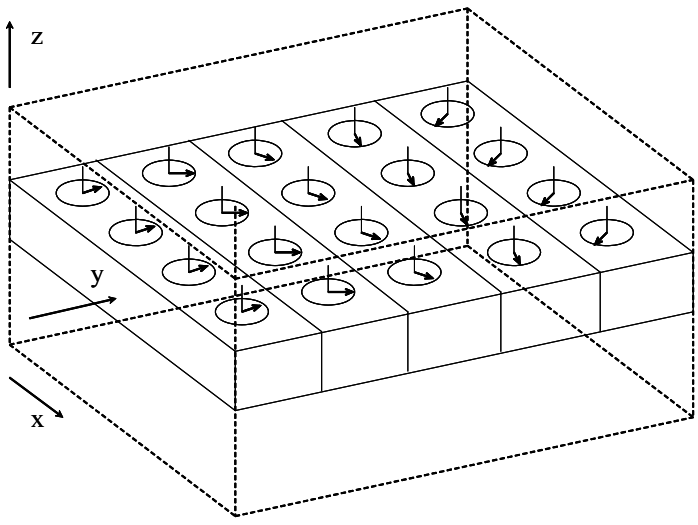


Figure 9. Principle of image acquisition, in particular, the phase encoding along the y-axis

After the interval T_y the gradient G_y is turned off and the measurement is started (third step). In order to achieve a resolution also in the x – direction, the gradient G_x is activated. The signal recorded in the measurement coil represents then a superposition of FID-pulses with different frequencies since the frequency of the FID signal $\omega(x_j)$ depends on x because of the gradient G_x (Figure 10). The recoded signal S, originating from the planar area at z_0 (without taking into account the relaxation terms), reads

$$S(z_0, t) \propto \sum_{i,j} \rho(x_j, y_i, z_0) e^{i\left[\omega(x_j)t + \delta(y_i)\right]} \quad \text{with} \tag{32}$$

$$\omega(x_j) = \gamma(B_0 + G_x x_j) = \omega_0 + \gamma G_x x_j \tag{33}$$

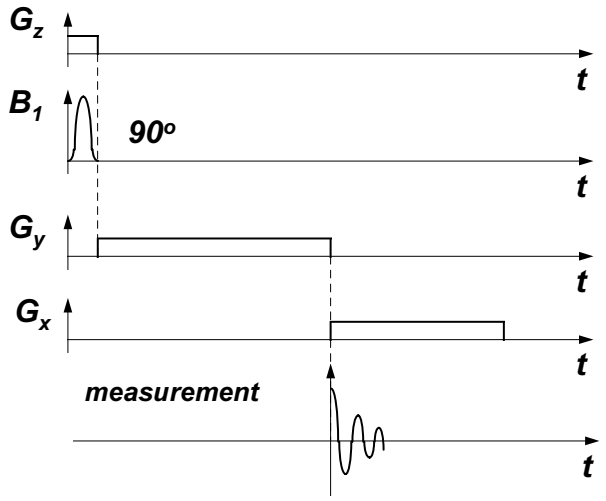


Figure 10. Sequence of gradients in the Fourier imaging scheme described here

The resolution Δx is again determined by the strength of the gradient and the bandwidth of the measurement system. After demodulation (multiplication with *cos* ω_0 and low-pass filtering) the signal reads

$$S'(z_0, t) \propto \sum_{i,j} \rho(x_j, y_i, z_0) e^{i\left[\gamma G_x x_j t + \gamma G_y y_i T_y\right]} \tag{34}$$

The signal is sampled in time steps T_x^m ($m = 1, \ldots, M$) in M channels. The measurement is subsequently repeated with $n = 1, \ldots, N$ different phase-encoding gradients G_y^n. The total signal is therefore finally composed of contributions of the form

$$S_{ij}^{mn}(z_0) \propto \sum_{i,j} \rho(x_j, y_i, z_0) e^{i\left[\gamma G_x x_j T_x^m + \gamma G_y^n y_i T_y\right]} \tag{35}$$

The summands represent the components of a discrete Fourier-transform such that the quantity $\rho(x_j, y_i, z_0)$, i.e., the image, can be obtained by application of Fourier methods.

Once the entire Fourier space (denoted as *k*-space in MR-jargon) is sampled, the image can be calculated. By an appropriate choice of the parameters G_x, T_x^m, G_y^n, T_y as well as of their temporal sequence the *k*-space can be sampled in various fashions which are adapted to the analysis to be performed.

3. MR Spectroscopy (MRS) and Further Methods

Due to the abundance of hydrogen nuclei in biological materials, ^{1}H imaging is used for routine examinations as the best signal/noise ratio is obtained with this element. Most routine applications are limited to morphology where spatial resolution is of primary importance. There are other, physiologically interesting elements however with non-zero spins that lend themselves for analysis and imaging, in particular also for functional imaging. ^{23}Na (spin 3/2), Larmor frequency (at 1 T) 11.3 MHz, and ^{31}P (spin 1/2), Larmor frequency 17.2 MHz, mentioned previously, are among these elements. Since the concentration of these elements is much lower that the one of hydrogen, signal/noise is correspondingly low. Image resolution is by far inferior such that a combination with ^{1}H imaging is usually made.

The chemical shift phenomenon allows for further analyses based on spectroscopy (MRS). A typical example is associated with ^{31}P as phosphor is a major component in physiological energy turnover. During this cycle, the phosphor atom appears in various chemical surroundings giving rise to associated chemical shifts (Figure 11).

MRI and MRS methods are still developing rapidly. A large number of advanced methods and medical applications are available; some of them are mentioned in the following. In particular, MR brain imaging is increasingly being applied in non-medical research, mostly related to the examination of human behavior patterns and psychology. For details and explanations the reader is invited to consult the literature, in particular the journal *Magnetic Resonance in Medicine*, where most advanced research results are being published.

With Echo-Planar Imaging (EPI) an entire image, i.e., the entire *k*-space is scanned during one single excitation. Imaging is extremely rapid, however, the method is associated with several drawbacks: The gradient coils have to be switched rapidly which is technically demanding and may furthermore induce adverse electrophysiological effects in the patient, susceptibility artifacts (caused by local variations in susceptibility due to the influence of the biological tissues) degrade image quality.

Sensitivity-Encoding (SENSE) is an efficient method to reduce imaging time. Thereby, the physical coil field characteristics are used to enable a parallelization of the *k*-space sampling.

Angiography is made by the application of pulse sequences that are sensitive to movement.

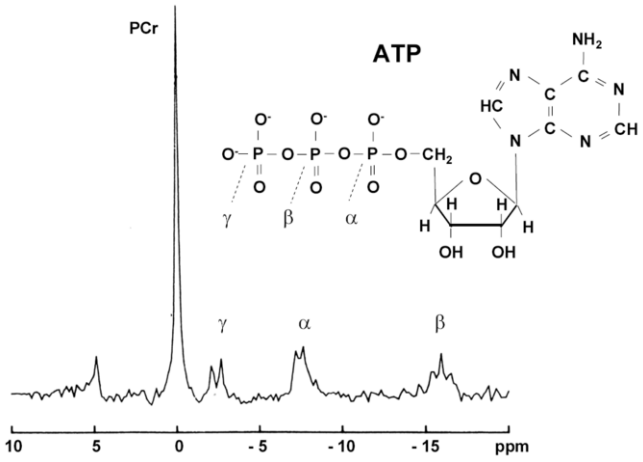


Figure 11. Phosphor spectrum. α, β, γ correspond to the three different positions of the phosphor atom in the ATP molecule. Due to the different chemical surrounding, Larmor frequencies are different (in ppm, parts per million deviation from frequency). PCr denotes phosphocreatinine, where the phosphor atom is again in different surroundings. The frequency of this strong line is chosen as reference (zero deviation).

By the application of pulse sequences which cause local saturation (no signal is thus obtained), patterns can be overlaid on biological structures (tagging, Figure 12). Muscular deformations can be visualized in this fashion.

Small-bore, high-field systems (up to 19 T) are used in small animal research (mice, rats). In combination with MR spectroscopy, fluorescence diffusion optical tomography (FDOT) or bioluminescence recording (both latter technologies are based on infrared diffusion) physiological and pathological processes are studies in animal models.

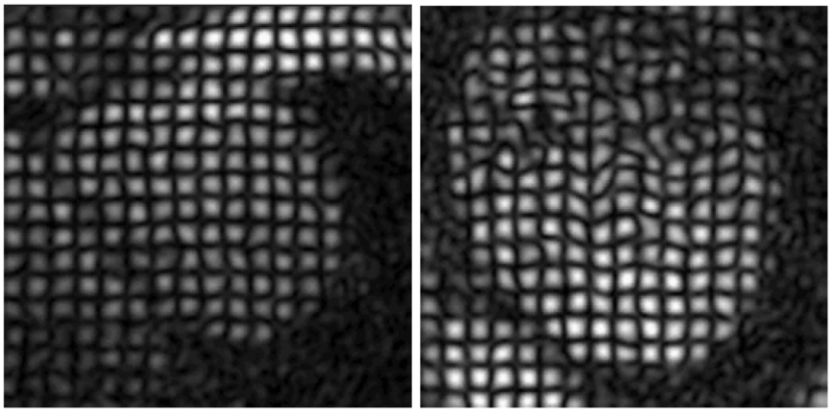


Figure 12. Tagging: The pattern induced by an appropriate pulse sequence allows the monitoring of tissue deformations as function of time, here of the heart (left: short axis section; right: long axis section).

Most biological tissues exhibit an anisotropic fiber structure. In Diffusion Tensor Imaging (MR-DTI) the components of the diffusion tensor of water molecules is measured. The primary eigenvector characterizes the average fiber direction in each voxel since water molecules diffuse faster along membranes than across. By concatenation of such vectors (fiber tracking), the fiber architecture of organs can be highlighted (Figure 13).

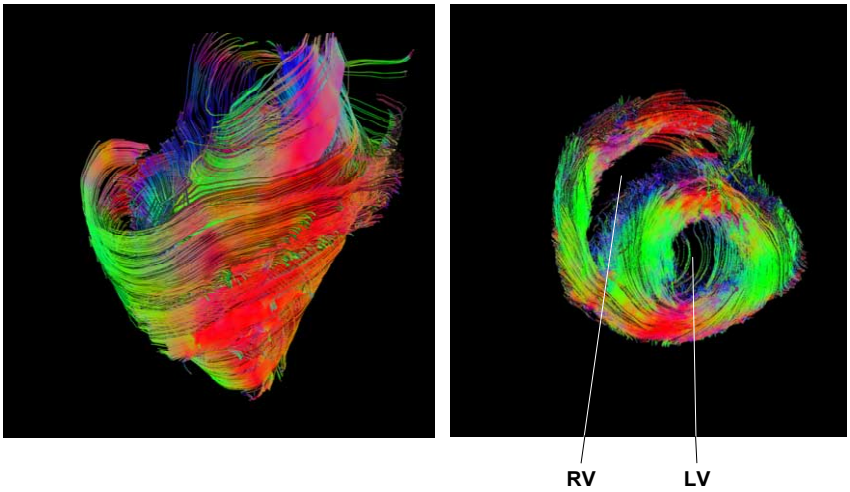


Figure 13. Fiber architecture of the human ventricles as determined by MR-DTI (ex vivo). RV: right ventricle, LV: left ventricle.

Further Reading

[1] C.N. Chen, D.H. Hoult, *Biomedical Magnetic Resonance Technology*. Medical Sciences, Taylor & Francis, 1989. ISBN 978-0852741184.

[2] A. Oppelt, *Imaging Systems for Medical Diagnostics: Fundamentals, Technical Solutions and Applications for Systems Applying Ionizing Radiation, Nuclear Magnetic Resonance and Ultrasound*, Wiley-VCH, 2006, ISBN 978-3895782268.

[3] M. Blaimer, F. Breuer, M. Mueller, R.M. Heidemann, M.A. Griswold, P.M. Jakob, SMASH, SENSE, PILS, GRAPPA: How to Choose the Optimal Method, *Topics in Magnetic Resonance Imaging* **15** (2004), 223–236. http://cfmriweb.ucsd.edu/ttliu/be280a_05/blaimer05.pdf.

[4] A.G. Filler, The history, development, and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, DTI, *Nature Precedings*. doi:10.1038/npre.2009.3267.2.

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-321

Introduction to Chapter VI: Rehabilitation Engineering

Tadej BAJD and T. Clive LEE (eds.)

Rehabilitation engineering is a rapidly evolving field that can be defined as the application of science and engineering to the development, design and application of assistive technologies and neurorehabilitation techniques that need to be matched with the specific needs of a person with a particular disability or a particular condition. This definition in its core assumes and requires a substantial degree of technical competence on the side of rehabilitation engineers that needs to be complemented with clinical competence of physicians and therapists all of whom form a team that assesses the needs of and determine the goals of a rehabilitation process for a particular client. This requires from all the rehabilitation team members not only specialized knowledge in their own field of competence but also the ability to successfully communicate with the other team members. In a successful rehabilitation team this may only be possible if rehabilitation engineers possess basic knowledge on anatomy, physiology, rehabilitation medicine and therapy while on the other hand the medical professionals understand basics of rehabilitation engineering. Rehabilitation of mobility and manipulation is the domain where a harmonic co-operation of the team members is critical for a successful application of an assistive device or a specific regime of training.

This contribution is divided into three standalone sections that nicely complement each other. In the first section we present basics of gait analysis and synthesis including orthotics and prosthetics that require an understanding of underlying biomechanics. The second section is devoted to the functional electrical stimulation (FES) of extremities, a technique which may be used as orthotic means or a temporary training aid, where the basic knowledge on electrical and biomechanical parameters associated with FES is presented. The third section is devoted to rehabilitation robotics, a field that saw a rapid development in the last decade, mainly due to introduction of virtual reality and haptic interaction in the rehabilitation of various neurological disorders. All three sections jointly provide basic concepts of biomechanics, kinesiology and technology needed to comprehend possible benefits when applied correctly to the patients.

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-323

VI.1. Gait Analysis and Synthesis: Biomechanics, Orthotics, Prosthetics

Zlatko MATJAČIĆ
University Rehabilitation Institute, Republic of Slovenia
University of Ljubljana
Linhartova 51
1000 Ljubljana
Slovenia

Abstract. This contribution presents and establishes the biomechanical principles that underlie human walkin. This is done by using a range of simplified biomechanical models of bipedal walking to explain the laws of movement and associated energetic requirements. Based on these simplified models, the measurements of normal walking are described. Selected pathological cases are used to illustrate the changes that occur in abnormal walking patterns. Finally, the basic design principles used when applying orthotics and prosthetics to enhance or restore impaired or missing function in walking are described for these case studies.

Keywords. Walking, bipedal locomotion, kinematics, kinetics, orthotics

Introduction

The analysis of human bipedal locomotion is of interest to many different disciplines, including biomechanics, human movement science, rehabilitation and medicine in general. Gait is the collective term for the two types of bipedal locomotion, walking and running. In this paper, we will deal only with walking. The objective is to alter pathological walking by means of orthotic or prosthetic aids.

Gait analysis involves basic measurements, such as walking speed, step length and cadence, and more detailed measures of the relative motion of body segments and joints, the patterns of forces applied to the ground and the sequence and timing of muscular activity. The movements of joints and segments, including position, velocity and acceleration, are termed *kinematics*. Kinematic data are obtained by contactless movement tracking systems that use video cameras for tracking reflective markers attached to predetermined anatomical positions on the selected body parts.

The pattern of forces is obtained from force platforms, sensors that are imbedded in the floor of a laboratory, that measure the forces exerted by the foot when in contact with the ground. These ground reaction forces, are used together with kinematic and anthropometric data, segment sizes, masses and inertial properties, to calculate net joint torques produced by muscular activity during gait. This calculation requires the

application of general equations of motion in the inverse dynamics model of the human body (Koopman paper, Chap1). These models constitute a part of the software of modern movement tracking systems. The ground reaction forces and the net joint torques are termed *kinetics*.

The patterns of muscular activity are termed *electromyographs* and can be obtained from surface electrodes applied over the skin of the muscle of interest. Kinematics, kinetics and electromyography represent important variables that help the clinicians to understand a particular gait pattern and to discriminate primary gait abnormalities, from the compensatory changes that are the consequence of the former. Based on an understanding of a pathological gait, a treatment procedure or application of suitable orthotics and prosthetics can be devised. However, understanding complex gait patterns is a challenging task for clinicians, as walking is largely defined by biomechanical laws. While understanding normal gait patterns can be learned through descriptions of events and observations during walking, the comprehension of individual pathological cases is challenging because they can be very diverse. In order to correctly interpret measured gait patterns, a basic knowledge of the biomechanics of walking is required.

The aim of this paper is to establish the biomechanical principles that underlie human walking. This is done by using a range of simplified biomechanical models of bipedal walking to explain the laws of movement and associated energetic requirements. Based on these simplified models, the measurements of normal walking are described. Selected pathological cases are used to illustrate the changes that occur in abnormal walking patterns. Finally, the basic design principles used when applying orthotics and prosthetics to enhance or restore impaired or missing function in walking are described for these case studies.

1. Bipedal walking in humans

1.1. Basic Definitions and Terminology

Walking is a cyclical activity that is composed of several succeeding cycles. The gait cycle is most frequently defined as a period from heel contact of one leg until the next heel contact of the same leg. Figure 1 illustrates the sequence of silhouettes of a walking person that constitute a gait cycle. The gait cycle consists of two distinctive phases: the stance phase in which the foot of the leg is placed on the ground and the swing phase in which the leg is in the air and advances to the position of the next heel contact. Each stance phase has a period of time at the beginning and at the end, termed double stance, in which both legs are on the ground. In normal walking, the stance phase extends over 60% of the gait cycle, while the swing phase occupies the remaining 40%. The duration of each double stance is approximately 10%. The stance phase is further sub-divided into: initial contact, loading response, mid-stance, push-off and pre-swing. The swing phase is sub-divided into: initial swing, mid-swing and terminal swing.

Step length is defined as the distance between the same points on each foot during double stance. *Stride length* is the distance between two successive heel strikes by the same foot and therefore represents the distance travelled within the gait cycle. *Walking speed* is the average speed over multiple strides, while *cadence* is the number of steps per time unit.

1.2. Simple Biomechanical Models of Walking

Walking is a cyclical activity where one of the important features is energy conservation. In the following section, we will derive a simple biomechanical model of bipedal walking that accounts for, and explains, all the major characteristics of bipedal walking.

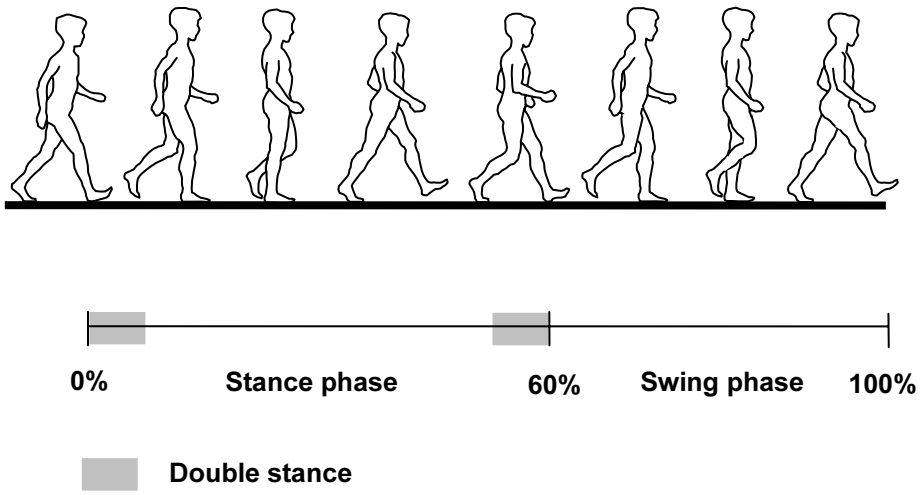


Figure 1. The gait cycle is divided into a stance phase, when the leg is in contact with the ground, and a swing phase, when the leg is in the air. Within the stance phase there are periods of double stance when both legs are on the ground.

Figure 2 shows a simple pendulum composed of a mass particle and a light, weightless rod. When the mass particle is displaced from the vertical position, work is done that increases the potential energy of the system. When the pendulum is released it starts moving along the circular arc. The level of *potential energy* is decreasing while the *kinetic energy* is increasing. When the mass particle passes the vertical position, all of the initial potential energy is transformed into kinetic energy. Later, when the pendulum continues moving along the arc, the velocity is decreasing as well as the level of kinetic energy, while the level of potential energy is increasing. If no dissipative force, such as friction, is present in the system, the swinging of the pendulum would continue indefinitely in accordance with Newton's First Law of Motion. In the ideal case presented, the level of mechanical energy, being a sum of potential and kinetic energies, remains constant at every time instant.

Figure 3 shows a simple inverted pendulum, where similar energetic considerations are applicable as in the case of the pendulum in Figure 2. Here, the mass particle initially possesses a certain velocity and, consequently, kinetic energy. When rotating upwards towards the vertical position, the level of kinetic energy is decreasing, due to a deceleration action of the gravitational force, while at the same time the level of potential energy is increasing, thus conserving the total mechanical energy of the

system. When the vertical position is reached, only a minimal horizontal velocity needed for passing the vertical position is present. Afterwards, when the gravitational force accelerates the mass particle along the arc, the potential energy level is decreasing while the kinetic energy level is increasing. The inverted pendulum model from Figure 3 can be considered as the simplest model of bipedal walking during the stance phase. The model is completed by adding another weightless rod articulated with the first one through a simple hinge joint at the location of the mass particle which constitutes a hip joint.

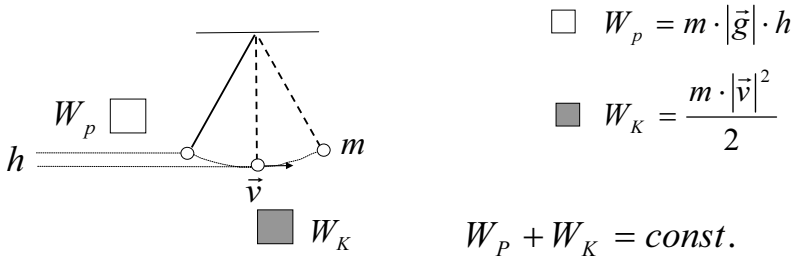


Figure 2. Qualitative model of a simple pendulum. Pendulum movement is illustrated with a sequence of three positions. When the pendulum is displaced from equilibrium it possesses the largest value of potential energy W_K, velocity of movement equals zero. When passing vertical position the pendulum possesses the largest velocity $\vec{v}$ and thus also kinetic energy W_K, which equals potential energy at the initial displaced position. (m = mass of pendulum bob; g = acceleration due to gravity, $\vec{v}$ indicates that velocity, v, is a vector quantity, as it has both magnitude and direction)

Figure 4 shows such a bipedal model, where one leg is in stance, providing a support to the mass particle on the top, while the other leg is in swing, in preparation for the next stance phase of the gait cycle. In the preceding treatment of bipedal walking we have considered only the gravitational force, which is considered an external force. Another external force that acts upon the inverted pendulum is a ground reaction force (GRF), that acts between the tip of the leg and the ground. In the simple inverted pendulum model, the GRF is entirely passive and acts along the leg throughout the circular motion. While the model from Figure 4 is conceptually clear, it misses a very important aspect of bipedal walking - the impact of the swinging leg and the ground at the moment of initial contact.

Figure 5 shows the situation at the moment of impact. In this passive inverted pendulum model, the impact time is very short and can be regarded as instantaneous. Just before impact, the mass particle possesses certain velocity $\vec{v}_-$ that is perpendicular with respect to the stance leg. When the leading leg impacts the ground a $\overrightarrow{GRF}_L$, which is directed along the leg, acts on the leg contacting the ground. In the short time of impact the force impulse, which equals the product of the magnitude of $\overrightarrow{GRF}_L$ and the time duration of the impact, acts in such a way that the direction as well as the amplitude of the velocity of the mass particle is changed. This change is depicted by a

vector composition, which results in a new velocity of the mass particle $\vec{v}_+$, which is directed perpendicularly with respect to the leg that just entered stance in order to be able to perform circular motion.

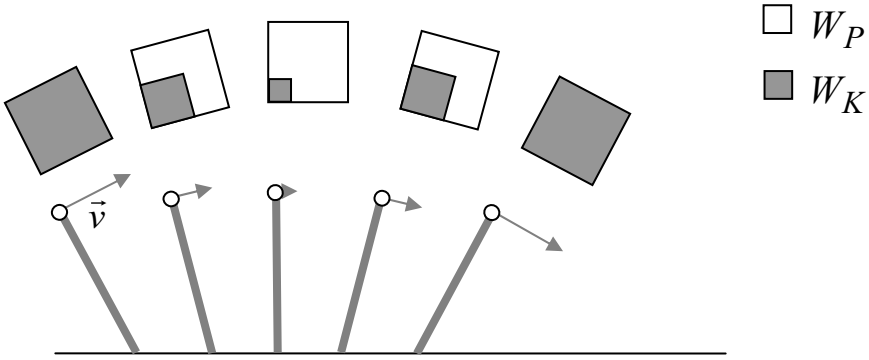


Figure 3. Qualitative model of an inverted pendulum. Movement of the inverted pendulum is illustrated with a sequence of five positions with indicated velocity vectors. Squares denote a share of potential and kinetic energy in each of five positions. Note that the sum of both energies remains constant in each case.

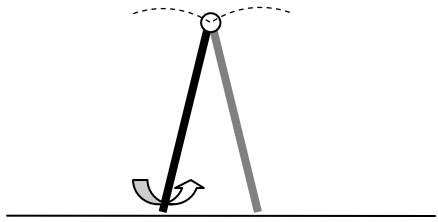


Figure 4. The simplest model of bipedal walking presented by two inverted pendulums articulated with a simple rotational joint.

Due to the geometry of both legs in double support phase, the consequence of the force impulse of the $\overrightarrow{GRF}_L$ at impact always decreases the total mechanical energy of the system. The change in the mechanical energy before and after the impact is proportional to the squared change in the velocities before and after the impact. The lost energy is transformed into a sound that we hear at the time of the impact and the deformation (which is so small that can not be noticed with a naked eye) of the leg that impacted the ground. The situation displayed on Figure 5 demonstrate one of the most important features of the bipedal walking; in every step there is a loss of mechanical energy due to the impact of the swinging leg with the ground. We can easily imagine that due to this loss of energy the simple mechanism would not be able to make another

step, because the kinetic energy of the system after the impact would be insufficient to reach the vertical position. It is clear that lost energy needs to be recovered within each step; this means that another important feature of bipedal walking is that in every step, work needs to be performed to recover the level of mechanical energy that is necessary for alternating circular motion of both legs.

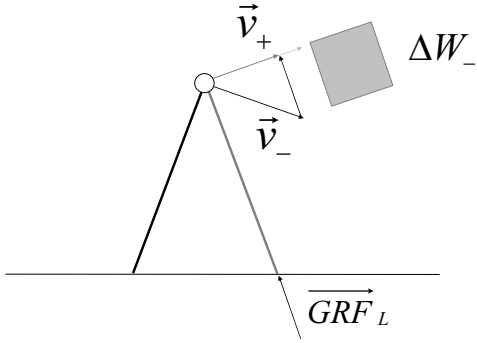


Figure 5. Energetics in double stance. A force impulse of $\overrightarrow{GRF}_L$ changes the direction of movement of mass particle m as well as reducing kinetic energy.

One possibility for efficient generation of mechanical energy in bipedal walking is shown in Figure 6. The geometry of the mechanism is identical to that in Figure 5. However, just before the impact of the swinging leg with the ground, an impulse of the $\overrightarrow{GRF_T}$ under the trailing leg is performed. Such an impulse could, for example, be achieved by releasing a pre-compressed spring (ideally, compressed through a prior impact, thus storing energy rather than dissipating it). Releasing the elastic energy from the spring would change the direction and the magnitude of the velocity of the mass particle in such a way that the mechanical energy of the system would be increased due to the increased magnitude of the velocity of the mass particle. Consequently, the impulse of the $\overrightarrow{GRF}_L$ of the leading leg impacting the ground would further change the direction and the magnitude of the mass particle velocity, similar to the case described in Figure 5. The overall effect of both force impulses would be such that the velocity of the mass particle before and after the impact would be exactly the same in magnitude while the direction after the impact would be perpendicular allowing for a passive rotation of the stance leg as an inverted pendulum as shown in Figure 3. Also, when comparing Figures 5 and 6, we can notice that a smaller amount of energy is dissipated when a push-off impulse preceded the impact.

A more realistic model of the double support phase is obtained if both legs are not rigid, but can retract and extend, i.e. change their length. In this way, changes in the energetics of walking need not be instantaneous, which is associated with high forces, but can be achieved in a more continuous manner. This is illustrated in Figure 7 where

the three snapshots are used to show the gradual changes in the direction of the velocity of the mass particle throughout the double support phase. We can see that, instead of the rather abrupt change in the trajectory of the mass particle as shown in Figure 4, a more fluid and continuous trajectory curve is achieved when simultaneous extension of the trailing leg and retraction of the leading leg are utilized. Also, the changes in the mechanical energy are not instantaneous but depend on i) the absorption of power by the leading leg (P_L), which in every time instant equals the vector dot product of the $\overrightarrow{GRF}_L$ and the $\vec{v}$ and ii) the generation of power by the trailing leg (P_T), which in every time instant equals to the vector dot product of the $\overrightarrow{GRF_T}$ and the $\vec{v}$. The time integral of both powers relate to changes in the mechanical energy of the system.

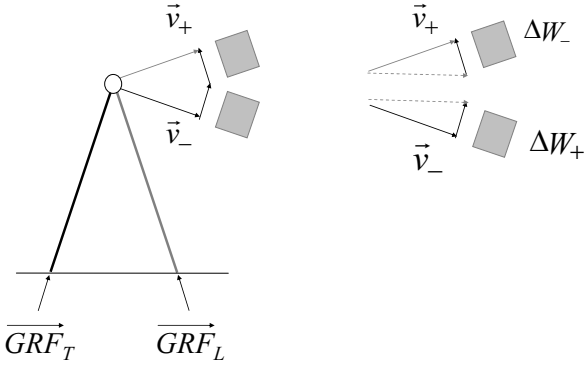


Figure 6. Active push-off is achieved by a force impulse of the trailing leg just before the impact of the leading leg with the ground. The overall effect of both force impulses is an appropriate change in the direction of velocity of the mass particle.

To summarize the above derivation of the simple bipedal walking model, the essential components are the two legs that can change their length, a mass particle representing the mass of the whole body and the two external forces; the gravitational force acting directly on the mass particle and the GRF acting on the mass particle through both legs. Most of the movement in the steady state is passive while energy absorption is performed by the leading leg following impact and energy generation is performed by the trailing leg just prior to entering the double support phase.

While this conceptual simple model of bipedal walking captures all the essential features of bipedal walking, it is only a first step toward understanding human walking, since our legs are not simple rods that extend and retract; rather, our legs are made of segments (pelvis, thighs, shanks and feet) that are articulated with joints (hips, knees and ankles). It is therefore necessary to extend our treatment to human-like biomechanical model in order to fully comprehend the mechanisms of bipedal walking. However, a proper and rigorous treatment of the biomechanics of such a model from the aspects of mathematics and physics is rather challenging and requires detailed knowledge of several engineering disciplines. Still, we can make certain simplifications

that will allow us a simplified treatment leading to useful results that relate to human walking.

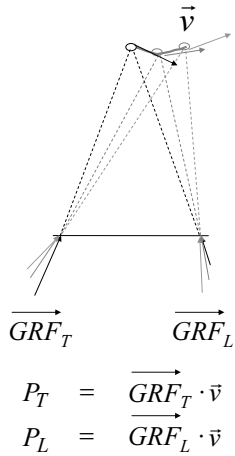


Figure 7. Realistic presentation of a double stance, where both ground reaction forces act in a way that gradually changes the direction of velocity of the mass particle. The leading leg is retracting and thus absorbing power and energy, while the trailing leg is extending and thus generating power and energy.

To summarize the above derivation of the simple bipedal walking model, the essential components are the two legs that can change their length, a mass particle representing the mass of the whole body and the two external forces; the gravitational force acting directly on the mass particle and the GRF acting on the mass particle through both legs. Most of the movement in the steady state is passive while energy absorption is performed by the leading leg following impact and energy generation is performed by the trailing leg just prior to entering the double support phase.

While this conceptual simple model of bipedal walking captures all the essential features of bipedal walking, it is only a first step toward understanding human walking, since our legs are not simple rods that extend and retract; rather, our legs are made of segments (pelvis, thighs, shanks and feet) that are articulated with joints (hips, knees and ankles). It is therefore necessary to extend our treatment to human-like biomechanical model in order to fully comprehend the mechanisms of bipedal walking. However, a proper and rigorous treatment of the biomechanics of such a model from the aspects of mathematics and physics is rather challenging and requires detailed knowledge of several engineering disciplines. Still, we can make certain simplifications that will allow us a simplified treatment leading to useful results that relate to human walking.

Figure 8 shows a model that resembles human skeletal system in a sequence of five snapshots that illustrate the conditions throughout the stance phase from initial contact until push-off. The leg segments are weightless as was the case in our previous models, while the mass particle is replaced by a large ellipsoid representing the head, arms and trunk. In the middle of this ellipsoid is a small circle that represents a centre of mass (COM). Since the legs have no weight, the COM remains in the same relative position within the ellipsoid. We shall first observe the sequence of humanoids from

Figure 8 in the same way as was presented in previous figures. We need to concentrate only on the COM, $\overrightarrow{GRF}_L$ and a "virtual" leg, which is represented by the dashed line connecting the point on the ground where $\overrightarrow{GRF}_L$ acts on the leg; this point is called the centre of pressure (COP). We can see that there is considerable resemblance with a simple inverted pendulum.

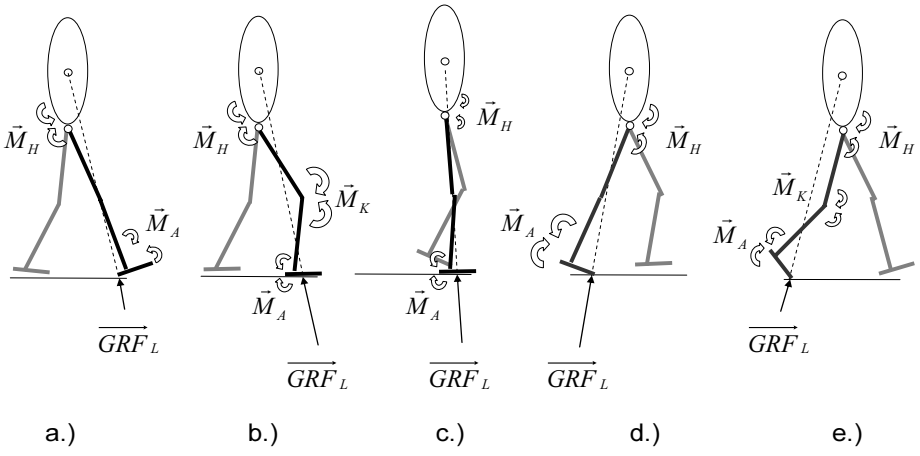


Figure 8. Human-like simple biomechanical model of walking in stance phase. Depending on the external moment in each joint produced by the GRF, the necessary balancing muscular moments are shown in the ankle ($\vec{M}_A$), the knee ($\vec{M}_K$) and the hip ($\vec{M}_H$) in all sub phases of stance: a.) initial contact, b.) loading response, c.) midstance, d.) push-off, e.) pre-swing.

The GRF always acts in the direction of the COM, while the 'virtual' leg rotates similar to the inverted pendulum in Figure 3. The length of the 'virtual' leg depends on the position of the thighs, shanks and feet of both legs and changes throughout the whole stance phase. In energetically efficient walking, these changes are such that the oscillations of the COM in the vertical direction are minimal. So far, the forces that we associated with bipedal walking were the *external* forces (gravity and GRF). Here, we need to introduce also the so-called *internal* forces that are produced by the leg muscles and result in equivalent net joint torques. The relationship between the external and internal forces in our model is rather simple. In each of the lower extremity joints, the net joint torques must equal to the product of the magnitude of $\overrightarrow{GRF}_L$ and the perpendicular distance from the line of action of the $\overrightarrow{GRF}_L$ and each individual joint. For example, at the initial contact, the $\overrightarrow{GRF}_L$ acts posterior to the ankle joint, thus producing the moment around the ankle that tends to plantarflex the foot onto the ground. Thereby, ankle dorsiflexors must be active to produce a net joint torque that nearly balances the external ankle moment produced by the $\overrightarrow{GRF}_L$. Similarly, in the

hip the $\overrightarrow{GRF}_L$ passes in front of the hip joint thus producing external moment that tends to flex the hip. Therefore, hip extensors must produce a net joint torque that nearly balances the external hip moment produced by the $\overrightarrow{GRF}_L$. In the knee joint no net joint torque is needed from either knee extensors or flexors, as the $\overrightarrow{GRF}_L$ passes through the joint, thereby producing no external moment. In the subsequent phases of stance, different demands arise for each joint muscles, depending on the requirements imposed by the $\overrightarrow{GRF}_L$. Here we need to stress that the only external force that is independent is gravitational force, while the $\overrightarrow{GRF}_L$ depends on gravity, mechanism dynamics and internal forces and moments produced by the muscles and other soft tissue. Through loading response, the greatest demand is on the knee extensors that act eccentrically (producing tension while being lengthened), thereby absorbing the energy due to the impact of the leading leg with the ground. Also, considerable activity is needed from the gluteal muscles to maintain the trunk erect. In midstance, there are rather small net joint torques needed, as the 'inverted' pendulum passively rotates around the ankle joint and the $\overrightarrow{GRF}_L$ passes in the vicinity of all three joints. During push-off, a strong concentric activity (producing tension while shortening) of plantarflexors is needed (to extend the 'virtual' leg), thereby generating the power and energy needed to compensate for the energy loss that will follow during the impact of the other leg. Here also, activity of hip flexors is needed to counteract the external $\overrightarrow{GRF}_L$ moment. During pre-swing, the activity of hip flexors, knee extensors and plantarflexors are needed due to the orientation of the $\overrightarrow{GRF}_L$

The double support phase is illustrated in Figure 9. The left side of Figure 9 shows both 'virtual' legs, both GRFs and all the moments in hips, knees and ankles of both extremities. The right side of Figure 9, similar to Figure 7, shows the evolution of the COM trajectory that results from shortening of the leading leg (flexion of all three joints) and the trailing leg extending (extension of all three joints). The relationships for power absorption (P_L) in the leading leg and power generation in the trailing leg (P_T) in the model from Figure 7 (simple rod shortening and lengthening) and in the model from Figure 9 ('virtual' leg shortening and lengthening resulting from changes in the angular positions of the leg segments) are given in the following equations:

$$P_T = \overrightarrow{GRF}_T \cdot \vec{v} = \vec{M}_{HT} \cdot \vec{\omega}_{HT} + \vec{M}_{KT} \cdot \vec{\omega}_{KT} + \vec{M}_{AT} \cdot \vec{\omega}_{AT} \tag{1}$$

$$P_L = \overrightarrow{GRF}_L \cdot \vec{v} = \vec{M}_{HL} \cdot \vec{\omega}_{HL} + \vec{M}_{KL} \cdot \vec{\omega}_{KL} + \vec{M}_{AL} \cdot \vec{\omega}_{AL} \tag{2}$$

The generated power/absorbed power in each joint depends on the net joint torques produced by the muscles and the angular velocity. If the direction of net joint torque and angular velocity is the same as is the case in Figure 9 for the trailing leg, the power and energy is generated in a particular joint, if the two directions are opposite as is the case in Figure 9 for the leading leg the power and energy is absorbed in a particular joint. Therefore, while the requirements that the leading leg is absorbing power and energy during loading while the trailing leg is generating power and energy during push-off is an inherent, non-negotiable feature of bipedal walking, this can be achieved in many different ways (with different joints contributing differently depending, for

example, on pathological conditions) resulting in different kinematic and kinetic patterns.

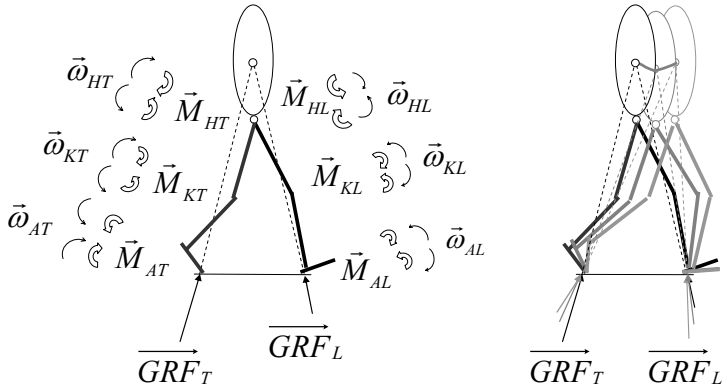


Figure 9. Human-like simple biomechanical model of walking in double stance phase. a.) Depending on the external moment at each joint produced by the ground reaction forces ($\overrightarrow{GRF}_L$, $\overrightarrow{GRF}_T$), also the necessary balancing muscular moments and angular velocities are shown in the ankle ($\vec{M}_{AT}$, $\vec{M}_{AL}$ and $\vec{\omega}_{AT}$, $\vec{\omega}_{AL}$), the knee ($\vec{M}_{KT}$, $\vec{M}_{KL}$ and $\vec{\omega}_{KT}$, $\vec{\omega}_{KL}$) and the hip ($\vec{M}_{HT}$, $\vec{M}_{HL}$ and $\vec{\omega}_{HT}$, $\vec{\omega}_{HL}$). b.) Realistic presentation of double stance where both ground reaction forces act in a way that gradually changes the direction of velocity of the mass particle.

The swing phase, which completes our theoretical treatment of bipedal walking, is depicted in Figure 10. For clarity, only the swinging leg is shown. Immediately after the leg leaves the ground (initial swing) net joint torques are needed in the hip (flexor moment) to propel the leg forward, while knee extensors need to eccentrically arrest the movement of the knee in the flexion that resulted from preceding push-off. During mid-swing, almost no muscular activity is needed as the leg segments are moving balistically. Terminal swing is characterized with eccentric activity of hip extensors and knee flexors to decelerate hip flexion and knee extension in order to prepare the leg for the next stance phase.

1.3. Normal Walking in Humans

In the previous section, we derived a conceptual biomechnical model of bipedal walking that was based on certain simplifications. In order to validate this model, we need to compare it with real measurements of walking.

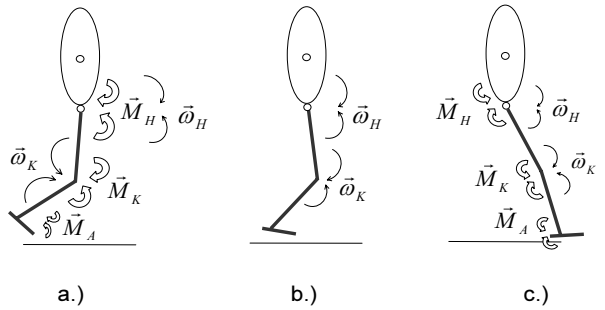


Figure 10. Human-like simple biomechanical model of walking in swing phase. Muscular moments and angular velocities are shown for each joint in the a.) initial swing, b.) mid-swing, c.) terminal swing.

Figure 11, which is similar to Figure 8, shows five snapshots of stance phase of walking in a neurologically and orthopaedically intact individual at a velocity of 1 m/s. Figure 11 shows the displacements of body segments and the magnitude and direction of GRF along the dashed line pointing into the COM. Figure 12 shows the horizontal and vertical components of GRF throughout the gait cycle. Consistent with our simple biomechanical model, we can notice that the first 30% of the stance phase is characterized by a negative value of horizontal GRF component (braking the movement of an inverted pendulum) while the remaining 30% is characterized by a positive value of horizontal GRF (propelling the movement of the inverted pendulum forward in the direction of progression). The first 20% of stance is characterized by a first peak value of vertical GRF component that exceeds the gravitational force, which is associated with the deceleration of the downward movement of COM and its redirection into an upward movement, as predicted by our simple energetic model of the double support phase. In the next 20% of stance, the inverted pendulum is almost passively rotating around the ankle joint, therefore the magnitude of vertical GRF component is lower than the gravitational force (body weight). In the last 20% of the stance the vertical GRF component again exceeds the gravitational force which is associated with strong forward and upward propulsion (leg is extending and generating positive power).

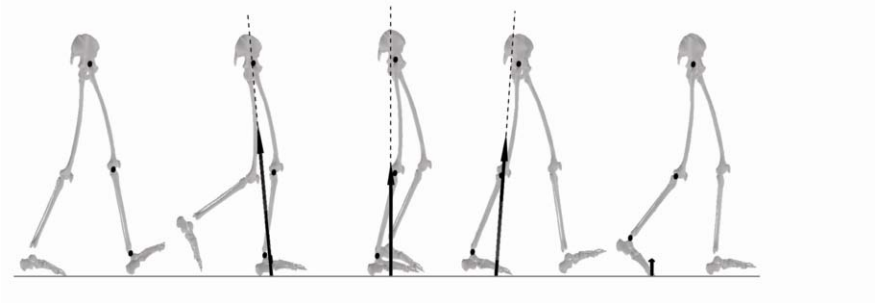


Figure 11. Representation of pelvis and lower extremities movement together with a GRF in five consecutive snapshots within the stance phase in normal walking.

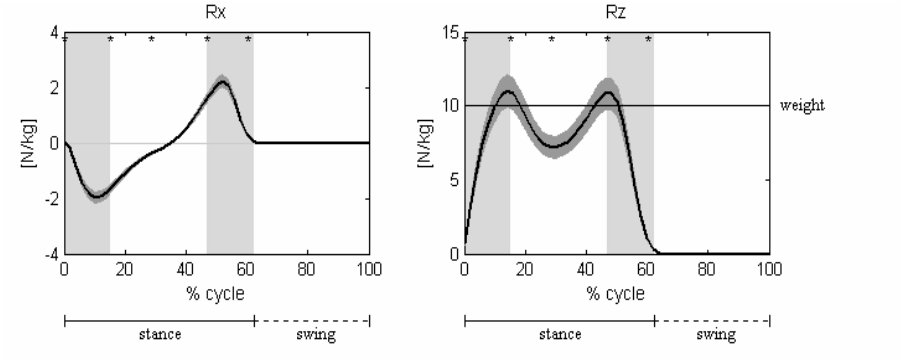


Figure 12. Graphs of horizontal component of GRF (R_x) and vertical component of GRF (R_z) throughout the gait cycle. Stars above the curves denote time instants that correspond to the snapshots from Fig. 11. Mean values and standard deviations are for normal walking (adapted from Winter, 1991). Both periods of double stance are also shown.

Figure 13 shows the joint kinematics, kinetics and energetics throughout the whole gait cycle. Consistent with the simple biomechanical model, we can observe that the ankle undergoes dorsiflexion (rotation of the inverted pendulum) throughout the majority of stance, ankle plantarflexion net joint torque slowly increases due to the requirements imposed by the GRF. During the loading response, the knee performs controlled movement into flexion and then returns to a neutral position, which is associated with leg shortening. This shortening is opposed by a strong knee extensor moment, resulting in considerable power absorption, which agrees with our simple biomechanical model. The hip undergoes a movement from flexed position at the initial contact into extension to facilitate the vertical position of the trunk. This is accompanied by a noticeable hip extensor moment required by the GRF passing the hip anteriorly. In this period there is a small burst of positive power generated in the hip that propels the COM forward. At the end of the stance phase, within the second double support phase the ankle goes rapidly into plantarflexion, which generates considerable positive power and energy as predicted by our simple biomechanical model. With a slight delay the power generation in the ankle is accompanied by a considerable power generation in the hip, which propels the leg into swing phase.

Figure 14 of the recorded EMG activity in the major lower limb muscles, shows timing and intensity of activity that corresponds well with the magnitude of joint moments throughout the gait cycle.

The comparison of the simple biomechanical model of bipedal walking and the measurement of human walking shows close resemblance. The magnitude of joint moments closely relates to the external moments generated around each joint by the GRF, while power generation and absorption just before and during the double support phase corresponds to energetic considerations of the inverted pendulum model.

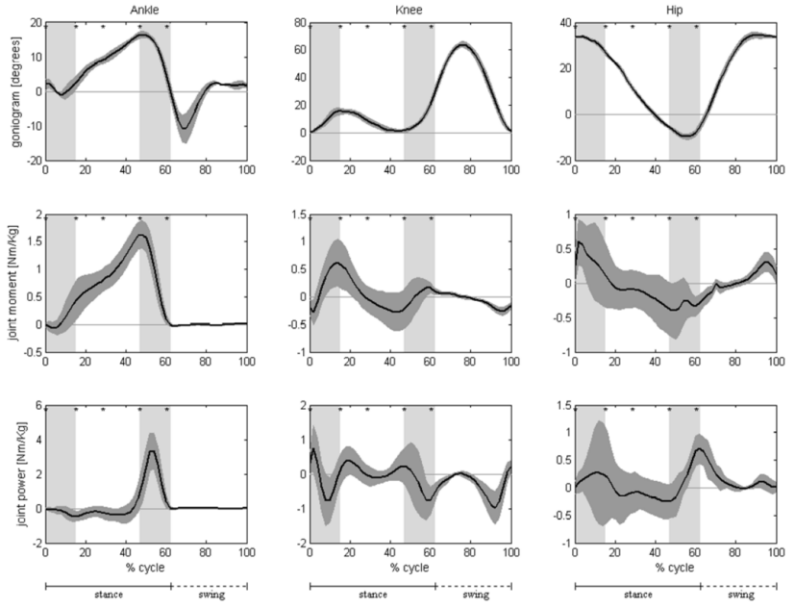


Figure 13. Graphs of joint goniograms, muscular moments and powers in the ankle, knee and hip joints throughout the the gait cycle of normal walking. Negative values in goniograms represent extension; negative values of moments represent flexor muscular moments; negative values of power represent absorption of energy. Stars above the curves denote time instants that correspond to the snapshots from Fig. 11. Mean values and standard deviations for normal walking were adapted from Winter (1991). Both periods of double stance are also shown.

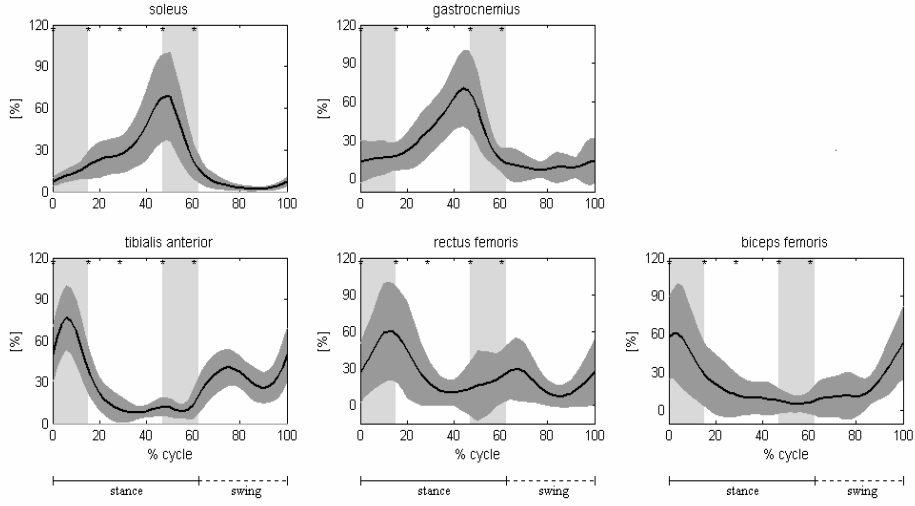


Figure 14. Graphs of electromyograms from the selected muscles of lower extremity throughout the gait cycle. Mean values and standard deviations are for normal walking were adapted from Winter, 1991. Both periods of double stance are also shown.

2. Orthotics

In pathological cases where normal walking function is either severely impaired or impossible due to diseases, injuries or other pathological changes (such as amputations of a part of the lower limb) of a neurological or musculo-skeletal nature, various orthoses and prostheses can be effectively used to regain adequate functional abilities.

Let us consider the stance phase of walking in a child with a cerebral palsy that resulted in pathological changes of the musculo-tendon structures, predominantly around the knee joint (Figure 15). We can see that the child is performing the initial contact with foot flat, rather than with the heel. This is primarily because the knee is forced to a flexed position due to muscle contracture, which occurs due to imbalance between the spastic hamstring muscles and the weakened knee extensors. The loading response is also very different from normal walking. The knee remains in a pronounced flexed position; therefore the child uses a considerable ankle joint moment to shift the centre of pressure (COP) where the GRF acts. By doing that, the GRF passes through the knee joint, thus minimizing the requirements for the weakened knee extensors. However, we can notice that this results in a considerable external hip flexor moment that needs to be balanced by increased output from the hip extensors that are also rather weak in such pathology. Therefore, this child is doing his best to find a feasible way to control the GRF and COP in such a way that enables him to walk. However, such walking is difficult; it requires considerably more energy, because throughout the whole of stance, the muscles need to produce considerable joint moments. In this situation one can think of applying a suitable knee orthosis that would externally and passively generate a knee extensor moment having two orthotic effects: 1.) the orthosis would stretch the knee flexor muscles that are spastic and in contracture, which brings a beneficial therapeutic effect and 2.) during the loading response, the orthosis would be able to provide more knee extensor moment that would reduce the requirements for ankle plantarflexion moment and hip extensor moment. The overall effect would be a more normal walking pattern. The knee orthosis could be a mechanical knee splint with an artificial joint approximately aligned with the physiological point of rotation in the knee and preloaded with a suitable spring.

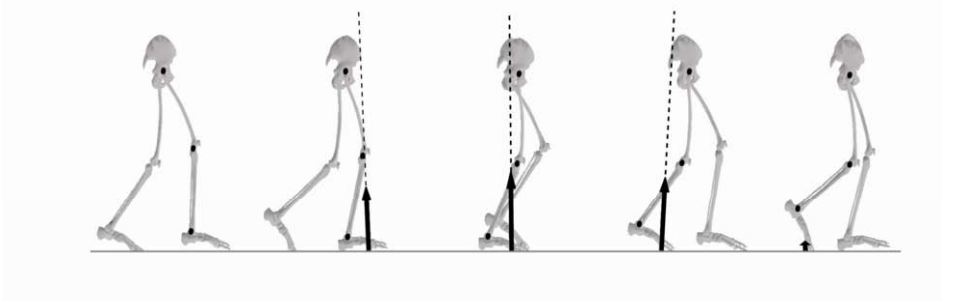


Figure 15. Presentation of pelvis and lower extremities movement together with a GRF in five consecutive snapshots within the stance phase of walking in a child with cerebral palsy.

By providing the above example, we can state the primary function of an *orthosis* is to change the biomechanics of the lower extremity. Application of an orthosis can have the immediate effect of substituting for the lost muscle function (in our case the weakened knee extensors). Also, an orthosis can bring about improvement indirectly by providing an opportunity to enhance or acquire motor skills. In our example, this means that the child whose walking pattern is enhanced could make initial contact on the heel, followed by the gradual weight transfer over the entire sole of the foot, which is not the case in his regular walking. Walking on tiptoes is rather unstable, therefore by enabling a more normal weight transfer the orthosis would also give the opportunity to the child to practice and improve his balance skills.

When designing an optimal orthosis for a given pathological case, one first needs to identify the primary cause for the observed changes in the walking patterns in order to determine what kind of orthosis would be appropriate, i.e. foot orthosis, ankle-foot orthosis, ankle orthosis etc. When this is determined, the mechanical properties of an orthosis need to be determined; the form, resting position-angle, the stiffness. Two major design principles need to be considered when making an orthosis. The first principle relates to GRF control. Basically, this principle derives directly from our simple biomechanical model of walking where we have shown that the GRF determines the required joint moments in the hip, knee and ankle joints. Alternatively, we can say that the joint moments developed by muscle activity shift the COP and GRF. Therefore, when designing an orthosis we need to anticipate what kind of changes the mechanical action of an orthosis will introduce in the kinematics and kinetics of walking in each particular case.

The second design principle relates to a so called three-point pressure control. The purpose of this design principle is to influence joint stability and/or mechanical properties by providing the first point of pressure above the axis of rotation, the second point of pressure below the axis of rotation and finally to provide the third point of pressure that acts in the opposite direction at or near the axis of rotation. This is illustrated in the Figure 16 where a KAFO (knee-ankle-foot orthosis) and AFO (ankle-foot orthosis) are shown with the indicated three points of pressure in each case. Special care must be given to the mentioned three points of pressure where the orthotic forces are transmitted onto the skeletal system. The areas needs to be large enough and shaped such as to enable even pressure distribution over the skin in contact with an orthosis. The mechanical action in each of the artificial joints of an orthosis may be different. For example, a KAFO may be designed in such a way as to lock the movement in the ankles and the knees thus providing stability during the stance phase to a person who is paralyzed but with some remaining function in the hips which are used together with the crutches for propulsion during ambulation. On the other hand, the action of a knee orthosis in the above example of crouch walking in a CP child must be such as to provide just enough stiffness to adequately modify the COP and GRF during walking.

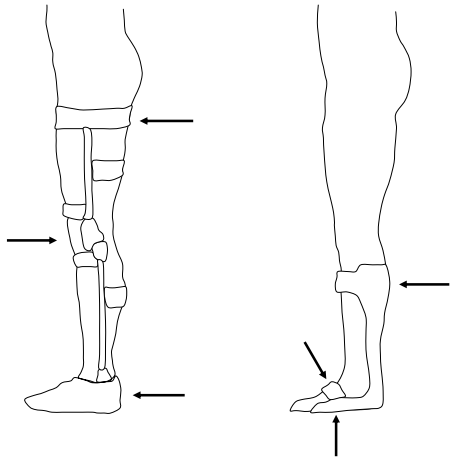


Figure 16. Three-point pressure examples in KAFO (knee-ankle-foot orthosis) and AFO (ankle-foot orthosis).

3. Prostethics

Unlike *orthoses*, which are generally used to augment function of the existing biological structures, *prostheses* are used to completely replace, not only lost function, but also a lost part of the biological structure – part of a limb. Although some similarities exist between the design principles for orthoses and prostheses, there are also major differences. Let us consider the application of an above-knee prosthesis, which is also called a trans-femoral prosthesis. Such a prosthesis is fitted after the amputation of the leg at the level between the hip and the knee joint. The remaining thigh is called a stump. Figure 17 illustrates the remaining leg and a trans-femoral prosthesis. The interface between the stump and the prosthesis is the socket, which is a critical element of prosthesis. It needs to provide a firm contact between the stump and prosthesis as well as even distribution of the pressure over the entire stump-socket contact area in order to avoid damage to the skin of the stump. Below the socket there is an artificial knee joint that can be a simple hinge joint or a more sophisticated polycentric knee that resembles the movement in a human knee joint. Below the knee joint there is a pylon that connects the ankle-foot complex with the knee.

Figure 18 shows the swing phase of walking with a trans-femoral prosthesis. Since the knee and the ankle-foot complex are passive, the user of a trans-femoral prosthesis needs to learn to control the movement appropriately. First, the hip flexor muscles accelerate the hip in flexion as in normal walking. However, this also initiates the acceleration of the artificial knee joint into flexion. The artificial knee joint has certain damping, which can be mechanical friction or viscous damping of the hydraulic fluid in the joint, that is set-up such to assure appropriate swing time of the artificial lower leg. Immediately after the leg is in the air, a brief activity of hip extensors is needed to reverse the motion in the knee joint in order for the leg to be fully extended when impacting the ground, which is vital for the stability of the limb in the following stance phase.

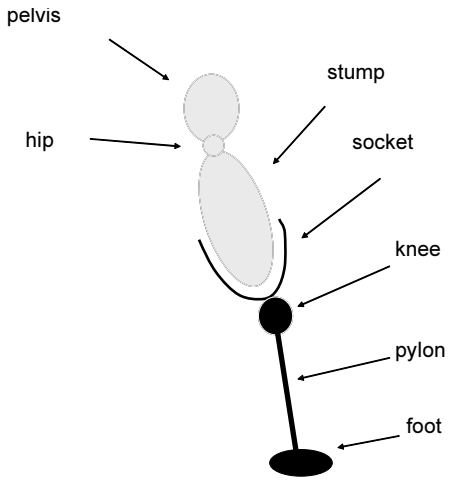


Figure 17. Schematic presentation of various components of a trans-femoral prosthesis – sagittal plane view.

Figure 19 shows the stance phase of walking with a trans-femoral prosthesis. At the moment of impact, the knee becomes locked. This is achieved by the mechanical design of the artificial knee, which incorporates a weight-dependent brake mechanism. The foot-ankle complex must be designed in such a way as to allow for a plastic deformation, which enables the power and energy absorption that is an integral part of bipedal walking. The control of GRF must be such that the GRF vector always passes in front of the artificial knee joint to guarantee stability of the artificial leg. During mid-stance, the leg rotates around the passive ankle-foot complex where some of the movement is enabled by a compliant ankle and most of the rotation is enabled by the circular shape of the foot sole. During this roll-over, the foot undergoes elastic deformation, thereby storing mechanical energy, which is released at push-off to provide some limited propulsion power. Most of the propulsion power comes from the hip joint, which generates positive power and energy during the first half of stance. This is possible because, at the beginning of the stance phase, the trunk is inclined anteriorly, which also shifts the GRF more anteriorly, thereby reducing the duration of the deceleration period in which the horizontal force is directed opposite to the walking direction. The biomechanics of walking with a trans-femoral prosthesis requires that a user learns a new motor control scheme for safety and efficiency, which is due to a fixed damping confined only to one, selected walking velocity. Computer controlled artificial legs also exist, where damping of the hydraulic knee joint is adaptively controlled in order to accommodate different walking speeds. Walking with a trans-femoral prosthesis is characterized by higher energy consumption (from 30% to 70% more than in normal walking) and higher vertical GRF (up to 30% more than in normal walking).

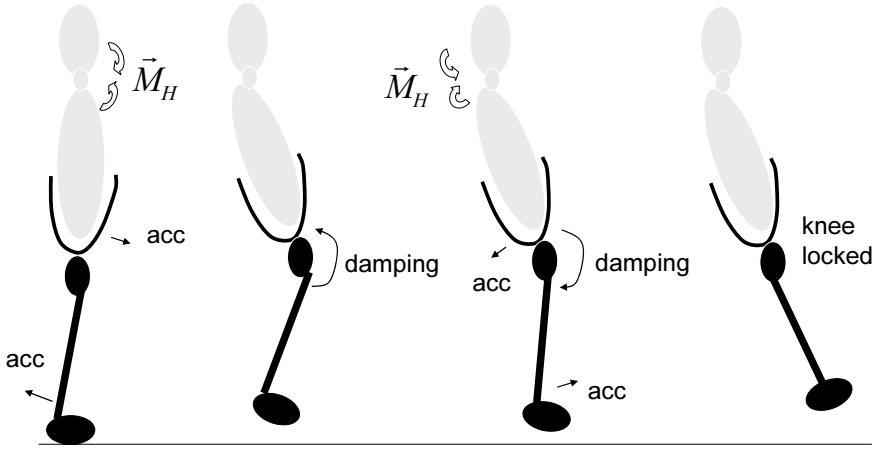


Figure 18. Control of a trans-femoral prosthesis during swing phase.

Walking with a trans-tibial or below-knee prosthesis poses similar gait characteristics as with a trans-femoral prosthesis, however they are less pronounced, it is more energy-efficient and walking appears more normal.

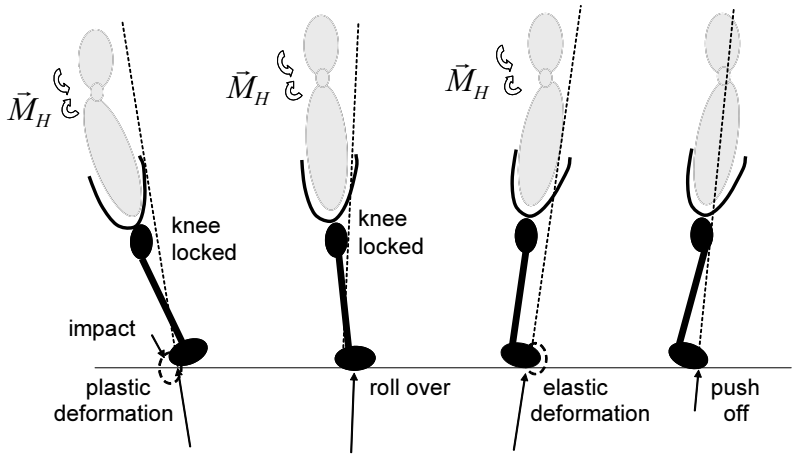


Figure 19. Control of a trans-femoral prosthesis during stance phase.

References

[1] Gage, J., R., (editor), *The Treatment of Gait Problems in Cerebral Palsy,* Mac Keith Press, London, UK, 2004
[2] Inman, V. T., Ralson, H. J., Todd F., *Human Walking,* Williams and Wilkins, Baltimore, USA, 1981
[3] Perry, J., *Gait Analysis: Normal and Pathological Function*, SLACK Incorporated, Thorofare, USA, 1992
[4] Winter, D. A., *The Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological,* University of Waterloo Press, Waterloo, Canada, 1991
[5] Zatsiorsky, V., M., *Kinematics of Human Motion. Human Kinetics,* Champaign, USA, 2002
[6] Zatsiorsky, V., M., *Kinetics of Human Motion. Human Kinetics,* Champaign, USA, 2002
[7] D. Popović and T. Sinkjaer, *Control of Movement for the Physically Disabled*, Springer, 2000
[8] Seymour, R., (editor), *Prosthetics and Orthotics: Lower Limb and Spinal,* Lippincott Williams & Wilkins, Baltimore, USA, 2002

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-343

VI.2. Basic Functional Electrical Stimulation(FES) of Extremites – an Engineer's View

Tadej BAJD and Marko MUNIH
Faculty of Electrical Engineering
University of Ljubljana
Slovenia

Abstract. The historical development of electrical stimulators producing contraction of paralyzed muscles is briefly presented. The influence of electrical stimulation parameters (amplitude of pulses, frequency, pulse duration, and duration of a pulse train) is explained. Special attention is paid to the description of the muscle recruitment curve. The phenomenon of reversed recruitment order, resulting in fatiguing of electrically stimulated muscle, is presented. The properties of surface electrodes (electrode size, polarity, resistance, and distance between electrodes) are examined. The use of surface electrodes made of metal plate or wire mesh, silicone impregnated with rubber, and conductive adhesive gel are discussed. The design of electrical stimulator circuits is also presented.

Keywords. Electrical stimulation parameters, muscle recruitment, muscle fatigue, electrodes, stimulators

Introduction: Brief History of FES

Medical applications of the discharge of electric fish for the treatment of headache and arthritis were first recorded by Scribonius Largus in the second century AD. However, the first electric devices making use of static electricity were developed in the eighteenth century. These were based on the principle of using friction of insulators, such as glass, to develop high voltages due to charge separation. The ninteenth century brought a more practical stimulator with an induction coil. The so-called 'Faradic stimulator' consisted of a mechanical interrupter in series with a battery and the primary winding of a transformer. The output from the secondary winding of the transformer was a series of pulses which were similar to the stimuli of the present day electrical stimulators. The Faradic stimulator was the first device that could produce controlled and sustained muscle contractions. The output frequency could be adjusted to easily exceed the stimulation rate of 100 pulses per second and the output level was also controllable. Finally, the twentieth century brought transistors, integrated circuits, and microprocessors permitting sophisticated electronic circuits to be incorporated into very small devices.

1. FES Parameters

Functional electrical stimulation (FES) is a rehabilitation technology that uses electrical currents applied to peripheral nerves. When a stimulating current is applied to the electrodes placed on the skin overlying sensory-motor structures, an electric field is established between two electrodes and ions will create a current in the tissue (Figure 1).

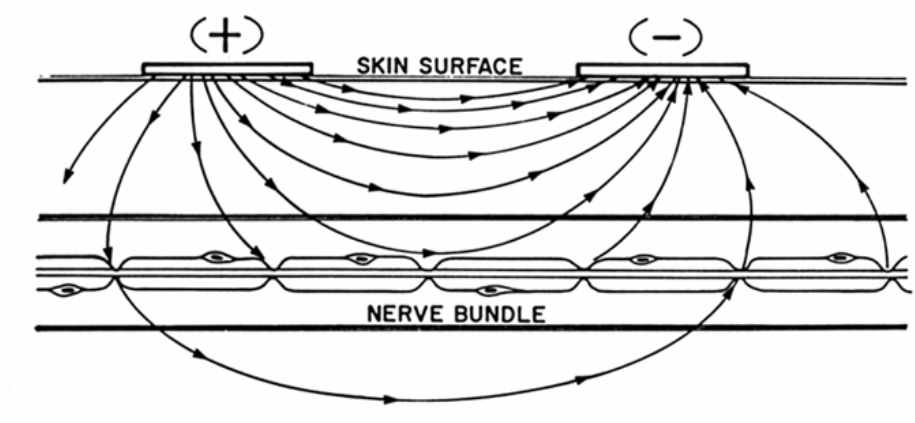


Figure 1. Electric field between a positive and negative electrode.

The ionic flow across the nerve influences transmembrane potential and can generate an action potential. The action potential propagates along the nerve causing contraction of a paralyzed muscle. In this way, FES provides restoration of movement or function, such as standing or walking by a person with a spinal cord injury.

FES is performed in a series of rectangular monophasic or biphasic (symmetrical or asymmetrical) electric pulses described by the following parameters: amplitude or intensity of pulses, frequency or pulse repetition rate, duration of single pulse, and duration of a pulse train. In most cases of surface FES applications, periodic monophasic or unidirectional pulses are used. Biphasic or bidirectional pulses prevent a slow deterioration of the electrodes, while the chemical conditions on the skin and in the muscular tissue remain unchanged.

1.1. Stimulator Circuit

With respect to the stimulator output circuit, electric pulses are either voltage- or current- controlled. Stimulators providing a *constant output voltage* can maintain the voltage desired, irrespective of resistance changes in the stimulated muscle tissue. Stimulators with current output stages make *constant current* pulses possible. An important difference between the two stimulator types becomes evident in the case of an improper contact between the electrode and the skin. In the case of a *constant current* stimulator, a smaller effective electrode surface results in greater current density, which can cause skin burns. With a *constant voltage* source, the resistance

increases due to insufficient contact, which results in a decrease of current and, consequently, of the muscle response, but causing no skin damage.

1.2. Muscle Recruitment

A muscle recruitment curve represents the dependence of *isometric* (measurement performed at a constant muscle length) joint torque upon the FES amplitude or pulse duration (Figure 2). The joint torque is not linearly dependent upon the stimulation intensity. Two nonlinearities occur - threshold and saturation. The increase in joint torque due to an increasing amplitude of electrical stimulation occurs as a result of activating new fibers in a nerve bundle laying in an electric field between the electrodes. The main reason why all nerve fibers do not react to the same stimulation amplitude is found in the differences in the stimulation threshold and various distances from the stimulation electrodes. First, the fibers closest to the electrodes are stimulated. In addition, the fibers with a greater diameter respond earlier. Beyond a certain stimulation intensity, the force of contraction no longer increases. At such a stimulation amplitude, all nerve fibers are excited, and a further increase of the stimulus does not increase contraction. In surface FES of knee extensors for example, the values of the stimulation threshold range between 20 and 60 V, while the saturation value is between 100 and 150 V.

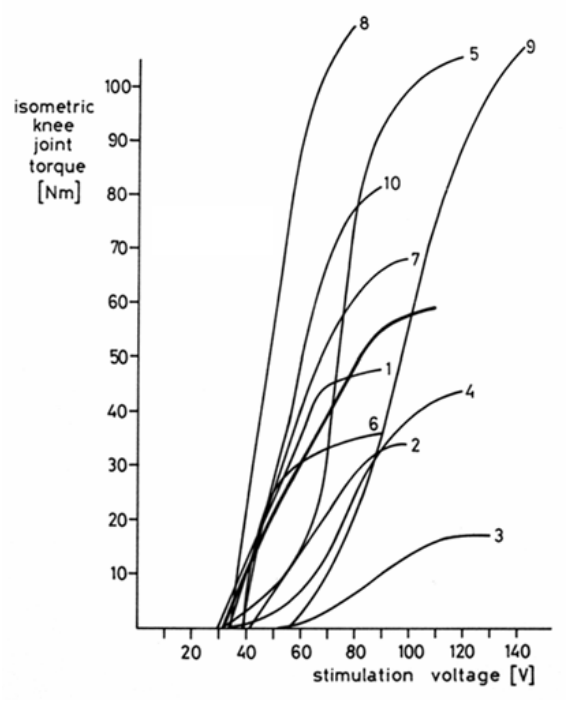


Figure 2. Muscle recruitment curves assessed in 10 paraplegic subjects after a FES restrenghtening programme.

A single stimulation pulse provokes only a short-lived muscle twitch of no more than 0.2 s. If electrical stimuli are repeated every second, a twitch occurs every second, between which the muscle relaxes. If the frequency of stimulation pulses increases up to 10 pulses per second (10Hz), between two twitches there is no time left for muscle relaxation. When measuring isometric contraction, we get twitching responses. This twitching is considerably reduced at stimulation frequencies between 15 and 20 Hz (Figure 3). At higher frequencies, the response is already smooth: this is known as *tetanic contraction*. The frequency at which the tetanic contraction occurs is called *fusion frequency*. It is not the same for all muscles and depends on properties of muscle fibers. Changes in stimulation frequencies also affect the intensity of the response. As regards the response intensity, slight losses are observed at lower stimulation frequencies. On the whole, the changing of frequency between 40 and 100 Hz causes small differences in the isometric torque as measured in the joint.

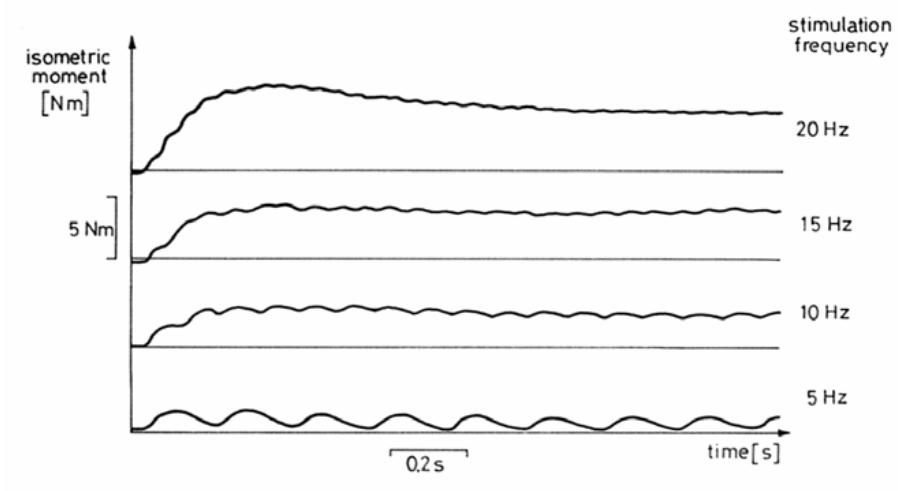


Figure 3. Influence of stimulation frequency on muscle response.

1.3. Muscle Fatigue

A low stimulation frequency results in less pronounced fatigue of the neuromuscular system. An electrically stimulated muscle fatigues more quickly than in the case of voluntary contraction. The main reason is the reversed recruitment order. It is not too serious an oversimplification to think of human muscle as a combination of slow fibers, capable of sustaining low levels of contractile activity without fatigue for prolonged periods, and fast fibers, capable of developing large forces, but fatiguing so rapidly that they can be used only in intermittent activities. Some muscles are predominantly made up of slow fibers, some predominantly of fast, and some of a given mixture of the two. All muscle fibers innervated by the same motoneuron have been found to be of the same type. Motoneurons innervating predominatly slow muscles have axons of small

diameter, and those supplying fast muscles have axons of a larger diameter. In a voluntary contraction of a normally innervated muscle, the slow fibers are recruited first, and as increased muscle force is required, the fast fibers are recruited. Slow fibers are, therefore, activated frequently, while fast fibers are employed only infrequently, during a burst of intense activity. When applying electrical stimulation, fibers with a greater diameter respond earlier. This are motoneurons innervating fast muscles. The normal order of recruitment is, therefore, reversed resulting in an increased fatiguing of electrically stimulated muscle. In addition, by electrical stimulation, the same nerve fibers are stimulated all the time, whereas with a healthy muscle the work is divided among different motor units of the same muscle. Also, due to relatively high stimulation frequency, the transmitter in the neuromuscular junctions is being exhausted, so the muscle stimulated soon shows signs of fatigue.

1.4. Duration of Stimulus

In a fashion similar to amplitude, pulse duration also exerts a direct effect upon the intensity of contraction. Here again, this is determined by threshold value and response saturation. When applying surface stimulation electrodes, the accompanying unpleasant sensation (when preserved) or even skin damage is due mainly to an excessively long duration of a stimulus, therefore short durations are used (0.1 to 0.5 ms), while the force of a paralyzed extremity is controlled by increasing the stimulus amplitude. Changing the pulse duration has little or no effect on stimulated muscle fatigue.

Functional movement of a paralyzed extremity cannot be obtained by a single electric stimulus, but requires a series of stimuli of a certain duration, following one another at an appropriate frequency. Such a series of stimuli is called a *stimulation pulse train*. In FES training of atrophied muscles, a stimulation pulse train is followed by a pause, and then by another stimulation train. The relationship of train duration and pause is often called the *duty cycle* and exerts an influence upon the fatigue of a stimulated muscle.

2. FES Electrodes

2.1. Anode & Cathode

A surface stimulation electrode is a terminal through which electrical current passes into the underlying tissue. At the electrode-tissue interface, a conversion occurs between the current of electrons driven through the wires coupled to the stimulator and the current of ions in the tissue. An electrode is usually made of metal. However, it may be made of a nonmetal, commonly carbon. The electrode through which current passes from the metallic or nonmetallic conductor *to* the tissue is called the *anode* and that through which current passes *from* the tissue to the conductor is the *cathode*. In electrical circuits, the current flows from the terminal at higher electrical potential to the terminal at lower electrical potential. In this way the *anode* is the *positive electrode* and the *cathode* the *negative electrode*.

2.2. Unipolar & Bipolar

Besides distinguishing between positive and negative electrodes, we also speak about *unipolar* and *bipolar* electrical stimulation techniques. With *unipolar* stimulation, one electrode is often considerably smaller than the other, whereas the electrodes used in *bipolar* stimulation both have the same size. In *unipolar* stimulation the smaller electrode is *negative* and is also called an *active electrode* due to the fact that in its vicinity there occurs depolarization of the membrane of nerve fibers. In motor nerve stimulation, the active electrode is positioned as closely to the *motor point* of the muscle as possible. The *motor point* is a site on the skin, where the amplitude of the *stimulus* required to fully activate the muscle is at a *minimum*, and where all of the motor nerve fibers are closest to the stimulating electrode. In multichannel electrical stimulation systems, it is possible to have a single anode and several independent cathodes or to have anodes and cathodes that are galvanically separated.

Let us examine four properties of surface stimulation electrodes and electrodes positioning, which influence the effectiveness of electrical stimulation: electrode size, polarity of electrodes, resistance, and distance between the electrodes.

(i) Electrode size: Electrical stimulation is applied to a nerve fiber, since muscle fibers have a considerably higher stimulation threshold. Thus we can say that larger electrodes are used to stimulate the nerve endings spreading all over the underlying tissue, whereas smaller electrodes are applied to influence the nerve when the latter come closer to the skin. Using larger electrodes, stronger contraction is obtained along with a reduced current density and a less pronounced unpleasant sensation on the skin. However, large electrodes permit no selective choice of a desired movement of the stimulated paralyzed extremity. The active areas of electrodes range between 2 cm^2 and 50 cm^2. Electrodes of 2 cm^2 to 4 cm^2 are used to stimulate the nerves near the surface, those of about 8 cm^2 for the stimulation of smaller muscles, while electrodes of 25 cm^2 or more are used in case of larger muscles.

(ii) Polarity:A positive and a negative electrode are placed along the muscle to be stimulated. Considering their polarity, the electrodes are positioned so as to provoke an optimal movement from the functional point of view. Stronger movement is usually obtained by placing the positive electrode distally.

(iii) Resistance:It is desirable that the resistance should be as low as possible in order to avoid energy losses before the stimulation has reached the neuromuscular tissue. The impedance between the electrode and the skin is frequency dependent. The DC (or low frequency) impedance tends to be several orders of magnitude larger than the impedance at higher frequencies. Nominal values of 2 kΩ are encountered. Contact conduction is increased by moistening the electrodes with water or special conductive electrode gels. Adipose tissue offers high resistance to electrical currents and so higher stimulation amplitudes must be used, causing pain in the skin as a side effect. Bones, too, are very bad conductors of electric current; electrical stimulation cannot reach muscles which are behind them.

(iv) Distance between electrodes:The greatest current density appears at the skin-electrode contact and tends to decrease with distance from the electrodes as the flow spreads out over a larger area. Closely spaced, small electrodes generally make the effective area of stimulation rather superficial due to the lower impedance of the current path through proximal tissue. Deeper tissues will be stimulated by using a greater distance between the electrodes. Increasing the electrode separation leads in general to an increase of the maximal achievable force. If the skin between the

electrodes is moist, this causes the current between the electrodes to flow to the skin which results in a burning sensation and a slight or no muscle contraction at all.

2.3. Electrode Design Criteria

The design criteria for surface stimulation electrodes are: physical comfort to the skin, electrical surface area greater than four square centimeters to prevent skin irritation, use of hypo-allergenic materials, flexibilty to follow body surface, ease of attachement and ability to remain in position for a duration of at least one active day, reusable, low cost, reliable means of connection to stimulator, resistant to medical solvents and electrode gels, low and stable electrical resistance.

The simplest surface electrodes consists of a metal plate or metal wire mesh coated with fabric or sponge. Common materials used are stainless steel, silver-silver chloride, platinum, or gold. For safety purposes, the upper part of the electrode is covered with a non-conductive material. The electrode is applied after having been moistened with water. Such electrodes are usually fixed on the extremity by means of velcro or elastic bands. With the development of longer term surface electrodes, these electrodes are often used for site evaluation either in pain treatment or stimulation provoking muscle contraction. Small, button-shaped electrodes of similar design are highly suitable for the stimulation of a single nerve. Here, the metal plate is coated with several layers of gauze so that the electrode might retain the moisture required for as long as possible.

Surface electrodes made of silicone impregnated with carbon are applied to the skin surface with conductive gels and held in place with adhesive patches. A too thinly or unevenly spread gel increases current density at certain points, thereby bringing about a danger of burns. Electrodes may be left on the skin for several days a time. Another important property of electrodes made of conductive rubber is their flexibility making them adaptable to any part of the body. These electrodes can be shaped by cutting, so as to adapt them as much as possible to a proper stimulation site.

Conductive adhesive gel electrodes provide self-adhesion when applied (Figure 4). Karaya gum was the first of these adhesive gels. Later a variety of electrically conductive polymers were developed, enabling good contact with the irregularities of the skin surface and a more uniform current density at the electrode-skin interface. These electrodes can be used for extended periods with minimal skin irritation. An important breakthrough in electrode design was made by Axelgaard. This is a flexible electrode which adheres well to the patient's skin, is easily removed therefrom, and is able to move with the patient's skin ensuring proper placement of the electrode. The electrode is in the form of a knit conductive fabric. Conductive fibers include a blend of stainless steel and polyester. The fabric can be stretched up to about 20%. A conductive adhesive fills interstitial areas of the knit fabric and adheres the electrode to the patient's skin. A non-conductive sheet on the other side of the knit fabric prevents undesired electrical contacts.

2.4. Problems with Electrodes

When improperly handled, electrodes can damage the skin in the contact area. Burns typically occur underneath the anode, but not the cathode, when using identical surface electrodes. Another problem resides in a precise electrodes positioning along a muscle. Sometimes a displacement of a few millimeters completely changes the muscle response. This happens when a selected nerve (e.g. peroneal nerve) ought to be

stimulated by surface electrodes. Surface electrodes may excite pain receptors in the skin, although patients sensibility may be reduced to such an extent that the sensation of pain is not critical. Another problem is undesired motion of the skin with respect to the neuromuscular tissue. Even though an electrode seemingly occupies the same place all the time, its distance from the nerve is not constant. This is one of the reasons why the movements caused by electrical stimulation cannot be easily repeated. Another limitation is that small muscles usually cannot be selectively activated and deep muscles cannot be stimulated without first exciting the superficial muscles. Relatively high voltages, sometimes in excess of 100V, between electrode pairs cause hazards for the patients and the personnel that treat them. Finally, the applicability of the surface stimulation electrodes depends on fixation problems. Stretchable garments with electrodes already mounted in appropriate locations have been developed by several manufacturers to simplify the application of electrodes to the skin surface. In the case of lower limb stimulation, fixation problems can be overcome by specially designed trousers carrying stimulation electrodes and cables. Such stimulation equipment is comfortable and easy to handle. In the non-invasive, upper limb neuroprosthesis, the surface stimulation electrodes were built into an elegant, self-aligning, and flexible splint. The splint provides additional fixation of the wrist joint and allows the entire electrode array to be positioned within a few seconds.

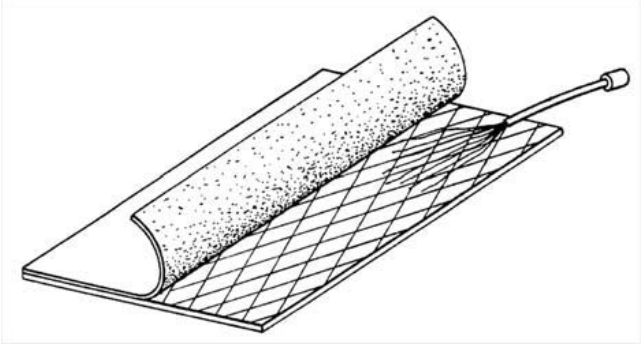


Figure 4. A conductive adhesive gel electrode with a portion of nonconductive sheet peeled back showing the knit fabric.

It is not difficult to realize that most of the inconveniences of surface stimulation electrodes can be overcome by the use of implanted electrodes. Nevertheless, because of their simple non-invasive application, surface electrodes will remain of use in therapeutic treatments.

3. Electrical timulators

Electrical stimulators comprise an input circuit, pulse generator, output stage, and power supply (Figure 5). The input into the electrical stimulator is represented by a control signal automatically switching electrical stimulation pulses on or off or may be under the patient's voluntary control. The output electric pulse current is led, via

electrodes, to the selected stimulation site. The type of stimulation pulse is determined by a pulse generator. The output pulses provided by the circuit are considerably lower than the surface stimulation pulses required for a functional movement of an extremity but are of an appropriate frequency and duration. The output stage ensures energy for the electrical stimulation of paralyzed muscle in either constant current or the constant voltage outputs. In case of current stimulation, the internal resistance of the end stage is considerably higher than the tissue resistance between the electrodes. The current source of stimulation pulses provides a constant current irrespective of the resistance of the skin and the tissue between the electrodes. In the case of constant voltage output stages, skin resistance is lower than that between the electrodes, and the stimulator provides a constant voltage independent of the skin and tissue resistance. A power supply provides the energy necessary for the operation of particular electronic circuits (low voltage) and the electrical stimulation itself (high voltage). Electrical stimulators are usually battery powered with stimulation pulses with an amplitude of more than 100 V. A high voltage for the output stage is obtained from low battery voltage by means of a voltage convertor.

The development of electronics made it possible for Wladimir Liberson to develop a stimulator to preventing foot-drop in hemiplegic patients. It was triggered by a heel switch in the shoe of the affected leg. Stimulation electrodes were positioned above the peroneal nerve in the popliteal fossa, behind the knee joint. Each time the patient raised the heel, the heel switch triggered the stimulator, causing the nerve to cause the extensor group of muscles to contract and dorsiflex the ankle and so lift the foot. Advanced versions of these dropped foot stimulators were developed in Ljubljana, Slovenia and Salisbury, UK. Both stimulators were applied to a large group of stroke patients. In the WalkAide peroneal nerve stimulator, a tilt sensor triggers ankle dorsiflexion.

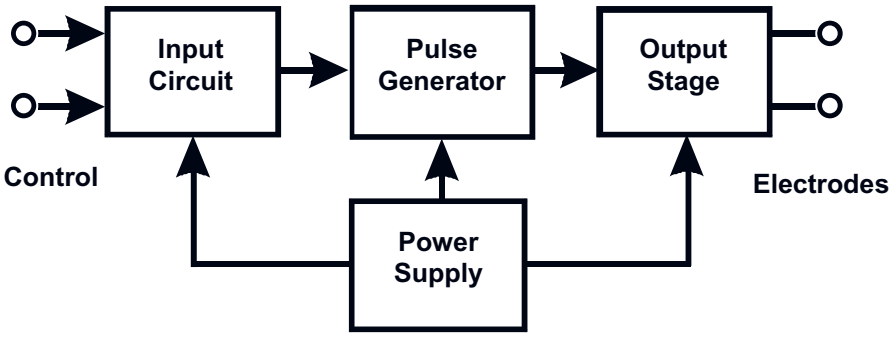


Figure 5. Block scheme of an electrical stimulator.

A minimum of four channels of FES were used for synthesis of a simple reciprocal gait pattern in completely paralyzed paraplegic subjects. During reciprocal walking, the stimulator must be controlled through three different phases of walking: right swing

phase, double stance phase, and left swing phase. This is achieved by two hand-push buttons built into the handle of the walker or crutches. When niether of the push buttons is pressed, both knee extensors are stimulated providing support to the body. On pressing the push button in the right hand, the peroneal nerve is stimulated in the right leg eliciting the flexion reflex presented by simultaneous hip and knee flexion and ankle dorsiflexion. The same is true for the left leg. The Ljubljana FES walking system consists of two small two-channel stimulators attached to each leg. Only three electrodes are applied to a single leg in order to produce knee extension and flexion responses. As both activities never occur simultaneously, the distal electrode placed over the knee extensors represents the common electrode for both stimulation channels. Using the same principles as the Ljubljana FES system, and adding two channels of stimulation to both hip extensors, the FDA-approved Parastep surface stimulation system was developed. A new multipurpose programmable transcutaneous electric stimulator, Compex Motion, was developed to allow users to design various custom-made neuroprostheses, neurological assessment devices, muscle exercise systems, and experimental setups for physiological studies. The Compex Motion stimulator can generate any arbitrary stimulation sequence, which can be controlled in real-time using any external sensor.

Despite the fact that correct application of a stimulator implies no danger either for the patient or therapist, we wish to emphasise some important safety points. When buying a new electrical stimulator, be sure that it carries the CE mark. Every stimulator has its own characteristics which must be indicated in the instructions of use, an obligatory accompaniment of any commercially-available stimulator. Special attention should be paid to stimulation parameters: pulse duration, frequency, and maximal current or voltage. Also of utmost importance is the information on the proper use of surface electrodes. The use of electrical stimulators might be dangereous in case of patients with an implanted pacemaker. Simultaneous use of electrical stimulators and high frequency surgical devices is prohibited. Electrical stimulators might not work properly in close proximity to microwave devices. Transthoracic positioning of FES electrodes may cause fibrillation of the heart. The stimulator should not be switched on in case of short-circuit of the electrodes. It is true, that electrical stimulators are made in such a way that a short-circuit of several minutes does not damage them. However, one should not take unfair advantage of this property.

Further Reading

[1] L.A. Benton, L.L. Baker, B.R. Bowman, R. Waters, *Functional electrical stimulation – A practical guide*, Rancho Los Amigos Hospital, 1980
[2] L. Vodovnik, T. Bajd, A. Kralj, F. Gračanin, and P. Strojnik, Functional electrical stimulation for control of locomotor systems, *CRC Critical Rev. Bioeng*. **6** (1981), 63-131
[3] A.R. Kralj and T. Bajd, *Functional electrical stimulation: Standing and walking after spinal cord injury*, CRC Press Inc., 1989
[4] L.S. Illis (ed.), *Neurological Rehabilitation*, Blackwell Scientific Publications, 1994
[5] D. Popović and T. Sinkjaer, *Control of movement for the physically disabled*, Springer, 2000
[6] P.J. Rosch and M. Markov (eds.), *Bioelectromagnetic Medicine*, Marcel Dekker Inc., 2004
[7] K.W. Horch, GS Dhillon (eds.), *Neuroprosthetics – Theory and Practice*, World Scientific, 2004
[8] J.D. Bronzino (ed.), *The Biomedical Engineering Handbook*, CRC Press and IEEE Press, 2000
[9] M. Akay (ed.), *Wiley Encyclopedia of Biomedical Engineering*, John Wiley&Sons, 2006
[10] L. R. Sheffler, J Chae, Neromuscular electrical stimulation in neurorehabilitation, *Muscle Nerve* **35** (2007), 562-590

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-527-3-353

VI.3. Rehabilitation Robotics

Marko MUNIH and Tadej BAJD
Faculty of Electrical Engineering
University of Ljubljana
Slovenia

Abstract. The paper presents the background, main achievements and components of rehabilitation robotics in a simple way, using non-technical terms. The introductory part looks at the development of robotic approaches in the rehabilitation of neurological patients and outlines the principles of robotic device interactions with patients. There follows a section on virtual reality in rehabilitation. Hapticity and interaction between robot and human are presented in order to understand the added value of robotics that cannot be exploited in other devices. The importance of passive exercise and active tasks is then discussed using the results of various clinical trials, followed by the place of upper and lower extremity robotic devices in rehabilitation practice. The closing section refers to the general importance of measurements in this area and stresses quantitative measurements as one of the advantages in using robotic devices.

Keywords. Robot, haptic interface, virtual reality, measurement

Introduction

The application of robotic approaches in neurological patient rehabilitation was introduced almost two decades ago [1]. Even though the number of robotic rehabilitation systems is large, the number of clinical trials remains quite limited. In fact, it is not yet clear what characteristics should be incorporated in a therapeutic robotic assistant platform.

Conventional therapeutic techniques and robot assisted techniques must not be perceived as two opposing modalities, but rather as two complementary approaches. Two very positive aspects of robotic therapy are high repeatability and automatic measurement during exercise. In contrast, the activities of a therapist unavoidably include many subjective elements, but an experienced therapist has an in-depth understanding of the individual patient which no high-tech device can ever possess. In future, robotic therapy will complement existing clinical practice: by reducing a therapist's workload, providing less costly and more extensive therapeutic programmes; by using quantitative measures of an intervention or injury and last, but not least, by new insights into the treatment process.

One of the natural common points of conventional and robotic therapy is related to haptic interaction between a patient and therapist. This approach gains specific importance for instance in the Bobath concept, which is based on an holistic approach to a patient, together with the International Classification of Functioning, Disability and Health (ICF) of the World Health Organization. At present, the emphasis is directed to problem solving, on predefining some intermediate goals and then adjusting theraphy

in gradual steps to finally learn complete movement, for example, in spasticity. This approach highlights and relies on the plasticity of the nervous system. Solving problems includes activities focused at a specific task, motivation, planning, interaction with the environment, selective placement of attention by stimuli and biomechanical aspects for effective and functional direction of movement. Thus, haptic interaction between a patient and a therapist must not be invasive and unilateral. On the contrary, it must evoke all the capabilities of a person leading to functional movement by minimum intervention.

To enable the best possible imitation of natural human interaction, while using a machine as a rehabilitator assistant, a robotic therapist must be highly consistent with human interaction. It is no coincidence that the original ideas which brought these systems to life did not arise in the rehabilitation field, but rather in neurophysiology and haptic perception, including sensimotor learning [7-9]. Robotic therapist devices differ in a number of ways from industrial robots. These are traditionally moved between two known points along a defined trajectory. Maintaining a known and highly accurate position is essential. In contrast, the robotic therapist must be programmable and adjustable. This can be achieved by impedance control algorithms or by passive manual control as in backdrivable robot manipulators including MIT-Manus and Braccio di Ferro. Other devices designed for robotic therapy might use industrial manipulator technology and have high manipulator arm mechanical impedance due to high gearing ratio and admittance control scheme.

Robotic devices may be used in positional trajectory mode for enforcing *passive movements*, for simple positional control of extremities, which, does not involve active participation of a patient on either neuromuscular or sensory levels. Such *passive* exercise, quite the opposite of the Bobath concept, using a robot contributes to rehabilitation, at least in specific clinical conditions [10-12].

Other studies [13-15] indicate better results with the application of techniques that consider the adaptive nature of the nervous system by solving problems with a suitable adaptability level. These techniques include *active assisted exercises* in which the robot moves the extremity along a predetermined trajectory. In the *active constrained* exercises, the robot provides higher opposing forces. The movement may be limited inside some 3D virtual space, with opposing forces applied when the subject tries to move outside this region. Similarly, one could imagine a ball object that limits movement toward the center (inside), with completely empty space all around. Furthermore, in *active resistive* exercises, the robot opposes the intended movement. In *adaptive exercises,* the robot is providing a previously unknown dynamic environment to which the subjects have to react and adapt.

Passive exercises need no input from the patient, while active constrained exercises require at least residual movement cababilities or subconscious sensory-motor co-operation. Active resistive and adaptive exercises require cooperation with sufficient volitional motor activity. Active resistive and adaptive exercises are therefore not a suitable choice for persons with larger movement deficits, since they usually cannot use them independently. Still, such persons may exploit robot-therapeutic exercises in which minimum co-operation on their side will help them in using and strengthening their remaining abilities.

1. Virtual Reality in Rehabilitation

Most of us are familiar with virtual reality (VR) technology from entertainment (e.g. games) and military simulations. Lately, its use has expanded to other areas, for instance computer-aided design (CAD), architecture, general virtual presentation of data and to medical applications. Medical applications of virtual reality include training in medicine and surgery, modelling of hard and soft tissue, image displays, remote surgery, ergonomy and rehabilitation. In rehabilitation, virtual reality is used for training and measuring the motor ability (e.g. in hemiplegia, paraplegia, Parkinson's disease).

Virtual reality is usually understood as a three-dimensional computer model which primarily defines the geometrical model (kinematic or dynamic model) of different virtual objects and their environment. It is possible to define not only the image of objects, but also their inherent physical characteristics. Such a virtual world can change in time - not only the position and orientation of particular items, but also dynamic characteristics of the surroundings, such as friction and gravity. Realistic visualisation and deep presence of person in VR are mostly dependent on high-quality graphics which take the user into the virtual world. The impression of virtual reality depends on the realistic appearance of the scene and activity. Simpler non-immersive visualisations include a 2D computer screen, a projection screen and 3D projection techniques in the environment. The immersive method is more realistic, enabling the user to percieve a full 3D visual field by use of special glasses.

The dynamic characteristics of objects, including mass, moment of inertia, forces, torques, and torques and forces resulting from the interaction of objects and, for instance, compliance (stiffness) of objects, roughness and smoothness of objects cannot be visually detected or presented. These require the sense of touch.

2. Hapticity

The verb ‘απτω’ - hapto is originating from Greek for reaching, holding and touching. If a person exerts a force F onto a mass m, which is integrated in the environment via the damper h and the spring k, the movement x is determined by the differential equation:

$$F = \left(m\frac{d^2}{dt^2} + h\frac{d}{dt} + k \right) x = m\ddot{x} + h\dot{x} + k\,x \qquad (1)$$

If the coefficients m, h and k are constant, the force depends on the position (distance to origin), velocity and acceleration. The first force term in the equation depends on mass and acceleration $m\ddot{x}$, with m representing the mass of e.g. a cube on icy surface. Significant force will be needed only during acceleration and deceleration with high m, while constant velocity movement requires no force. Only a small force is needed for the acceleration and deceleration of a light styrofoam cube with a small m. The force can, as a second example, depend on the term $h\dot{x}$, where h stands for damping of e.g. of an oar in water. If the velocity $\dot{x}$ is low, only a small force is needed, whereas at higher velocities, a considerably greater force is required. Another medium, like oil or a spoon in honey or air, represents a different damping coeficient

value h and thus a different force at the same velocity $\dot{x}$. The damping force is small when an oar moves in air, but if the velocity is high, in the case of an airplane propeller, the pull/push force becomes significant. The impact of stiffness k is presented by means of a spring. The greater the deviation x, the greater the force. Also, a higher spring stiffness coeficient k value, with x unchanged, requires greater force.

A general body in a virtual environment, onto which a person exerts a force, is not a point mass, pure damper or spring but rather a combination of all three terms in the above equation that becomes more complex. This is typical for haptic touch in a virtual environment, where none of the three coeficients of the equation is constant. The values of m, h and k change locally in the virtual environment. Actually, one of the three parameters, e.g. k, has six values at the same point (or position) in space: k_x, k_y and k_z along individual axes of space, and k_{Rx}, k_{Ry} and k_{Rz} around individual axes of rotation – thus it has *six degrees of freedom* (DOF) (Fig. 1). An empty space has nil or small values of all six k. A flat wall in a virtual environment is represented as a high k horizontally. A wall can also be slippery as ice or non-slippery if coated with rubber, which is set with the other two (transversal) k parameters of the wall. The remaining three k elements stand for the three rotational stiffness coeficients. Similar six-dimensional understanding applies to the other two parameters - h and m.

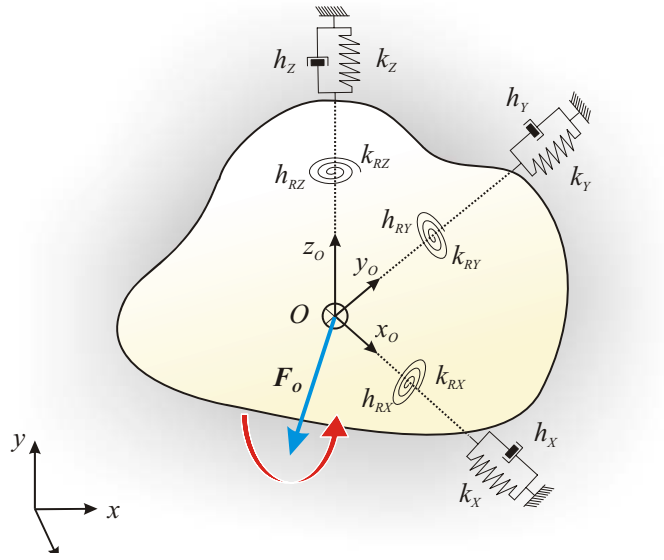


Figure 1. Second-order mechanical system with six DOF, three translations and three rotations and m, h and k

In rehabilitation haptic rendering is often used to denote a virtual tunnel that connects two extreme points of movement, the starting and the final one. The trajectory between them can run over a straight line or a curve having a different shape. All

elements m, h and k along the direction of movement equal zero, whereas, perpendicular to the direction of movement, the stiffness coefficient k increases by selected functions. This is reflected in the virtual pipe or tunnel that forces the user to the central curve line, where this force component does not exist.

In a series of attempts, should the user appear at exactly the same coordinate of space, carrying the same coefficients, the current values of $\ddot{x}$ and $\dot{x}$ (current acceleration and velocity) are very likely to differ, meaning that the force F felt by the user in a virtual environment with a haptic interface is also different. A haptic interface is, therefore, a robot capable of functioning in line with the above equation, contrary to the classic understanding of robot functioning. These are usually position-controlled, as referred to under *passive exercise*. Haptics in a virtual environment is achieved in real time by setting the interaction force in the point between the user and interface, considering all the numerous parameters mentioned above. It is understandable why this technology has become accessible only in the last two decades.

The connection of quantities F and x represents an energy contact, in reality the transfer of power between man and environment $F\dot{x}$ and thus controlled transfer of energy via man-machine interaction. In addition to the vision sensory pathway, the user participating in haptic interaction in a virtual environment also engages and uses touch, position, force and texture sensation in their body as well as the mechanisms of visual and tactile recognition, along with reflex and volitional control mechanisms, including the entire motor chain.

Further to vision and haptic modalities, further information can be supplied to the user via sound. Particular cases would be sound produced during movement along rough or smooth surfaces, along a ladder, or when typing on virtual keyboards.

3. Passive Exercise

CPM (Continuous Passive Motion) devices have been used in the postoperative rehabilitation of joints since approximately 1960. It was later established that joint immobilisation in a test group of rabbits led to great problems in restored mobility [16]. After five to six weeks of immobilisation, most joints developed moderate to extreme changes, including those in joint cartilage and bone mass, with changes already apparent in the second week. In clinical practice, this presented as joint contractures and reduced range of motion (ROM). To avoid this, they tried to maintain a good ROM by simple CPM devices which moved the limb over an arc corresponding to joint movement. Such devices consist of a simple motor with a mechanically or electronically adjustable range and velocity of movement. Most modern CPM devices adopt ideal joints with a fixed point of rotation, and can provide movement in one vertical plane (2D). Usually they cannot be reprogrammed and are controlled using the open loop principle, which means that their movement angles and forces are not measured immediately but are corrected and reset a few 100-times in a second.

Such CPM devices are very efficient in preserving range of movement, they reduce stiffness in joints, decrease the need for drug administration and shorten the length of hospitalization. Comparison of CPM and physiotherapy reveals no major differences as regards the above parameters, however in the majority of cases, CPM results in weaker muscles, delays in activation of extensors and stiffness of flexors [17]. It would be

difficult to expect anything else, since the CPM devices only move a person's extremities without activating the muscles.

4. Active Exercises

Several research groups developed robotic devices for rehabilitation of the upper and lower extremities. These devices, designed for exercise, use various methods of operation. *Active constrained* movements are made possible by most of the mentioned devices, while movements with *active resistance* are provided by MIT-Manus, Bi-Manu-Track and MIME [18-20]. *Adaptive exercises* are possible using Bi-Manu-Track and MIME [19, 20], in which case the healthy arm leads and the injured arm imitates movement, both moving simultaneously.

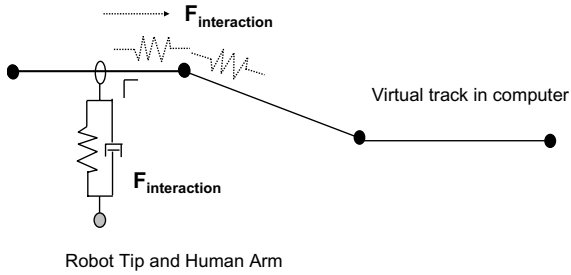


Figure 2. Trajectory in active constrained exercise

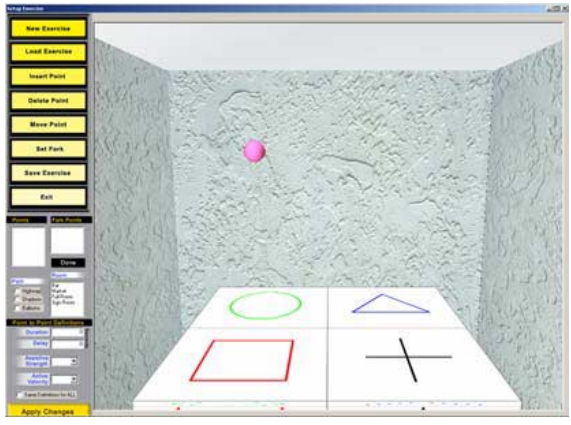


Figure 3. Example of environment in active constrained exercise, GENTLE/S project

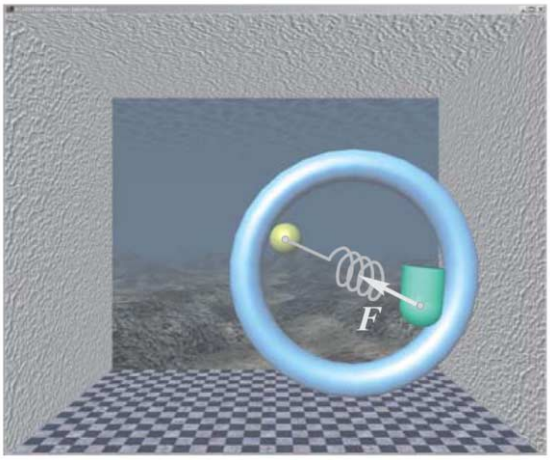


Figure 4. Example of active exercise with active resistance (torus), HIFE project

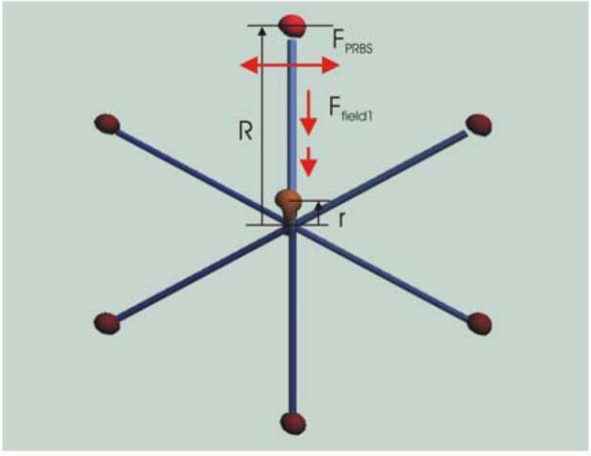


Figure 5. Example of active exercise with active resistance (pipe with a spring), I-Match project

The literature often contains descriptions of technical approaches and solutions. However, there is a limited number of clinical trials with clear goals and methodology that would systematically present the influence of robotic devices and existing modes of operation on the rehabilitation process. Among 17 known clinical trials, for a variety of reasons, only a handful can be used for direct comparison. None of them systematically investigates the influence of the method of applied exercise. Most trials have so far employed all three modes of operation, including: passive, active constrained and active resistive, regardless of the probability that one of these methods

might be more successful than others. Only Fasoli et al. [27] and Stein et al. [28] tried to specifically evaluate particular exercise versions. The findings of the Fasoli trial disclose that *active resistive* exercises are more useful than *active constrained* exercises as regards the upper extremities. Still, the repeated trial covering the same group did not reveal any differences between the exercises involving *active constrained* movements and those involving *active resistive* movements, perhaps because of the manner of data calculation. In the first trial Fasoli et al. [27] included in the group of an *active constrained* movement those who did not have sufficient motor function for the *active resistive* group. This was not the case in the trial conducted by Stein et al. [28]. Thus, the results of *active resistive* exercises were overrated in advance. These findings point to the fact that some robot-assisted therapies are less suitable for specific groups of patients. Therefore, the precise contribution of individual exercise methods to the rehabilitation of upper extremities in a hemiplegic person remains unexplained.

Other methods of exercise are perhaps equally or even more important for robotic rehabilitation, but they have not been investigated sufficiently, or at all. One such method is *gravity compensation* of the upper extremity. Most robotic devices provide some sort of support for the arm. The evolution of gravity compensation devices has been going on for decades – for example Sanches [29]. Beer et al. specifically investigated the implementation and reach of gravity compensation for the upper extremity [30-32]. Their preliminary research showed that poor coordination of muscle activation led to unexpected torques in the hemiplegic arm joints (e.g. shoulder abduction prevented the ability of elbow extension). Further trials revealed immediate improvement of motor abilities in persons after a stroke, if gravity did not influence the upper extremity. The active abduction of the shoulder was reduced and consequently extension in the elbow increased in static conditions. The latest results point to a similar mechanism functioning in dynamic circumstances. Such findings indicate improved motor abilities due to the use of gravity compensation. The second trial involving exercise with a sling revealed no significant difference. Even though the principle of gravity compensation is at present rather unexploited, faster recovery can be expected along with new trials. Recently, three arm gravity compensators were put on the market: Armon[1], Dynamics Arm Support[2] (DAS) and Armeo[3]. The first two devices provide lower arm support, while the third is an exoskeleton type mechanism allowing free arm movement for people with little muscular arm strength, using passive principles, thus providing for static and not accounting for dynamic compensation. More research in this field can clarify the mechanisms of the compensation impact, as well as indicate where it would be reasonable to apply these mechanisms.

Besides the above approaches, other methods of rehabilitation can additionally promote fast recovery in hemiplegic patients. Interesting assertions were published by Kahn et al. [35], claiming that equally positive changes in motor function were observed in a group undergoing robot-assisted reaching exercise (*target is always reached)* and in a group undergoing robot-assisted reaching exercise *without obligatory reaching of the target.* Some other possible techniques deserve mention: functional electrical stimulation, pharmacology, intensive exercise, including many repetitive exercises, loading therapy - automated use of this method - and sensor-motor exercise.

[1] www.mginside.info
[2] www.exactdynamics.com
[3] www.hocoma.ch

A systematic review of the effect of robot-aided therapy on recovery of the hemiplegic arm after stroke, collecting results from eight selected studies is provided by Prange et al. [36]. This indicates that robot-aided therapy of the proximal upper limb can improve short- and long-term motor control of the paretic shoulder and elbow. This statement is supported by quantitative analysis of short-term pooled data in chronic stroke patients and indicates that increased motor recovery of chronic patients is possible after robot-aided therapy. However, no consistent effect on improvement in functional ability has been reported, although the training modalities were not directly designed for this. Restoration of motor control appears greater after robot-aided therapy than conventional tehrapy. It was not possible to establish which aspects of robot-aided therapy (e.g. increased intensity of movements, more effective training modalities) were responsible for the beneficial influence on recovery. Clinical experience seems to show that robot-aided therapy can improve motor control of hemiplegic upper limbs, perhaps even more than conventional therapy, in both sub-acute and chronic stroke patients.

5. Devices for Robot-Aided Therapy

A number of research or commercial platforms exist today designed specifically for the tasks of rehabilitation robotics, while in some other platforms the primary design goal has been aimed at something else, but the resulting device also suits the needs of rehabilitation. Regardless of their origin, the devices can be of either the exoskeleton or the end-effector type. An exoskeleton is an external framework, close to, or in contact with, human limbs.

Exoskeletons contain rigid and resistant components for each of their segments, being able to providing passive or active translation or rotation in a particular joint. The rotation axes of the robot must correspond to the rotation axes of the human skeleton, and the limb may be connected to the exoskeleton at several points. Powered exoskeletons may be designed, for example, to assist and protect soldiers and construction workers, or to aid the survival of people in other dangerous environments. Wider medical markets exist for providing mobility assistance and rehabilitation for aged and infirm people. A weak point in exoskeleton devices is the limited amount of Z-bandwidth and the transparency of the haptic interface – the parameters that tell how good is the touch (impedance) compared to the ideal model in the virtual environment. A strength is the good fit to each human body segment and the ability to move each limb and interact with supervised known parameters.

Perry et al. [39] is using an exoskeleton robot with seven degrees of freedom in the most important joints of human arm. L-Exos is a tendon-driven, wearable haptic interface with 5 DoF, optimized to obtain a solution with reduced mass and high stiffness, by employing special mechanical components and carbon fiber structural parts [41]. Neural control of an upper limb powered exoskeleton system has 8 DoF [44]. ARMin represents an interesting later design that at the moment enables movements in 6 DoF [26].

In most cases, robotic haptic arms that interact with the environment only via end-effectors are in contact with the human arm via the wrist or with the human leg via the foot only. As a consequence, the trajectory of each body segment under motion is not supervised and determined by a machine, as above, but also depends on the person. This can be good, but may also lead to a less determined environment. Good points in

the end-effector approach are better values for Z-bandwidth and transparency of the haptic device.

Examples of end-effector upper extremity devices include MIT Manus [18], Assisted Rehabilitation and Measurement (ARM) Guide [19], Mirror Image Motion Enabler (MIME) [20], Bi-Manu-Track [21], GENTLE/S [22], Neurorehabiltation (NeReBot) [23], REHAROB [24], Arm Coordinating Training 3-D (ACT^{3D}) [25], Braccio di Ferro, [42], NEDO project device [43] and Baltimore Therapeutic Equipment Co. (BTE), produce several training devices.

Most devices have been designed for training of the proximal joints (shoulders and elbows) of the hemiplegic arm. Some allow for two-dimensional plane movement only, while in most cases, movement is possible in three dimensions within the limited section of the entire working area of the arm. On the contrary, Bi-Manu-Track includes the forearm and wrist, which is enabled also by the recent version of the MIT-Manus device. Over and over again we encounter new robotic devices and evolutional upgrades of the existing ones (e.g. Furusho et al. [33] and Colombo et al. [34]).

There are also devices available for the lower extremities, Gait Trainer GT I [37] and HapticWalker [38] most frequently in combination with treadmills, and also separately for ankles. In the lower extremities, the end-effector devices only guide foot motion, while the multiple degrees of freedom of the rest of the body (e.g. leg, hip) remain completely unrestricted. In the case of a patient with an unstable knee joint, the physiotherapist has to stabilize the knee manually. Studies on the Gait trainer GT I revealed comparable muscle activation patterns as in treadmill walking.

Research devices are also available for exercising some other joints of the body, like ankle, wrist, individual fingers [45] or several fingers simultaneously.

Figure 6. ARMin robot is one of the recent exoskeleton devices (ETH Zürich)

Figure 7. GENTLE/S project rehabilitation environment as an example of end effector approach

6. Measurements in Rehabilitation

In neurology, standard rating scales are used for measurements, mainly for the purpose of early systemisation. On the other hand, uniform, quantitative, traceable procedures for measurements in rehabilitation have not yet been introduced. One of the models supposedly establishing a logical, understandable and thorough system in rehabilitation was adopted by the World Health Organisation in 1980. Even though the model, known as ICIDH (*International Classification of Impairments, Disabilities and Handicaps*) is very detailed and subjective (rheumatologic diseases), it provides a solid framework for understanding and treating neurological diseases and injuries. The most important concept of the model is the assumption that every disease can be evaluated on four levels, namely:

- *pathology level,*
- *impairment level*,
- *disability level*,
- *handicap level*.

In practice, the border between individual levels can be quite blurred. In general, a distinction has to be made between measurements and scoring scales. Measurement involves the use of a specific, traceable standard and the comparison of an actual value of a certain quantity against this standard. Scales often rely on subjective observations or rough measurements leading to a very few quantum descriptors. In rehabilitation, measurements are much rarer than scoring scales. The simplest scoring of motor

abilities and impairment of e.g. the upper extremity includes the squeezing of the dynamometer for estimating muscle strength and measuring the range of movement of individual joints. Self-scoring by patients is also present. The criteria of a good scoring scale include:

- validity (the result actually refers to the aspects for which the test was carried out),
- reliability (competent observer and time stability of results of such observer),
- sensitivity (detection, differentiation of sufficiently small changes that are still relevant).

Standardised rehabilitation tests are for the most part tailored to a specific disease and consider also the condition of other parts of the body and activity in general. Some of these tests are *Fugl-Meyr*, *Barthel Index* and the widely-used *Barthel Index*, *Katz index*, *Nottingham Adl Index*, *Jebsen*, *FIM-test*, *Box and Block Test*, *Nine-Hole Peg Test*, *Action Research Arm Test*, *Franchay Arm Test* and others.

Unlike the scoring approach, haptic, robotic technology provides for the next stage in rehabilitation, objective measurement during exercise. Quantities that vary with time (position and orientation, velocities, accelerations, forces, torques) can often be measured directly or derived. The presentation of these quantities for individual, relevant points of the body offers an objective basis for evaluation, as in kinesiology, and derived indexes can answer significant questions. Observation of measured data in a frequency domain can result in new insights, for instance, in the case of Parkinson's disease, the amplitude and frequency of shaking are determined by both place and time.

The measured quantities can be used as entries into various physical or physiological models and through them new important parameters may be acquired, e.g. active torques of the muscles of individual joints, passive torques in joints or even mass and evaluation of moment of inertia of some body segments. In doing so, we have to be aware of the fact that the more complex models usually comprise a higher number of parameters, some roughly evaluated, carrying high measurement uncertainly, which in the end may contribute to an even more uncertain final result.

7. Conclusion

Recent technological developments have made many things possible for the first time. However, actual systems are only now being introduced into our lives in the fields of computing, virtual reality visualisation, medicine in general and a narrow field of rehabilitation. Previously most simple devices, accessories and physiotherapists' hands will in the future be complemented by computerized devices. These will help in existing and, hopefully, also some new aspects of rehabilitation. Robotic approaches undoubtedly allow for equally fast, perhaps even faster, but certainly more interesting and entertaining rehabilitation methods. It has to be expected that at least some good laboratory prototypes will develop into widely-available products. These products are expected to have a suitable modular design, and developments in other areas, e.g. improvement in speed of computer processing, improvement in video techniques, software support, etc., will further contribute to the advancement in this area. Unfortunately, the size of the rehabilitation robotic market will be considerably smaller than the size of the market for computer games, which is why the volume of investment will be smaller and the pace of development is expected to be slower.

Acknowledgements

The described research work is financed by the Slovenian Research Agency in the form of a programme group Analysis and Synthesis of Movement in Man and Machine, and by providing grants to young researchers.

References

[1] D. Khalili, M. Zomlefer, *An intelligent robotic system for rehabilitation of joints and estimation of body segment parameters*, IEEE Trans Biomed Eng **35** (1988), 138-46
[2] B. Gjelsvik, E. Bassoe, *Form und Function*. Stuttgart, New York: Thieme Verlag, 2002.
[3] J. Howle, *Neuro-Developmental Treatment Approach: Theoretical Foundations and Principles of Practice*, Laguna Beach, CA, NDTA, 2002
[4] B. Bobath, *Observations on adult hemiplegia and suggestions for treatment*, Physiotherapy 45 (1959), 279-289
[5] E. Panturin, *The Bobath concept*, Clin. Rehabil. **15** (2001), 111-113
[6] S. Raine, *Defining the Bobath concept using the Delphi technique*, Physiother. Res. Int. **11** (2006), 4-13
[7] F.A. Mussa-Ivaldi, N. Hogan, E. Bizzi, *Neural, mechanical, and geometric factors subserving arm posture in humans*, J. Neurosci.. **5** (1985), 2732-2743
[8] R. Shadmehr, F.A. Mussa-Ivaldi, *Adaptive representation of dynamics during learning of a motor task*, J. Neurosci. **14** (1994), 3208-3224
[9] T. Tsuji, P.G. Morasso, K. Goto, K. Ito, *Human hand impedance characteristics during maintained posture*, Biol. Cybern..**72** (1995), 475-485
[10] P. Lum, D. Reinkensmeyer, R. Mahoney, W.Z. Rymer, C. Burgar, *Robotic devices for movement therapy after stroke: current status and challenges to clinical acceptance*, Top Stroke Rehabil **8** (2002), 40-53
[11] G. Fazekas, M. Horvath, A. Toth, *A novel robot training system designed to supplement upper limb physiotherapy of patients with spastic hemiparesis*, Int. J. Rehabil. Res. **29** (2006), 251-254
[12] S. Hesse, G. Schulte-Tigges, M. Konrad, A. Bardeleben, C. Werner, *Robot-assisted arm trainer for the passive and active practice of bilateral forearm and wrist movements in hemiparetic subjects*, Arch. Phys. Med. Rehabil. **84** (2003) 915-920
[13] C.D. Takahashi, D.J. Reinkensmeyer, *Hemiparetic stroke impairs anticipatory control of arm movement*, Exp. Brain Res. **149** (2003), 131-140
[14] J.L. Patton, F.A. Mussa-Ivaldi, *Robot-assisted adaptive training: custom force fields for teaching movement patterns*, IEEE Trans. Biomed. Eng. **51** (2004), 636-646
[15] J.L. Patton, M.E. Stoykov, M. Kovic, F.A. Mussa-Ivaldi, *Evaluation of robotic training forces that either enhance or reduce error in chronic hemiparetic stroke survivors*, Exp. Brain Res. **168** (2006), 368-383
[16] R.B. Salter et al., *The effects of continuous compression on living articular cartilage*, J. Bone Joint Surg. **42-A** (1960)
[17] M.A. Ritter, V.S. Gandolf, K.S. Holston, *Continuous passive motion versus physical therapy in total knee arthroplasty,* Clin. Orthop. Relat. Res. **244** (1989), 239-243
[18] H.I. Krebs, N. Hogan, B.T. Volpe, M.L. Aisen, L. Edelstein, C. Diels, *Overview of clinical trials with MIT-MANUS: A robot-aided neuro-rehabilitation facility*, Technol. Health Care **7** (1999), 419–423
[19] D.J. Reinkensmeyer, L.E. Kahn, M. Averbuch, A. McKenna-Cole, B.D. Schmit, W.Z. Rymer, *Understanding and treating arm movement impairment after chronic brain injury: progress with the ARM guide*, J. Rehabil. Res. Dev., **37** (2000), 653–662
[20] C.G. Burgar, P.S. Lum, P.C. Shor, H.F. Machiel Van der Loos, *Development of robots for rehabilitation therapy: the Palo Alto VA/Stanford experience*, J. Rehabil. Res. Dev., **37** (2000), 663–673
[21] S. Hesse, G. Schulte-Tigges, M. Konrad, A. Bardeleben, C. Werner, *Robot-assisted arm trainer for the passive and active practice of bilateral forearm and wrist movements in hemiparetic subjects*, Arch. Phys. Med. Rehabil. **84** (2003), 915–920
[22] S. Coote, E. Stokes, B. Murphy, W. Harwin, *The effect of GENTLE/s robot-mediated therapy on upper extremity dysfunction post stroke*, Proc. 8th International Conference on Rehabilitation Robotics; 2003, Daejeon, Korea, Human-Friendly Welfare Robot System Engineering Research Center, 59–61, 2003
[23] C. Fanin, P. Gallina, A. Rossi, U. Zanatta, S. Masiero, *NeRe-Bot: A wire-based robot for neurorehabilitation*, Proc. 8th International Conference on Rehabilitation Robotics, Daejeon, Korea, Human-Friendly Welfare Robot System Engineering Research Center, 23–26, 2003.

[24] A. Toth, G. Fazekas, G. Arz, M. Jurak, M. Horvath, *Passive robotic movement therapy of the spastic hemiparetic arm with REHAROB: Report of the first clinical test and the follow-up system improvement*, Proc. 9th International Conference on Rehabilitation Robotics, Chicago, Illinois. Madison (WI), Omnipress, 127–130, 2005.
[25] J. Dewald, M.D. Ellis, B.G. Holubar, T. Sukal T, A.M. Acosta, *The robot application in the rehabilitation of stroke patients*, Neurol. Rehabil. **4** (2004), 7
[26] T. Nef, R. Riener, *ARMin—Design of a novel arm rehabilitation robot*, Proc. 9th International Conference on Rehabilitation Robotics, Chicago, Illinois. Madison (WI), Omnipress, 57–60, 2005
[27] S.E. Fasoli, H.I. Krebs, J. Stein, W.R. Frontera, N. Hogan, *Effects of robotic therapy on motor impairment and recovery in chronic stroke*, Arch. Phys. Med. Rehabil. **84** (2003), 477–482
[28] J. Stein, H.I. Krebs, W.R. Frontera, S.E. Fasoli, R. Hughes, N. Hogan, *Comparison of two techniques of robot-aided upper limb exercise training after stroke*, Am. J. Phys. Med. Rehabil. **83** (2003), 720–728
[29] R.J. Sanchez, E. Wolbrecht, R. Smith, J. Liu, S. Rao, S. Cramer, T. Rahman, J.E. Bobrow, D.J. Reinkensmeyer, *A pneumatic robot for re-training arm movement after stroke: rationale and mechanical design*, Proc. 9th Int. Conf. Rehabilitation Robotics, Chicago, Illinois, Madison (WI), Omnipress, 500–504, 2005
[30] R.F. Beer, J.D. Given, J.P. Dewald, *Task-dependent weakness at the elbow in patients with hemiparesis*, Arch. Phys. Med. Rehabil. **80** (1999), 766–772
[31] R.F. Beer, J.P. Dewald, W.Z. Ryme, *Deficits in the coordination of multijoint arm movements in patients with hemiparesis: evidence for disturbed control of limb dynamics*, Exp. Brain Res. **131** (2000), 305–319
[32] R.F. Beer, J.P. Dewald, M.L. Dawson, W.Z. Rymer, *Targetdependent differences between free and constrained arm movements in chronic hemiparesis*, Exp. Brain Res. **156** (2004), 458–470
[33] J. Furusho, K. Koyanagi, Y. Imada, Y. Fujii, K. Nakanishi, K. Domen, K. Miyakoshi, U. Ryu, S. Takenaka, A. Inoue, *A 3-D rehabilitation system for upper limbs developed in a 5-year NEDO project and its clinical testing*, Proc. 9th Int. Conf. Rehabilitation Robotics, Chicago, Illinois. Madison (WI), Omnipress, 53–56, 2005
[34] R. Colombo, F. Pisano, S. Micera, A. Mazzone, C. Delconte, M.C. Carrozza, P. Dario, G. Minuco, *Upper limb rehabilitation and evaluation of stroke patients using robot-aided techniques*. Proc. 9th Int. Conf. Rehabilitation Robotics, Chicago, Illinois, Madison (WI), Omnipress, 515–518, 2005
[35] L.E. Kahn, M. Averbuch, W.Z. Rymer, D.J. Reinkensmeyer, Comparison of robot-assisted reaching to free reaching in promoting recovery from chronic stroke, In: M. Mokhtari (ed.), *Integration of assistive technology in the information age*, Amsterdam (the Netherlands), IOS Press, 39–44, 2001
[36] G.B. Prange, M.J.A. Jannink, C.G.M. Groothuis-Oudshoorn, H.J. Hermens, M.J. Ijzerman, *Systematic review of the effect of robot-aided therapy on recovery of the hemiplegic arm after stroke*, J. Rehabilitation Research & Development, **43** (2006), 171–184
[37] S. Hesse, D. Uhlenbrock, *A mechanized gait trainer for restoration of gait*, J. Rehabilitation Research & Development, **37** (2000), 701–708
[38] H. Schmidt, D. Sorowka, F. Piorko, N. Marhoul, R. Bernhardt, *Entwicklung eines robotergestützten Laufsimulators zur Gangrehabilitation (Development of a Robotic Walking Simulator for gait rehabilitation)*, Biomed. Eng, **40** (2003), 281-286
[39] J.C. Perry, J. Rosen, *Design of a 7-Degree-of-freedom upper limb powered exoskeleton*, BioRob 2006, The IEEE/RAS-EMBS Int. Conf. Biomedical Robotics and Biomechatronics, Pisa, Toscany, Italy, 805-810, 2006
[40] H. Schmidt, D. Sorowka, F. Piorko, N. Marhoul, R. Bernhardt, *Entwicklung eines robotergestützten Laufsimulators zur Gangrehabilitation (Development of a robotic walking simulator for gait rehabilitation)*, Biomedical Engineering **40** (2003), 281-286
[41] A. Frisoli, F. Rocchi, S. Marcheschi, A. Dettori, F. Salsedo, M. Bergamasco, *A new force-feedback arm exoskeleton for haptic interaction in Virtual Environments*, First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems WHC, 2005.
[42] M. Casadio, V. Sanguineti, P.G. Morasso, V. Arrichiello, *Braccio di Ferro: A new haptic workstation for neuromotor rehabilitation*, Technol. Health Care **14** (2006), 123–142
[43] K. Koyanagi, J. Furusho, U. Ryu, A. Inoue, *Development of rehabilitation system for the upper limbs in a NEDO project*, Proc. 2003 IEEE Int. Conf. Robotics & Automation, Taipei, Taiwan (2003), 14-19
[44] J.C. Perry, J. Rosen, S. Burns, *Upper-limb powered exoskeleton design*, IEEE Transactions on Mechatronics **12** (2007), 408-417
[45] U. Mali, M. Munih, *HIFE-haptic interface for finger exercise*, IEEE/ASME Transactions on Mechatronics **11** (2006), 93-102

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

Subject Index

This page intentionally left blank

Basic Engineering for Medics and Biologists
T.C. Lee and P.F. Niederer (Eds.)
IOS Press, 2010

Author Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank