TITLE: Modelling in Healthcare

AUTHOR: Complex Systems Modelling Group (CSMG)

AUTHOR AFFILIATION: The IRMACS Center, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

AMS Codes 00A06, 00A71, 97Mxx

Preface

As healthcare systems worldwide face the challenge of delivering quality services while maintaining control over escalating costs, there is growing support for the view that conventional approaches to the organization of healthcare systems are failing. The questions arising are extremely complex, and in most cases it is not acceptable to rely on simple intuition to answer a given question. In order to develop solid, defendable, evidence-based answers to the complex questions arising in modern healthcare, modelling is being increasingly applied. Yet, to many healthcare policy makers, the development, tuning, testing, validation, and eventual application of a model is considered a foreign art.

A model is a simplified representation of a real world situation used to help answer a specific question. The main role of a model is to help steer decisions in the right direction. In most cases a model cannot give the "right" answer to a problem, but it can be a useful tool in characterizing the problem and finding ways to resolve it. In this book we provide a first look into the world of modelling, with particular focus on modelling in healthcare. We provide this as it is our belief that decision-makers and modellers must share a certain level of knowledge to maintain a healthy relationship. Modellers working on healthcare problems cannot be blind to the less mathematical issues of healthcare, and decision-makers engaging modellers should not view models as mystical crystal balls from which answers emerge. As the Complex Systems Modelling Group, we have worked extensively with healthcare policy-makers, and developed expertise in developing and analyzing of healthcare models, and in explaining models and modelling to those unfamiliar with the subject. With this book we hope to use these expertise to help strengthen the bonds between the worlds of modelling and healthcare.

This book is by no means a complete text on the subject of modelling in healthcare. In fact each chapter in this book (with the exception of the introductory chapters) could easily be extended into a complete textbook. We hope readers will view this as a handbook of modelling in healthcare, and use it to provide themselves with a broad overview of how modelling works and what it is capable of.

As a handbook of modelling techniques, each chapter of this book has been written in a self-contained manner. Any given chapter can be read without having read previous chapters. (Although we strongly recommend reading the introductory chapters before tackling other chapters.) With the exception of the introductory chapters, each chapter focuses on a particular style of modelling that is applicable to healthcare. To keep the book as self contained as possible, most chapters contain enough background that they are accessible to anyone with a solid high school level of mathematics. To ease reading, each chapter is written using the same basic template consisting of: *Model Overview, Common Uses, Model or Mathematical Details, Examples,* and *Related Reading.* Readers can quickly scan the model overview and common uses sections to determine if a model is applicable to the problem they are interested in, and study the mathematical details and examples sections if more detail is desired. The related reading section points readers to further literature of interest.

It is our hope that this book will provide a stepping stone for people interested in the world of modelling in healthcare, while remaining an excellent reference guide for those more familiar with the subject. As such, this book should be of use to

anyone, academic or professional, who is interested in broadening their knowledge regarding modelling in healthcare.

Acknowledgments

The creation of this book began when the British Columbia Ministry of Health Services commissioned the Complex Systems Modelling Group (CSMG) at Simon Fraser University to produce a report on the many different mathematical options for modelling health care demand. For some of us this was a first step into the world of mathematical modelling specifically for healthcare, and a step that has become a defining moment in many of our lives. We would like to thank the British Columbia Ministry of Health Services for funding the original report, and their continued support during the completion of this book.

Like any large team project, the production of the final manuscript for this book was a long and complicated process. Without the assistance of the many contributing authors (listed on page iv) the final product would have been less than it is. Nonetheless, as with any large team project, there are certain individuals whose roles were more pivotal in the completion of this work. In this respect we would like to acknowledge authors Hare, Rutherford, and Vásárhelyi, as principle authors responsible for the original report and this book.

Finally, we would like to express our gratitude to the Center for Interdisciplinary Research in the Mathematical and Computational Sciences (IRMACS, http://www.irmacs.sfu.ca) at Simon Fraser University. We are deeply indebted to IRMACS for their continual support during the production of this work. Many barriers could never have been crossed without the aide of the strong team of IRMACS administrative and technical support staff.

Peter Borwein, Executive Director, CSMG
Oct. 2009.

Complex Systems Modelling Group

Authors

Azadeh Alimadad, M.Sc.
Biostatistics

Alex Borwein
Health Sciences

Peter Borwein, Ph.D.
Mathematics

Vahid Dabbaghian, Ph.D.
Mathematics

Chiaka Drakes, M.Sc.
Mathematics, Education

Ron Ferguson, Ph.D.
Mathematics

Amir H. Ghaseminejad, M.Sc.
Engineering, Social Modelling

Yuri Gusev, Ph.D.
Theoretical Physics

Warren Hare*, Ph.D.
Mathematics, Operations Research

Jenny Li, M.Sc.
Mathematics

Snezana Mitrovic-Minic, Ph.D.
Operations Research

Alexander Rutherford*, Ph.D.
Queueing Theory, Physics

Alexa van der Waall, Ph.D.
Mathematics

Krisztina Vásárhelyi*, Ph.D.
Epidemiology, Genetics

Les Vertesi, M.D.
Epidemiology, Health Policy

* Principle Authors

Contents

List of Figures

List of Tables

Part 1

Modelling in Healthcare

CHAPTER 1

The Whys, Whats, and Whens of Modelling in Healthcare

All this will not be finished in the first one hundred days. Nor will it be finished in the first thousand days, nor in the life of this administration, nor even perhaps in our lifetime on this planet. But let us begin. John F. Kennedy (1917-1963)

1. Why Model in Healthcare

Formally healthcare is defined as any effort made to maintain or restore health. Taken to the extreme, this definition encompasses almost everything we do. Breathing can be viewed as an effort to provide sufficient oxygen to the lungs, maintaining health. Eating supplies the body with nutrition to support and repair itself, restoring health. In a more practical sense, healthcare refers to efforts made by trained healthcare practitioners to maintain or restore health, and to efforts made by individuals to make contact with healthcare practitioners.

Healthcare is one of the oldest and largest professions in the world. Paintings discovered in the Lascaux caves in France, radiocarbon dated at over 15,000 years old, are interpreted to show the use of plants as healing agents. The Edwin Smith papyrus, dated between 3000 and 4000 years old, describes the examination, diagnosis, and treatment of numerous trauma injuries. Traditional Chinese medicine has origins dating back to the 5th century BC, and it is still in use today.

The size of the healthcare profession is also easily demonstrated. The healthcare expenditure per capita in the United States is over $5,000 per year. Healthcare expenditures total to over 15% of the United States' Gross Domestic Product (GDP). Although the United States are in fact the extreme end of the scale, Australia, Canada, France, and the United Kingdom, have all reached the 10% mark for healthcare expenditures as a portion of the GDP [1]. As life expectancies increase and population demographics shift, these numbers are expected to increase.

Aside from showing the size of the healthcare industry, the above numbers have sparked great debate and concern over the sustainability of the healthcare system. In 1970, the United States' total health expenditures only measured 7% of the GDP. Australia's, Canada's, France's, and the United Kingdom's healthcare expenditures measured approximately 5%, 7%, 5%, and 5% (respectively). Whether this growth is due to changing age demographics, increased cost of medical supplies, or a surplus of disposable income, it is clear that the healthcare systems of most modern countries are undergoing a time of change.

[1] All numbers based on 2003/04 fiscal year.

To cope with the rapid changes in the field of healthcare, governments and policy-makers worldwide must seek methods to better understand healthcare systems and the individuals who access them. The questions arising in modern healthcare are extremely complex, and it is no longer acceptable to rely on simple intuition to answer a given question. In order to develop solid, defendable, evidence-based answers to these complex questions, nathematical modelling is becoming increasingly important. In order to understand and interpret results model results, it is important for policy-makers to have a solid grasp of the fundamentals of modelling in healthcare. In this book we hope to provide many of these fundamentals, to allow policy-makers and modellers alike to quickly step into the exciting world of mathematical modelling in healthcare.

2. What is a Model

In this book the word "model" means a *simplified representation* of a real world situation used to help answer a *specific question*. As the focus of this book is modelling in healthcare, the situations and questions we discuss will tend to be those that arise in the healthcare industry.

There are two important aspects to the definition of a model. First, a model is a *simplified representation* of the real situation. Consider the scientific endeavor of modelling a collection of building designs in order to determine which design stands up best in the event of a fire. One option would be to build the entire collection of buildings and then burn them all down. Although this "model" would answer the specific question, it would not save any resources in the process. Instead it would be more reasonable to build small scale replicas of the buildings and burn them down in more controlled environments. This would answer the question faster and more accurately, as many more tests cases could be examined.

The second important aspect of a model is its capacity to answer a *specific question*. Models tend to answer the questions they are designed to answer, and as such, designing a model with no particular question in mind provides no insight into the situation of interest. This may be a useful exercise for a young academic student, but for a healthcare policy-maker the result is generally just a waste of resources.

When simplifying the real situation for the purposes of modelling it is important to preserve the properties of the system that are relevant to the question. For example, a model of an airplane may take on many forms depending on the purpose it is designed to serve. To study the aerodynamic properties of airplanes, a physical model preserving the shape of the airplane is built. To allow passengers to select seats on a commercial flight, a graphical seating plan may be produced by the airline. The latter model retains entirely different characteristics of the airplane than does the former.

This raises an interesting note on the distinction between *detail* and *complexity* in modelling. The goal of modelling is to clarify concepts, but models attempting to reproduce a real situation by introducing a large number of variables tend to accomplish the opposite. Models aim to expose pertinent relationships between variables, but unnecessary information can conceal these. As such, a good model has as low a complexity as possible while retaining the details necessary to approach the specific question the model is designed to examine. In general, models with a focused question and a limited number of conditions are more likely to be useful.

There are many different models that are applicable to solving questions in the field of healthcare, and there is no such thing as a unique "best" model for a given problem. In fact, in most cases, more than one model discussed in this book are applicable in solving a single question. In these cases different modelling methods are often complementary, with the best results obtained through an approach that integrates multiple methods. In general, modelling is most convincing when various different kinds of models lead to the same conclusion .

3. When to Use Modelling in Healthcare

Modelling can be a valuable tool to aid healthcare management, as long as it is used appropriately and with awareness of its limitations. It is most useful to think of modelling in healthcare not as a specific method, but rather as a process where modellers combine techniques and skills in mathematics and computation with the specialised knowledge of healthcare experts to arrive together at appropriate approaches to problems in healthcare. However, with this said, it is prudent to temper expectations on what modelling in healthcare can deliver. The main role of a model is to steer decision-makers in the right direction. In most cases a model cannot give the "right" answer to a question, but it can be a useful tool in characterizing the problem and finding ways to resolve it. Furthermore, modellers (and decision-makers who examine modellers' results) must always remain aware of the various biases influencing personal opinions and experiences. Models should not be blindly used, but validated both mathematically and by the solicitation of experts. A model that is contradictory to the real situation should be held in doubt and its conclusions should be examined carefully.

Despite these limitations, modelling techniques stand to make a significant impact in the field of healthcare. In this book we examine a number of current modelling techniques and how they have been applied to healthcare. This book is not a complete text on the subject of modelling in healthcare. Each chapter herein, with the exception of the introductory chapters, could be extended into a complete textbook in its own right. However, one might consider this a handbook of modelling in healthcare. It provides an introduction to many modelling methodologies, and references to further reading on each. It touches on many of the problems that are currently of interest in healthcare, and provides examples of when modelling has been used to approach these problem. For example, the book examines problems such as,

- Predicting and adjusting the future demand for healthcare (see Examples 7, 4.3 and 4.3);
- Examining demographic factors which relate to health (see Examples 4.2, 4.2, and 4.3);
- Understanding and adjusting patient health behaviour (see Examples 4.1, 4.1, and 4.3);
- Decreasing wait times and understanding bottle-necks for healthcare access (see Examples 4.1, 4.3, and 4.2);
- Understanding and controlling the spread of communicable diseases (see Examples 4.2, 4.2, and 4.2);
- Optimizing healthcare delivery (see Examples 4.1, 4.3, and 4.2).

There are many more problems which could fall under the heading of modelling in healthcare than those listed above. Some of these are covered in this book,

others are not. Many of those which are not covered are problems which are highly specialized in nature. (For example, the mathematical modelling employed for the delivery of radiation therapy or analysis of MRI data.) Such problems generally require a level of mathematics well beyond the scope of this book.

4. Related Reading

Detailed information on changes in national healthcare expenditures can be found in reference [**200**].

CHAPTER 2

How to Use this Book

I can't work without a model. I won't say I turn my back on nature ruthlessly in order to turn a study into a picture, arranging the colours, enlarging and simplifying; but in the matter of form I am too afraid of departing from the possible and the true. Vincent van Gogh (1853-1890)

Every man is wise when attacked by a mad dog; fewer when pursued by a mad woman; only the wisest survive when attacked by a mad notion. Robertson Davies (1913–)

This book can be viewed and used in a number of different ways. Primarily, it is a handbook of modelling techniques with an emphasis on how to apply them to current issues in healthcare. However it may also be used as an introductory undergraduate text on the subject. An excellent undergraduate project would be to select a chapter from this book, read it and its corresponding references, and then perform a literature search for additional examples of the model's application in healthcare.

As a handbook of modelling techniques, chapters pertaining to modelling techniques are written to be entirely self-contained, and focus on a specific type of model. The chapters, not focusing on a specific style of model include:

- Chapter 1, which introduces modelling in healthcare as a whole,
- Chapter 2, which describes how to use this book,
- Chapter 3, which discusses the modelling process,
- Chapter 4, which discusses the issues around collecting data for analysis, and
- Chapter 9, which discusses some general issues about constructing and analyzing models.

Throughout the book one will occasionally see margin notes, such as this one. The purpose of these is to highlight information that may be of interest to the reader.

The remaining chapters discuss particular models that can be applied in the field of healthcare. Each of these chapters is given an artistic title that provides some insight as to where the models discussed might be used, and a scientific title, which provides the standard name for the model examined in the chapter. A table of the models considered in this book can be found in Chapter 3 (Table 1, page 11).

In order to ease reading, the layout for chapters on specific models is uniform. Each chapter is divided into five sections, entitled: *Model Overview, Common Uses, Model Details* or *Mathematical Details, Examples* and *Related Reading.* In the *Model Overview* section we give a brief description of the model. These overview sections avoid mathematical language and should be readable by anyone with a solid high school background in science. The next section, *Common Uses*, provides a list of

example questions that the modelling technique could be used to address. These lists are not complete, but intend to provide an idea of what kind of problems the type of model is capable of answering. In the *Model Details* or *Mathematical Details* section we give a more detailed mathematical description and analysis of the model. Wherever possible, we provide all the necessary scientific background to read these chapters, however in some cases the models are complicated enough that this is impossible. Sections that may require a substantial undergraduate level of knowledge are:

- Section 3, Game Theory, which requires differentiation,
- Section 3, Network Theory, which discusses Graph Theory,
- Section 3, Markov Models, which requires matrix manipulation,
- Section 3, System Dynamics, which requires the use of differential equations, and
- Section 3, which requires differentiation and matrix manipulation.

In each of the *Examples* sections we provide two or three examples of how the modelling technique is applied in practice. Often the first of these examples is an artificially created example designed to demonstrate the model without burdening ourselves with the complications that arise in real examples. The remaining examples are taken from actual applications of the modelling technique in healthcare. Some of these examples demonstrate successful uses of the modelling technique in healthcare, others demonstrate how the model can fail if it is used inappropriately. The final section of each modelling chapter, *Related Reading*, provides details of the references used in the chapter, as well as several references that provide more detailed reading on the model discussed.

Regardless of the model discussed, the "Model Overview" section can be understood by a reader with a basic high school science background .

This book also contains an appendix that may be of use to the reader. The appendix lists and reviews of some of the modelling software that we have come across during our research. This can be quite technical at times and is intended for those who are interested in using software for producing models.

1. The Language of Modellers

There is a specialized technical language assocated with mathematical modelling. Most words are defined upon their first use within a given chapter. For now we would like to highlight several words that are frequently used throughout this book and comment on their meaning in healthcare modelling.

Model: a *simplified* representation of a real world situation used to help answer a *specific question*.

Quantitative Models: Models that use the language and tools of mathematics to describe the behaviour of a system. Such models make numerical predictions about how the real system is likely to behave.

Qualitative Models: Models designed to provide insight about why a given situation exists and what its driving factors might be. Such models do **not** provide numerical results pertaining to a given situation.

Disease: any *negative* health effect (for example, viral and bacterial infections, genetic disorders, increased chances of accidents causing harm, etc.).

Risk (Factor): any action or situation, be it beneficial or detrimental, that affects the probability of experiencing disease.

CHAPTER 3

The Modelling Process

You must see your goals clearly and specifically before you can set out for them. Les Brown (1945-)

Do not quench your inspiration and your imagination; do not become the slave of your model. Vincent van Gogh (1853 - 1890)

From drawing up the optimal staff schedule for a hospital emergency room to exploring how the global airline industry impacts the spread of disease, models are finding applications in almost every area of the healthcare industry. Yet, to many, the development, tuning, testing, validation, and application of modelling is a mysterious and an overwhelming task. In this chapter we provide a very broad outline of the modelling process, with specific emphasis on modelling in healthcare.

It is impossible to define a concise step-by-step process for selecting, designing, tuning, and applying an appropriate model to answer a given question. However, it is possible to outline some guiding principles that can help modelling projects achieve good results, and to list some general steps that one should expect during the modelling process. We begin with a quick overview of some guiding principles in modelling.

Guiding Principles in Modelling.

The question should be clearly defined: Models intended as "multi-purpose" tools that start without a clearly defined question generally end up without any clear conclusions. Conversely, models designed with a clear purpose in mind, once validated, can often be easily adapted to other purposes.

Models should be simple and transparent: In building models, one of the most difficult tasks is selecting the relevant details. Once the relevant details are uncovered, the model design should be as simple as possible while incorporating these details.

On a related note, there are many software applications that may aid in modelling. Although software can reduce the time involved in repetitive tasks, the modeller must still have a thorough understanding of what the software (and subsequently the model) actually does. Otherwise, it is easy for errors to arise in the model.

All assumptions should be clearly stated: All models are built on a set of assumptions, some of which are testable, and others that are not. These assumptions must be clearly stated, and whenever possible tested. Assumptions that are not testable should be discussed with experts from within the field.

Variables and measures should be clearly defined: A quantative model is useless if the numeric result is uninterpretable. As such the numerical variables and output measures should be clearly stated for each specific model.

Use the best data available: Clearly, the quality of data imposes a limiting factor on the quality of mathematical models and their results. Although the model may be designed and tested with most data, final implementation and results should always use the best quality data available.

Interpret results carefully: After a model is created and final results are obtained, it is a common mistake to over-interpret the importance of the results. One of the most common errors is to assume causality where only association is present. Most statistical models are only capable of showing correlations between two events, not explaining the causality. (This is discussed further in Chapter 5; other common errors are discussed as they arise within this book.)

In Figure 1 we provide our view of the modelling process. Notice that it is a long process with many "feedback" loops. This suggests that many initial approaches to a problem will be unsuccessful. With practice and experience the number of unsuccessful approaches will decrease, but one should never expect the first attempt at modelling a system to work flawlessly.

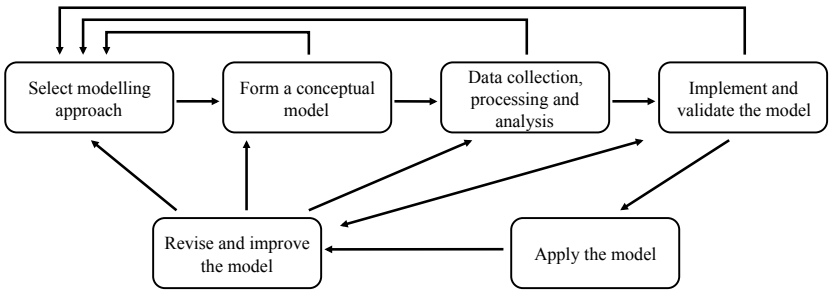

FIGURE 1. **The modelling process:** One View of the Modelling Process

In the remainder of this chapter we elaborate on each step of the modelling process. The chapter ends with some references where one can learn more about the modelling process.

1. Selecting a Modelling Approach

Contrary to what is commonly taught in high-school, very few problems have a unique solution. Indeed, in our experience, most questions (be it healthcare related or not) can be solved in more than one manner and have more than one reasonable solution. Likewise, for most problems, more than one modelling approach is possible, and each will have advantages and disadvantages. Therefore one of the first concerns a modeller will have to deal with is selecting which modelling technique to apply. This selection process will generally be driven by many factors, including the type of data available, the nature of the situation to be modelled, and the type of

question posed. In general, the most convincing results are obtained when multiple modelling techniques are applied and their results support each other.

In Table 1 we list the models discussed in this book, and provide a brief explanation for the main usage of each with respect to healthcare. We further split the models into the two broad modelling categories of qualitative models and quantitative models. We describe these categories below.

Model	Qualitative or Quantitative	Main Usage	Chapter
Descriptive Statistics	Quantitative	Summarizing data sets.	5
Regression Analysis	Quantitative	Predicting future trends based on statistical data.	6
Epidemiological Risk Models	Quantitative	Developing relationship bewteen risk factors and diseases, and explore the impact of interventions on population health.	7
Psychosocial Risk Models	Qualitative	Exploring how the public can be swayed into better health behaviour.	8
The Health Belief Model	Qualitative	Providing a psychological framework to help understand patient behaviour.	10
The Behavioral Model for Healthcare	Qualitative	Providing a psychological framework to help understand patient behaviour.	10
Game Theory	Both	Understanding rational decision making (often to determine when healthcare clients are acting irrationally).	11
Human Capital Models	Qualitative	Examining health decisions from an economic perspective.	11
Network Models	Both	Describing social or physical interactions within society, and how they impact health.	12
Markov Models	Quantitative	Exploring the properties of objects in a system that move through a series of states (such as patients moving through disease states).	13
Systems Thinking	Qualitative	Developing flow-box diagrams that view a system as a whole (usually to better understand feedback loops and interactions between various parts).	14
System Dynamics	Quantitative	Quantifying systems thinking models.	14
Queueing Models	Quantitative	Understanding wait times and bottle necks for objects moving through a system (such as patients waiting for surgery).	15

TABLE 1. Categorization of the models covered in this book, along with locations.

1.1. Qualitative Models. Many models in healthcare are not designed to provide specific numerical results, but instead are designed to provide insight into

why a given situation exists and/or what its driving factors are. Such models are generally referred to as *qualitative models*.

Qualitative models come in many forms. Sometimes they rely on psychological analysis of a situation, other times these models focus on examining how various aspects of a company interact. However, all qualitative models share a common property; they do not attempt to produce a quantified output as a solution to a problem. Instead, they attempt to determine the factors that impact a given problem in order to provide guidance on how the situation might be adjusted.

Consider for example the advertising industry and its continual goal of convincing the public to spend their money. Over time some clear trends have developed. More toy commercials appear near the Christmas holidays, and more weight loss commercials shortly thereafter. The reasons for these trends may appear clear (people are interested in buying gifts before Christmas, and interested in fulfilling New Year's resolutions of weight loss after Christmas), but some very bright minds were involved in developing and answering the question of when people are most susceptible to a given form of advertisement.

Similar ideas can easily be applied to the healthcare question of how to increase attendance at blood banks, immunization clinics, and various other healthcare services that decrease the overall burden on the healthcare budget. By developing a qualitative model of the factors that affect an individual's interactions with the healthcare system, we can better understand why certain groups of people are less likely to maintain a regular schedule of mammography for example. By building qualitative understanding for situations, such as the above, we can develop interventions that are better designed for the given situation.

It should be noted that statistical data is usually not the starting point of qualitative models in healthcare. For this reason it is extremely important to validate qualitative models using scientific experiments. That is, before applying the results implied by a qualitative models, one should always use quantitative modelling to confirm its validity.

A list of qualitative models discussed in this book can be found in Table 1.

1.2. Quantitative (Mathematical) Models. Many of the models described in this book use the language and tools of mathematics to describe the behaviour of a system. In these cases the system is described by a set of variables and equations that establish relationships between these variables. We refer to such models as *quantitative* or *mathematical models*.

Mathematical models come in many different forms, all sharing the common feature of quantifying something. Typically, mathematical models take an input of data and produce an output of conclusions. Therefore, mathematical models can only be as good as the data used.

It is instructive to consider a simple example such as an emergency room queue. This demonstrates some interesting characteristics of models and modelling. On one level, modelling an emergency room queue could be very simple. Patients could be treated as a single queue of customers that are served by several physicians. The assumption of "first come first served" could be employed and assumed "fair." However, whether this is fair actually depends on what one seeks to accomplish. Are we simply interested in maximizing the number of patients served, or are we interested in efficient use of resources? Is a single queue equally fair to all patients, or should some prioritization be employed? If all physicians are occupied with complicated

cases that take a long time to resolve, the waiting time for those remaining in the queue should increase substantially. The complexity of this example increases manyfold as greater detail is brought into the model. The harmonic operation of multiple healthcare services, all relying on the same pool of resources, for example, can be even more complex. One task of mathematical models is to make complex situations more manageable.

If a quantitative model is chosen, the modeller must also make several further choices about the modelling technique to be used. For example, should the model be

Stochastic or **Deterministic**: *Stochastic* models are models that incorporate random events and behaviours. For example, prescriptions for a specific medication at a pharmacy are filled at random times, although the average number of prescriptions may be constant over time. Useful stochastic models allow for long term patterns and average properties to be determined. *Deterministic* models are models where events proceed in a fixed and predictable fashion. As a result, the same set of initial conditions will result in the same outcomes every time. Despite this, deterministic models can exhibit extremely complicated behaviour, and are often useful in studying how changes in one part of a system impact other parts of the system.

Static or **Dynamic**: A *static* model is a model that provides a snapshot of the system at a specific point in time. As such, static models do not allow for time to effect the variables of the system. Making predictions based on such models is usually done via basic extrapolation, and therefore limited in its accuracy. However, static models are often sufficient and generally easy to construct. Static model are also well suited for developing case strategies to deal with a given situation. In contrast, in *dynamic* models the states of variables change over time. Because of the time component, dynamic models can provide a representation of the evolution of the system, which generally allows for more accurate predictive properties. However, dynamic models are more difficult to design.

Discrete or **Continuous**: For each variable in the model, one must decide whether the variable is discrete or continuous. *Discrete* variables are variables that can only take on values from a list of possible values. The list may be finite (such as days of the week) or infinite (such as the list of integers). Alternately, *continuous* variables are chosen from the real number line, so any two values always have a third value in between them. Continuous variables may still have upper and lower bounds ($5 \leq x < 7$ for example), or may be unbounded ($x \geq 9$). In most cases what the variable represents will provide insight as to whether it is discrete or continuous. For example, the number of patients in a queue should be discrete, while the arrival rate of patients into the queue should be continuous.

If a dynamic model is used, whether time is modelled discretely or continuously, has a profound impact on the model, its implementation, and the type of mathematics required to analyze the model. It should also be noted that all computer simulation models proceed in discrete time due to the digital nature of computation. However, the time step may be specified to be so small that continuity is essentially preserved.

A list of quantitative models discussed in this book is displayed in Table 1.

2. Forming a Conceptual Model

Once a modelling approach is selected, the modeller proceeds by forming a conceptual model of the problem. This is a cognitive process of translating external events into internal models, similar to what humans automatically engage in more or less every day in order to make sense of the world.

When a conceptual model is formed, it becomes a theoretical construct that represents, often visually, the processes, relationships, and variables considered to be important within a system. This construct should be examined by experts and practitioners from within the system to determine a first level of validity. In particular, if the experts and practitioners from within the system do not trust the conceptual model it will remain unused, regardless of its quality.

The conceptual model both drives and is driven by the variables that are considered important in the system. Since the variables that are considered important may change as data analysis is performed, one may have to reform the conceptual model several times before the modelling process is complete. Moreover, in building the conceptual model, it may become clear that the chosen modelling approach is not appropriate. Thus, one may have to select a new modelling approach in order to develop the conceptual model into a usable model.

3. Data Collection, Processing, and Analysis

Throughout the modelling process, the modeller relies on data. For qualitative models, data is used to test and support the model, for quantitative models data is used to tune the model to allow for predictions. Overall, data provides descriptive information about the system, and suggests which variables should be considered important within the model. Examples of possible variables include the demographic structure of a population, the transmission rate for a communicable disease, or the rate at which surgical procedures are completed.

Data Collection. The classic *GIGO* axiom of modelling stands for "Garbage In, Garbage Out." GIGO captures the idea that a model is only as good as the data used to test and tune it. In some problems, the data requirements are easy to define, and the data is easy to collect. For example, determining the future distribution of population age groups can be easily accomplished by examining past age distributions and extrapolating. Of course, birth rates, death rates, immigration rates and emigration rates all have to be taken into account, but overall this data can be easily and accurately obtained.

Further discussion regarding the collection and cleaning of data can be found in Chapter 4. Discussion on Statistical Analysis can be found in Part 2 of this book.

In many problems in healthcare data collection is a limiting factor in model development and analysis. This is beginning to change as computerized patient tracking is developed and implemented, but even then confidentiality issues cause data collection challenges. Ignoring the issue of patient confidentiality, data collection in healthcare remains a resource-intensive undertaking that often requires conducting surveys or population studies. Such surveys can be extremely expensive and time consuming to complete, and even on completion the data may be corrupted by survey bias.

Data Processing (Cleaning). Ideally data collection is carried out with a specific modelling problem in mind. In this way the right kind of data can be collected to help solve the problem in question. In practice information on model variables is often extracted from data collected for other purposes. As a result, data may be biased and contain errors or inaccuracies. Another potential problem in healthcare modelling is that the question may be too difficult to define initially. In this case, the modelling process begins as an exploratory learning process, with a conceptual model of the problem as its result. It may not be clear at the outset what data is appropriate for describing the system. In these circumstances extensive cleaning of the data is often necessary to improve quality. *Data cleaning* can involve data entry, checking data for errors, identifying sources of bias, removing duplicate entries, and merging or linking databases.

Statistical Analysis. Once data of adequate quality is available, it is then possible to study the system through statistical analysis. *Statistical analysis* may include the use of descriptive statistics (see Chapter 5), regression analysis (see Chapter 6), risk analysis (see Chapters 7 and 8), or some combination thereof. (Many other forms of statistical analysis may also be employed, but these are not detailed in this book.)

The results of the data analysis are used to determine which variables are most important for the problem, to test a model's validity, and to tune a model for making predictions. Often statistical analysis will inform the modeller that some of their basic assumptions about the system were wrong, forcing the modeller to take a step backwards and form a new conceptual model for the problem. This may occur when a modeller determines that a variable assumed to be insignificant turns out to be significant or vice versa.

4. Implementing and Validating the Model

Once a model is specified, it has to be implemented in such a way as to produce predictions about the system under study. Model implementation may involve a computer or may proceed using more analytical approaches.

Computer simulation is a software-based method of implementation. By simulating a system, it is possible to examine how a model behaves without understanding the all of the analytic details of the system. In this regard simulation is often referred to as a *black box*; input and output are visible, but how the output is generated might not be fully understood. There are both advantages and disadvantages to simulation methods. On the one hand, even highly complicated problems can be captured in a simulated model, without detailed knowledge of the mechanics of the system and without the requirement for mathematical expertise on the part of the modeller. On the other hand, this lack of transparency can mask logical errors in the model, often producing false conclusions.

A second approach to implementing a model is provided by the tools of *mathematical analysis*. If the modeller is able to describe the system in terms of equations, then analytic or numerical solutions may be sufficient to "solve" the model without the need for simulation. This provides several strong advantages over simulation. For example, analytical methods produce exact reproducible solutions without the need for (often expensive) software. Furthermore, analytical methods often provide deep insights into the workings of a system. However, analytical solutions for complex models are often difficult or even impossible to achieve.

A third approach, which lies somewhat between simulation and analysis, is *numerical analysis*. In numerical analysis, the modeller uses mathematical techniques to develop equations to represent the model and then simplifies these equations if possible. The modeller then turns to computers to numerically approximate solutions to the equations. This approach retains some of the robustness of mathematical analysis and is especially useful if there is no known analytical method for solving the particular problem.

After a model is implemented it has to be tested for validity. Validation is carried out to substantiate that the model performs with satisfactory accuracy within the domain of its applicability. The simplest test of validity of a model is to compare the model output with actual data about the system. However, in doing this one must be warned that using the same data to tune and validate the model can easily lead to false positives. That is, the resulting validation may only be showing that the model works on this data set, because it was designed to work on this data set. Unless the model contains strong assumptions to simplify the data then this is not a particularly strong conclusion.

5. Applying the Model

Once a finalized model is tested, tuned, and implemented, it can be used to explore properties of the system *as described by the model*. It should be reinforced that no matter how well a model is tested, tuned, and implemented, it can only examine the aspects of the system it is designed to study. A seating chart of an airplane is the perfect model for allocating seats, but no matter how accurate the model is, it can never be used to test if the airplane will actually fly.

Exploring properties of the system can take many forms. For example, models do not usually display equal sensitivity to all input parameters. Determining which parameters have the greatest impact on the system can be useful in determining where to make interventions and where to focus further data collection efforts. Parameters that make little impact on the system do not have to be quantified as accurately as parameters which have major impacts on the system. Analysis that focuses on determining which parameters have the greatest impact on the system is generally called *sensitivity analysis*.

Another popular use of models is to determine some sort of optimal behaviour. For example, if a healthcare ministry has a budget for only a fixed number of physicians, they may wish to know where to locate those physicians in order achieve optimal patient care. In general, *optimization* is the application of this type of question to a model. Often the question can be written as "which selection of parameters minimizes the cost such that the desired result occurs?" Finding the answers to such questions has become a field in itself, and can be accomplished by a number of different means. Chapter 16 of this book is devoted to some optimization problems in healthcare.

6. Revising the Model

Here we come full circle and begin another modelling cycle. The modelling process is not merely a search for a solution, but also a learning process. New knowledge about the system is incorporated into new versions of the model.

7. Example: Modelling Healthcare Demand

Predicting the future demand for healthcare is of utmost importance to many healthcare policy-makers for the purpose of setting budgets and developing future coping strategies. Here we use the example of modelling healthcare demand to demonstrate that one problem can be approached by a plethora of different modelling techniques.

To do this we identify four groups of targeted models for healthcare demand, which we label: *population models*, *behavioural models*, *operational models* and *global models*.

Population Models. Population models focus on the healthy population and explore ways to reduce the number of individuals that become ill through disease prevention and health promotion interventions. Thus, the output of these models is largely used to inform policy decisions on public health interventions and prevention strategies.

One of the major focuses of population models is the growing rates of chronic diseases. Diseases such as cardiovascular disease, diabetes and cancer are becoming of great concern worldwide. Data from the United States indicate that preventable illness constitute approximately 70% of the illness burden and the associated costs. Preventable causes, such as cigarette smoking and obesity, represent eight of the nine leading causes of death in the United States [**80**]. This represents a huge challenge in finding ways to effectively promote lifestyle modification and prevent disease [**182**]. By reducing the number of people advancing from the healthy population to the at-risk population, the overall demand for healthcare resources may be reduced.

Examples of population models in healthcare can be found in Examples 4.2, 4.2, 4.3, 4.1, 4.2, 4.2, and 4.2 of this book.

Behavioural Models. Behavioural models address how people interact with healthcare providers and their peers to receive both expert and lay advice for managing their health. This is one of the main goals of the psychosocial models described in Chapter 10. Behavioural models aim to understand the social dynamics that contribute to fluctuations in service utilization. Thus, like population models, behavioural models can be used to look at how demand may be reduced before it is generated.

Patient-doctor interactions have been studied at the individual level using game theory [**63**] [**68**] [**211**] (see Chapter 11). Social network theory has also been applied to understanding the important roles that interactions with both peers and professionals play in determining healthcare demand [**175**] (see Chapter 12 Example 4.2).

Examples of behavioural models in healthcare can be found in Examples 4.1, 4.2, 4.1, 4.1, 4.2, 4.3, and 4.1 of this book.

Operational Models. Operational models are concerned with finding the most efficient strategies for processing the prevalent level of service requests. Often these are highly focused models studying exactly one aspect of healthcare. For example, models of staffing schemes and resource utilization within an emergency department. Thus, unlike population and behavioural models, operational models seek to manage demand as it arises, instead of trying to reduce demand before it is generated.

Operational models are most frequently applied within hospitals, clinics and other healthcare facilities and measure demand in terms of wait times or blocked patients (patients who sought healthcare but could not access it). Queueing theory (Chapter 15) and discrete event simulation (Section 4.3) have largely been the methods of choice for such problems, since random arrivals and queueing heavily influence the demand for service. Another method that is becoming more common in this field is system dynamics (Chapter 14).

Examples of operational models in healthcare can be found in Examples 4.2, 4.1, 4.3, 4.2, 4.1, 4.3, and 4.2 of this book.

Global Models. Global models are complex system models that may incorporate any of the previous three types in order to study interactions between multiple components of a healthcare system. They focus on understanding how changing one aspect of the global healthcare system (such as improving access to knee surgery) may effect demand in another aspect of the system (the requests for physiotherapy). This helps policy-makers determine whether a change in the system will have a positive or negative global affect.

Cost-containment is perhaps the toughest problem facing the healthcare system. Healthcare is absorbing a growing proportion of government budgets, but demographic explanations fail to fully account for this growth. Considering the healthcare system as a complex dynamical system is a potentially powerful means of analysing how a large number of components interact to produce unexpected outcomes. For example, an improvement in healthcare delivery in a specific setting may be accomplished by pulling resources from another setting. Disruptions of this type may result in excess demand and growing costs as a consequence of a large number of complex interactions.

Modelling at the global level is not simple, and much work remains to be done in this area. Recently, global models have often been implemented in terms of system dynamics (Chapter 14) [**111**] [**112**]. Other examples of global models in healthcare can be found in Examples 4.3, 4.2, 4.1, 4.2, 4.3, and 4.2 of this book.

8. Related Reading

For another description of the modelling process see [**42**]. Reference [**52**] discusses the concept of mental models. References [**63**], [**68**], and [**211**] examine game theory and some of its applications to healthcare. Reference [**175**] investigates social network theory. Reference [**13**] provides a taxonomy of 77 verification and validation techniques for conventional simulation models. References [**80**] and [**182**] develop models for use in health resource allocation. Reference [**42**] provides an introduction to simulation and modelling. References [**111**] and [**112**] look at system dynamics modelling in healthcare.

Part 2

Data Collection and Statistical Models

CHAPTER 4

Issues of Data

He used statiststics as a drunken man uses lampposts; for support rather than illumination. Andrew Lang (1844-1912)

There are two kinds of statistics, the kind you look up, and the kind you make up. Rex Stout (1886-1975)

Data Collection and Data Errors

Regardless of what is being modelled, or what modelling technique is being applied, at some level every model should be grounded in reality. Sometimes this grounding comes from consultation with experts in the field, or from logical deductions of how things work. Frequently, this grounding is established by performing some form of experiment or data collection regarding the system of interest.

The collection, or experimental creation of good data is an extremely difficult task in healthcare. In some aspects of healthcare, such as drug testing, data can be created in a "controlled" scientific manner, however in the vast majority of situations, data must be collected from historical events. Even in the case of controlled drug testing, tests can easily miss side effects which are slow in arising. This makes data collection in healthcare an extremely difficult task, which in turn makes grounding a healthcare model in reality a challenging task.

In this chapter we discuss possible methods for collecting data and some of the errors that can arise. We begin with some terminology to describe different types of data.

1. Types of Data

When dealing with the data aspect of modelling it is often useful to be able to describe when, how, and where the data was collected in a concise manner. In this regard, it is useful to provide some terminology on these concepts.

1.1. Data Collection Methods. The manner in which data is acquired is of key importance in determining the quality of the data. In the field of healthcare, there are three common methods of data collection: *experiments*, *health records*, and *surveys*. Table 1 summarizes the differences between these types of data collection, and highlights some advantages and disadvantages of each. First we discuss each data type in more detail.

Experimental data. The least common, but most reliable, source of data is generated via scientific experiments. By scientific experiments we refer to experiments that hold to the scientific principle of reproducibility. That is, the experiment can be repeated in a different time and place with the same (approximate) results.

Data Type		Advantages	Disadvantages
Experimental	Data collected through blind clinical trials.	- accurate and reproducible	- highly expensive in time and money - unethical or impractical in many cases
Health Records	Data collected by healthcare providers detailing when, where, and how patients access the healthcare system.	- accurate and contains technical health information	- generally does not contain information on personal health habits - biased against individuals who have not used the system
Survey	Data collected by contacting participants and requesting that they report the answers to certain questions.	- relatively quick and easy to collect - can be tailored to answer any desired question	- contains the highest room for error - can be expensive in some cases

TABLE 1. **Common Data Collection Methods:** Three common methods for collecting data, and their advantages and disadvantages.

In Physics or Chemistry a reproducible experiment is often an achievable goal, however in the field of healthcare reproducibility is extremely difficult to achieve. The problem lies in the fact that results in healthcare often hinge around how a single person reacts. Since no two groups of people are the same, expecting the same outcome from different groups of people is overly optimistic. Nonetheless, some experimental data exists. In healthcare, most experimental data is created in the form of so-called *blind clinical trails.*

The idea in a blind clinical trial is to give a random group of people either a drug or a placebo pill. The goal is to test if the drug has an effect that the placebo does not. The subjects given the drug are termed the *experimental group* and the subjects given the placebo pill are termed the *control group.* Blind clinical trails may be *single blind, double blind,* or *triple blind*, as described below.

Generally, if nine double blind tests show no correlation between a risk factor and a disease, but one double blind test shows a correlation, only the one "interesting" test will become public.

single: the subjects are unaware if they are in the experimental or control group.

double: neither the subjects nor the experiment administrators are aware of who is in the experimental or control groups.

triple: neither the subject, the experiment administrators, not the statisticians who analyze the data are aware of who is the in experimental or control groups. Although the statisticians are told whether a subject is group A or group B, they are not told whether A or B is the control group.

In most drug testing situations double blind tests are considered the minimal acceptable level of experimentation to produce accurate results. However, even results of double blind tests can be skewed by experimental error. We discuss this and some of its implications further in Section 2. One example of this is the fact that only experiments which produce "interesting" results tend to get published. For example, if nine double blind tests show no correlation between a certain drug

and disease, but one double blind test shows a correlation, then it is quite plausible that only the results of the test that shows a correlation will become public. This is generally referred to as *publication bias* (see Subsection 2.3).

Nonetheless, experimental data is generally considered the best possible source of data for modelling research. However, for a variety of reasons, experimental data is seldom collected. One of the more compelling reasons for this is that often the hypothesis is that a certain object is a risk factor to health. Ethically one cannot intentionally submit a collection of people to something that one believes will cause people harm. (Imagine, for example, performing a double blind test to determine if smoking cigarettes with a filter is less harmful than smoking cigarettes without a filter.)

Two other strong reasons for avoiding experimental data is the high cost of performing the experiment and the time required to perform the experiment. In healthcare, the costs of experimental data arise primarily from paying the participants, supplying the drugs, providing the test environment, and supplying the necessary expertise to ensure the experiment is performed safely. These costs very quickly add up to staggering numbers. Moreover, in healthcare, the time required to perform proper experimental tests is often highly impractical. When researching in terms of health, time-scales are generally in terms of lifetimes, or at least years. For example, although one could theoretically create an experiment that tests how the consumption of a vitamin supplement pill alters the probability of participants catching the flu, the experiment would have to be run in a controlled environment for years before completion.

Finally, in many cases in healthcare, the factors one wishes to collect data on cannot be altered in an experimental manner. A prime example of this is a participants socioeconomic status. Clearly an experimenter cannot provide a participant with a fixed income for the period of a lifetime to determine how this impacts their health.

Health Records and Survey Data. Given the difficultly in generating experimental data, healthcare research generally relies on other sources of data from which to draw its conclusions. Currently there are two common sources outside of experimental data: *health records* and *survey data*.

Health records and survey data both work on the principle that historical trends provide a naturally formed experiment. The difference lies in how the data has been collected. Health record data refers to data that is maintained by healthcare providers regarding when, how, and why individuals access given points in the healthcare system. Survey data is data is collected by contacting participants and requesting that they report the answers to certain questions through interviews or questionnaires.

Since health records are collected and maintained by health professionals, the data is often accurate and contains important health facts that the average individual cannot understand. However, health record data seldom includes information on an individual's personal habits, such as the frequency at which a person exercises. Another difficulty with health record data is that it only involves individuals who have actually used the healthcare system, while healthy individuals are unseen.

Survey data can get around both of these problems, as anyone can be asked to participate and any question can be asked. However, survey data has often been criticized as inaccurate, as participants tend to over-emphasize their "good" traits

and under-emphasize their "bad" traits. For example, it is widely accepted that the self-reported mass of an individual is usually lower than the actual mass of an individual. This is enhanced as an individual's mass increases.

Survey data can be collected in a variety of ways. Telephone surveys, mail-in surveys, and, more recently, web-based surveys are all common practice. Each method has advantages and disadvantages. We refer interested readers to Section 3 for references and further reading.

1.2. Time Series Data. In many cases, one wishes to examine how time is changing certain factors in a community. In order to do this, data must be collected at a series of points in time. Such data is called *time series data* or *serial data*.

Serial data may be collected in either a cross-sectional or longitudinal manner. *Cross-sectional* data refers to when a new collection of individuals is surveyed at each point in time[1]. This provides a series of "snapshots" of the population at various points in time and therefore provides some insight as to how the population dynamics are changing over time. *Longitudinal* data (sometimes called *panel data*) refers to when the same collection of individuals is surveyed at each point in time. This type of data is considerably harder to collect (as people must be recontacted several times over many years), but provides a higher level of insight into a population. Specifically, longitudinal data allows researchers to study how an individual changes over time. This type of data is necessary to properly tune some models.

In time series data, one separates the participants into a collection of cohorts. A *cohort* is a group of individuals from a given population that is defined by experiencing a common event during a particular time span. In healthcare the most common manner of grouping cohorts is by year of birth. However one might define cohorts by the year a mother gave birth (to examine changes in the impact of child birth on health) or the year of an individuals first entry into residential care (to study changes in the expected life-span of individuals in residential care).

A cohort *is a group of individuals from a given population that is defined by experiencing a common event (typically birth) during a particular time span.*

In health record data, patients are generally given a "health number" when they first access the system (or when they first enter the country). This allows for longitudinal data to be extracted from health records easily.

1.3. Electronic Health Records. It is easily arguable that electronic health records are one of the keys to modernizing the health system and improving access and outcomes. As electronic health records are implemented, high quality comprehensive data sets will become more readily available.

In Canada, a drive for the standardization of electronic health records is being headed by the *Canada Health Infoway*. Standardized health records will automatically ensure that any data originating from electronic health records will be of high quality. This has the potential to bring tremendous benefits to studies employing mathematical modelling and health data analyses.

2. Data Quality and Data Biases

The accuracy of data sources is often a major concern. Data errors can be broadly grouped into two categories: *sampling errors* and *non-sampling errors*.

[1]Occasionally the term cross-sectional data is used to refer to data that is not serial data. This can be thought of as the degenerate case where the series of points in time consists of only one point.

2.1. Sampling Errors. Sampling errors are errors that arise from estimating a population characteristic by looking at only one portion of the population rather than the entire population. That is, *sampling errors* are errors that result from a poor selection of the *representative sample*. For example, suppose one is collecting data on how reading is related to health. To do this one sets up a survey at the local library which asks patrons to state how many books they read each month and how they perceive their health status. On the surface this seems fine, but digging a touch deeper one realizes that this data set would consist of a very biased sample. The major flaw is that anyone who does not read would not go to the local library and therefore would be ignored in the study. Moreover, people who find it difficult to get to the library will be under represented, this would quite likely include the portion of the population that has lower than average health status.

Even when a "good" selection of the representative sample is considered, it usually contains sampling errors of some form or another. For example, the selection of random phone numbers from a telephone book will result in the exclusion of the unlisted portion of a population. Random door to door surveying will result in a bias towards people who are home more often. And regardless of sample selection, one will always be biased towards people who are more open and therefore more likely to respond to surveys. This final point is side-stepped with the use of health record data, but health record data is biased towards people who have accessed healthcare.

To avoid representative sample bias one should always try to obtain as large a sample as possible. One should also include demographical statistics in the data, and use these statistics to ensure that a "good" representative sample is selected. For example, if a data set has a significantly skewed gender ratio, then one should be careful to normalize the data with respect to gender before using it. Fortunately there is a large collection of literature on how to detect and deal with sampling errors. Unfortunately, there is very little known about how to prevent them.

2.2. Non-sampling Errors. *Non-sampling errors* are errors that result from reasons other than poor representative samples. These errors include poor survey design, poor survey or experiment implementation, and poor data storage.

Survey Design Errors. Whenever a data set is created through a survey, one runs the risk of a poorly designed survey skewing the results. Recall our example of performing a survey to help determine if reading is related to health (from Subsection 2.1). Such a survey might compose of any of the following questions:

- *Do you read at least two books a month?*
- *On average how many books do you read in a month?*
- *How many books did you read last month?*
- *What books did you read last month?*

All four of the above questions essentially estimate the number of books read within one month. However, the different phrasing may lead respondents to answer in very different manners. In the first question, a respondent may be led to believe that reading at least two books a month is the correct answer and therefore lean towards answering yes (even if they only read one and a half books a month). In the second, the respondent is no longer led to believe 2 is correct, but may still tend to answer with a higher number than true. Moreover, when respondents are asked for answers of which they are unsure, they tend to estimate and round to

"nice" numbers. For example a person is highly unlikely to answer 1.5 to the second question, even though this may be closer to the actual average number of books they read each month. The third and fourth questions are better in that respondents will more likely answer truly, but may cause problems in sampling error as people may read more books in July than in February. Moreover, it is unclear if you must have started reading the book, finished reading the book, or both during the given month for the book to count.

As complicated as this seems, it gets worse. For example it has been found that survey results depend on the order in which the questions are asked, and that there are differences between the way that people respond to written surveys versus oral surveys. The reasons for this have been attributed to people "learning" about themselves through the course of the questionnaire, and people being more candid about sensitive questions in written surveys.

Survey and Experiment Implementation Errors. One of the strongest concerns with survey data is that the vast majority of it is *self-reported*. That is, the respondent is asked a question and the answer is recorded without any effort to verify that the correct answer is given. In general, self-reported data tends to overemphasize what society considers to be "good". For example, self reported weight tends to be lower than actual weight, and self reported height tends to be higher than actual height. Similarly, people tend to under report their role in illegal activities (such as drug use) and over report their role in charitable activities.

Self-reported health status is often considered to be a more important determinant of healthcare utilization than actual health status.

It should be noted that, in some cases self-reported data is more important than scientifically accurate data. For example, self-reported health status (rather than actual health status) is often considered to be an important determinant of healthcare utilization.

Another challenge in implementing surveys is the characteristic low response rate. In practice, most surveys receive response rates significantly less than 50%. To combat this, marketing firms have developed various methods to increase response rates to a survey. Some simple suggestions include:

- keep the survey brief,
- provide an incentive (the respondent receives cash, gifts, lottery entry, etc. for completing the survey),
- guarantee anonymity, and
- explain to potential respondents how the survey can help society.

Many more techniques for improving survey response rates exist, but we leave those for curious readers and marketing firms to research.

One of the greater problems resulting from low response rates in surveys is interviewer frustration. If respondents are slow to understand or answer certain questions, the interviewer may begin to skip these questions, or lead the respondent to an answer. Note that this is not necessarily a sign of a corrupt interviewer, but is often an honest interviewer just trying to be helpful.

In terms of experimentation implementation, the most common error is due to an experimenter biasing the results as a result of not being blind to the participant's status. This is best avoided by performing double or triple blind experiments.

Data Storage Errors. From the above discussion it may appear that survey data is highly unreliable and whenever possible health record data should be used. However health record data can also have major errors in its collection. Most common is the fact that the health record data must be recorded by a human.

Aside from the random typographical errors that all humans are likely to produce, there is the tendency for health record data to overemphasize the positive results of an institute. For example, health record data on waitlists is generally computed by the difference in the time a person entered the waitlist to the time they exited the waitlist. Patients who do not exit the waitlist are therefore excluded from the calculations, which skews the data.

Another problem is that much of the health record data for hospitals is often inputed by nurses who have other duties. When hospitals are busy these other duties take precedence, and the data is not entered. Later it may be found that some data is lost by the time the nurse has enough free time to enter it. Thus data accuracy is only assured if the people who input the data are given sufficient time to do so.

2.3. Other Errors: Publication Bias and Data Interpretation. *Publication bias* is not truly a data error in that the accuracy of a data source in not the reason for the error. Publication bias is the result of the natural inclination for people to highlight the results they find most interesting.

Journals and researchers have a limited amount of resources. As such, they naturally seek to optimize those resources by focusing on results that appear most interesting. For journals, this means publishing papers that show a strong relationship between a risk-factor and a disease. In healthcare research this often means focusing on studies that showed strong correlations between risks and diseases. One problem is that random sampling error means that every study is flawed to some level or another. If ten studies are performed examining the health benefits of eating a daily serving of cucumber, odds are at least one study will find some relationship between eating cucumber and health status. The problem lies in the fact that the one study that found a correlation will likely be published, while the other nine studies that find nothing will likely be ignored.

Another source of publication bias comes from the natural desire to feel useful. Performing a data collection study is often a long and difficult process. Researchers wish to be rewarded for the work involved. Thus, one data collection study will often gather information on a large variety of topics, and researchers will then analyze the data thoroughly looking for any correlations within the results. On the surface this appears reasonable, but the same problem as before once again occurs. Due to data sampling errors, it is likely that at least one false positive will result in any data set. Searching the data set for this false positive causes the publication of results that are not truly representative of real life. We refer to such publication bias as *publication bias in situ* and discuss it further in Example 4.2.

There is no easy way to detect and unravel publication bias. The world of research has a certain professional drive to publish. Researchers who publish more are rewarded, while researchers who don't publish are often left behind.

2.4. A final note. In spite of the difficulties in collecting accurate data, whether it be experimental data, health records data or survey data, it is highly important to ground every model in reality. In general, if the data set is large, and some care is taken to ensure that it is a representative sample of the population, then data is an excellent way of doing this.

Finally, some researchers attempt to increase the size of their datasets by pooling data from various surveys or experiments. Results of studies based on pooling

of dataset should be examined carefully because of the difficulty in controlling for the varying biases within the datasets.

3. Related Reading

References [**1**] and [**26**] discuss a wide variety of data collection techniques, and how to cope with sampling errors.

There is a large collection of literature on how to detect and deal with sampling error; unfortunately, there is very little research on how to prevent it.

References [**174**], [**189**], and [**130**] compare self-reported height and weight to measured height and weight, and demonstrate a natural tendency for individuals to emphasis their positive attributes. Reference [**53**] looks at the reliability of self-reported health status as a measure of health, including analysis of how survey results depended on the order in which the questions were asked, and that there are differences between the way that people respond to written surveys versus oral surveys. Reference [**37**] provides an analysis of the predictive power of self-rated health for mortality in different socio-economic groups.

Reference [**121**] discusses various types of survey data, and appropriate techniques for analyzing each type. Reference [**225**] examines statistical and econometric techniques for the analysis of count data.

CHAPTER 5

The Basics

Smoking is one of the leading causes of statistics. Fletcher Knebel (1911-1993)

Statistically speaking, the probability of any of us being here is so small that you'd think the mere fact of existing would keep us all in contented dazzlement of surprise. Lewis Thomas (1913-1993)

Descriptive Statistics and Distributions

1. Model Overview

The basis concepts of statistics and probability have become such everyday part of life that most people have a natural feel for them. Listening to the daily weather report we hear the "p.o.p.", probability of precipitation, and decide whether to carry an umbrella that day. The sports news provides us with statistical updates on our favourite teams and players. Even the democratic electoral system is reported in terms of percentage of votes received. Yet despite the commonness of statistics and probability in everyday life, many people do not have a real understanding of these two disciplines and many results are under-reported, misstated, and misunderstood.

Statistics can be thought of as a collection of tools to analyse, interpret, and present data. On the other hand, *probability theory* is the field of mathematics devoted to determining the likelihood of a potential event occurring. In a sense, the fields of statistics and probability are like two sides of the same coin. Probability uses statistics to develop numbers on which to base predictions. Statistics uses probability theory to describe the underlying processes of potential events. The key difference is in the direction of thought. Probability theory examines the future, and provides tools to examine what might happen. Statistics examines the past, and provides tools to better describe what has happened. The overlap arises from the basic assumption that past behaviour is a good predictor of future behaviour.

After collecting or acquiring a large dataset, it is often very useful to provide some summary information regarding the data. This summary information is what we refer to as *descriptive statistics*. The goal of descriptive statistics is to provide an easy-to-read description of how the data is clustered.

In particular, researchers are often interested in the data's central tendency and degree of separation. The *central tendency* of the data is what many people call the *average*, and refers to descriptions of the data's most likely outcomes. Often this is reported in the form of the *mean*, but in some cases one might report the *median* or *mode* of the data. Mathematical definitions of these notions appear in Section 3, but for now it suffices to say they each provide a different manner of capturing the average.

The *degree of separation* refers to descriptions of how close the data elements are to the central tendency. This is an extremely important piece of information, as it tells you how well the central tendency describes a given data set. Unfortunately, many day-to-day sources of descriptive statistics (such as the newspaper) do not report the degree of separation. Most commonly, degree of separation is provided in terms of the *standard deviation*. A quick interpretation of the standard deviation is that approximately 68% of the data lies within one standard deviation of the mean, and 95% of the data lies within two standard deviations of the mean. Thus if the standard deviation is large, the mean provides a poor description of the data.

By reporting the degree of separation, the researcher provides confidence that the reported summary statistics are representative of the data. Another method to do this is through the form of *confidence intervals*. Confidence intervals provide an upper and lower bound for each summary statistic provided. For example, a researcher might state "we are 95% confident that the correct mean lies between 6.4 and 8.3." Generally, for analysis to be considered complete enough for publication, researchers are required to report a 95% confidence interval for their results. Unfortunately, outside of academic literature confidence intervals are seldom reported, making it difficult to determine if the supplied summary statistics are useful information. Policy makers should be wary of any descriptive statistics that do not include a report of either degree of separation or confidence intervals.

Another common approach to descriptive statistics of data is through the form of charts or graphs. Typically these are clear and easy to interpret, but it is possible to use graphs to display the data in a misleading manner (see Example 4.1).

Sometimes simple descriptive statistics are the goal of a research project, however in this book we assume the goal to be the development of a model to help answer a specific question. To develop a model one must switch their viewpoint from the past to the future, and therefore switch from statistics to probability theory. This causes us to switch from talking about how data looks, to how a random event might behave. In order to apply data to the creation of models one generally needs to determine the *probability distribution* of the data. The probability distribution describes the probability of occurrence for each possible outcome to a random event. The most famous probability distribution is the *normal distribution*, commonly called the *bell curve*. The normal distribution does a nice job of describing the outcome of a random event that is equally as likely to be too high as to be too low. However, many random events, such as the arrival time of the next patient, do not follow this pattern. Indeed any random event that has an achievable strict lower bound cannot follow a normal distribution. For example, the number of patients in a waiting room cannot be below 0 and is at least occasional 0, therefore cannot follow a normal distribution.

Some other distributions are discussed in Section 3.3 below. These distributions are more complicated and less commonly used than the normal distribution, but often very useful in developing healthcare models.

2. Common Uses

At some level statistical models are used in almost every form of modelling. Any time a researcher collects data, they usually begin by generating and providing descriptive statistics. Descriptive statistics are perfect for answering quick, trivia-like questions such as:

- *What is the most common age for an individual to take up smoking?*
- *What percentage of the population was obese in 1975, 1985, 1995, and 2005?*
- *What is the average age at which a woman first contracts breast cancer?*

To answer questions that are more general than these, one should turn to one of the more advanced forms of statistics discussed in Chapters 6, 7 and 8.

3. Mathematical Details

In mathematics the fields of statistics and probability are intimately intertwined. We begin this section with a discussion of the standard mathematical techniques used to summarize statistical data. In order to discuss the mathematics of confidence intervals, we next review basic probability theory and develop the ideas of probability distributions. We end with a discussion on confidence intervals and statistical significance.

3.1. Descriptive statistics. In order to provide an "at-a-glance" summary of data, most researchers will at some point rely on descriptive statistics. Often descriptive statistics are simple, commonly understood values that give the reader a sense of the data's central tendency and degree of separation. In other cases, researchers rely on charts and graphs to help describe the data.

Central tendency refers to descriptions of the data's most likely, or most commonly occurring outcomes. That is, if a random data sample was chosen, what type of value would one expect to see. The three most common measures of central tendency are *mean, median,* and *mode.* Given a data set $\{x_1, x_2, x_3, \ldots, x_N\}$ the *mean* of the data (often denoted by $\bar{x}$ or μ) is defined by the well known formula

$$\mu = \frac{(x_1 + x_2 + \cdots + x_N)}{N}.$$

where N is the number of elements in the data set. Thus the mathematical word "mean" corresponds to what is commonly referred to as "the average". To compute the *median* we begin by sorting the data: $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_N$. If N is odd then we define the median to be data element $x_{(N+1)/2}$. If N is even then we define the median to be the mean of data elements $x_{N/2}$ and $x_{(N/2)+1}$. Finally, the *mode* of the data is the most commonly occurring element. If the most commonly occurring element is not unique, the data is said to have multiple modes.

Degree of separation refers to descriptions of how close the data elements are to the central tendency. If the central tendency of mode is employed, this is best done by simply stating what portion of the data elements agree. If mean or median is used, then one often provides the *variance* or *standard deviation* of the data. For a given data set $\{x_1, x_2, x_3, \ldots, x_N\}$ the variance is denoted by σ^2 and defined by

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N},$$

where μ is the mean of the data. The standard deviation is then denoted by σ and is the square root of the variance.

Occasionally, some researchers will also provide a measure referred to as the *standard error of the mean.* This is defined as the square root of the variance divided by the sample size ($\sigma/\sqrt{N}$). This measure is useful for building confidence intervals, but should not be used as a descriptive statistic.

To interpret the standard deviation (or variance) we rely on the "Central Limit Theorem." Loosely the central limit theorem states that if the sample size is big and the data is selected from a consistent random distribution, then the result is a *normal distribution*. The normal distribution is formally defined in Subsection 3.3. For now it suffices to say the consequence of the Central Limit Theorem is that, approximately 68% of the time the random variable will lie within one standard deviation of the mean, and 95% of the time it will lie within two standard deviations.

As mentioned before, it is also common for descriptive statistics to take the form of charts and graphs. Typically these are clear and easy to interpret, but occasionally a researcher will display the data in a misleading manner. (Some examples of this are given in Subsection 4.1.) Policy-makers should be careful in interpreting graphs, and question any conclusions that are not clearly supported.

3.2. Basic Probability. The probability of the occurrence of an event E is denoted $\Pr(E)$. It is defined as the number of ways the event can occur divided by the number of possible outcomes,

$$\Pr(E) = \frac{\#\text{ of ways } E \text{ occurs}}{\#\text{ of possible outcomes}}.$$

The probability of an event E given a known set of factors F is the number of ways the event can occur given the factors divided by the number of possible outcomes where the factors are present,

$$\Pr(E|F) = \frac{\#\text{ of ways } E \text{ occurs given } F \text{ is present}}{\#\text{ of possible outcomes where } F \text{ is present}}.$$

These two definitions form the basis of probability theory in mathematics, while the remainder of the theory is largely focused on how to determine the number of ways events can occur with or without certain factors present.

A classical example is the rolling of a pair of 6 sided dice and then examining the sum of the numbers produced. To simplify this discussion let us paint the first die green and the second die red. The green die can take on any one of six possible outcomes, the red die can do the same. As such the total number of possible outcomes is 36. However, if we consider the sum of the two rolled dice, then the total number of possible outcomes becomes 11: $(2, 3, 4, \ldots, 12)$. These outcomes are listed in Table 1.

Green Die ⇒ Red Die ⇓	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

TABLE 1. Possible outcomes resulting from the summing two rolled dice

From Table 1 we can easily count to see there are 6 ways to get a sum of 7. Therefore the probability of totalling 7 is

$$\Pr(total = 7) = \frac{6}{36}.$$

Similarly we can see the probability of totally 5, 6, or 9 to be

$$\Pr(total = 5) = \frac{4}{36},\ \Pr(total = 6) = \frac{5}{36},\ \Pr(total = 9) = \frac{4}{36}.$$

Notice that the probability of totalling 5 is equal to the probability of totalling 9, this means both of these events are equally likely to occur.

The study of probability is largely attributed to Blaise Pascal, who is rumoured to have developed it for the purpose of winning at games of dice.

From the table we may also compute the probability of totalling 5 given that the green die rolls a 4. Notice since the green die is fixed at 4 there are now 6 possible outcomes, exactly one of which is a total of 5. Thus,

$$\Pr(total = 5|green = 4) = \frac{1}{6}.$$

Similarly we find

$$\Pr(total = 7|red = 3) = \frac{1}{6},$$

$$\Pr(total = 5|green = 5) = \frac{0}{6},$$

$$\Pr(total = 12|total \geq 9) = \frac{1}{10}.$$

Notice that in some cases (such as $\Pr(total = 5|green = 5)$), the probability can be 0. When this occurs we say the event and factors are *mutually exclusive,* that is the event cannot occur given the factors listed.

In the case of rolling two dice the probabilities were simple enough to work out by writing out the entire table of possible events. In most cases this is not true (consider for example rolling 3 dice, 4 dice, 10 dice, etc.). Instead researchers rely on the well developed fields of combinatorics and probability. Although the majority of these fields are beyond the scope of this book, we do take a brief look at probability distributions.

3.3. Probability Distributions. In the previous subsection we discussed a simple example of a random event, rolling a pair of dice. In this example we developed a table (Table 1) that described all possible outcomes for rolling the dice. Using this table we were able to determine the complete list of outcomes and their probabilities. We present this in Table 2.

x	2	3	4	5	6	7	8	9	10	11	12
$\Pr(total = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

TABLE 2. Probabilities for each outcome of the sum of two rolled dice

Table 2 represents what is called a *probability distribution.* If the event E takes on a finite number of values then the function

$$f(x) = P(E = x)$$

is the *finite probability distribution* for the event E. Finite probability distributions have a number of distinctive properties. For example, since each value is a probability, we have $0 \leq f(x) \leq 1$ for all x. And, since the sum of $f(x)$ over all x represents the probability of some event occurring we have $\sum_x f(x) = 1$.

Probability distributions allow us to compute the *expected value* of a random variable. This is, as the name suggests, the value that is expected to occur on average if the distribution is sampled from repeatedly. More mathematically, if n random variables are selected from a given probability distribution, then the mean of these values should approach the expected value of the distribution as n grows to infinity. The expected value for a finite probability distribution can be computed from the formula $E(X) = \sum_x xf(x)$. For the example provided in Table 2 this evaluates to

$$\begin{aligned} E(X) &= \sum_x xf(x) \\ &= 2\tfrac{1}{36} + 3\tfrac{2}{36} + 4\tfrac{3}{36} + 5\tfrac{4}{36} + 6\tfrac{5}{36} + 7\tfrac{6}{36} + 8\tfrac{5}{36} + 9\tfrac{4}{36} + 10\tfrac{3}{36} + 11\tfrac{2}{36} + 12\tfrac{1}{36} \\ &= 7. \end{aligned}$$

In healthcare, the most commonly arising finite distribution is the *Poisson distribution.* The *Poisson distribution* expresses the probability of a number of events occurring in a fixed period of time, if these events occur with a known average rate and are independent of the time since the last event. A perfect example is the number of newly arriving patients into a hospital in a given hour. The function defined by

$$f_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, \ldots$$

is the finite probability distribution for the Poisson distribution. This distribution has one parameter λ that represents the expected number of occurrences during one time interval. The Poisson distribution is discussed further in Subsection 3.4.

Although finite probability distributions are simple to understand, they are not practical in many situations. For example, if we consider the random variable of an individual's mass, it is clear that there is not a finite number of options. (One could make a finite number of options by restricting mass to the nearest kilogram, but realistically an individual can be 70kg, 70.5kg, 70.00343kg, etc.). To deal with random numbers that can take on any value, continuous distributions are used.

Recall, $\int_b^a f(x)dx$ is the integral of $f(x)$ from a to b, and is equal to the area under the curve defined by $f(x)$ from $x = a$ to $x = b$.

Unlike finite distributions, continuous distributions cannot take the form of a function. (Consider, if x can take on an infinite number of values, then the properties $f(x) \geq 0$ for all x and $\sum_x f(x) = 1$ will be very difficult to achieve.) Instead we use what is referred to as a *probability density function.* A function f is is a *probability density function* (**pdf**) for a continuous distribution if $\Pr(y_0 < E < y_1) = \int_{y_0}^{y_1} f(x)dx$. What this means is that the probability of event E lying between y_0 and y_1 is equal to the area under the curve $f(x)$ between y_0 and y_1.

Probability density functions come in many shapes and forms, but like distribution functions they all satisfy two conditions. First, like distribution functions, **pdf**s are always positive: $f(x) \geq 0$ for all x. Second, similar to distribution functions, the area under the entire **pdf** must be equal to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$. Beyond this, **pdf**s may be as simple or as complicated as one requires to describe the event.

We can also use **pdf**s to compute the expected value of a distribution. The expected value for a continuous probability distribution given by the **pdf** $f(x)$ can

be computed from the formula

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

This is the area under the curve defined by the formula $xf(x)$ (when the curve is below zero, the area is subtracted).

If E has a finite number of outcomes then the function $f(x) = P(E = x)$ is the finite probability distribution for E.

There are many different **pdf**s that are studied and used in probability theory, and it is beyond the scope of this book to go into them all in detail. However, there are several that are of particular interest in healthcare, and we go into these now.

Multinomial distribution: The *multinomial distribution* results from having a **pdf** that is formed from a series of steps (see the leftmost graph in Figure 1). The use of this distribution function is simply to recreate finite distributions in the framework of **pdf**s.

If E has an infinite number of outcomes then the function $f(x)$ is a probability density function of the distribution if $\Pr(y_0 < E < y_1) = \int_{y_0}^{y_1} f(x)dx$.

Empirical distributions: Related to multinomial distributions, *empirical distributions* are created to exactly match observed data. Such distributions are useful for testing if the type of distribution used has a significant impact on the system being researched. However, when implementing complicated models however, using the empirical distribution limits our implementation options.

Normal distribution: The **pdf** for the *normal distribution* is given by the classical "bell curve:"

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};$$

(see the center graph in Figure 1). This distribution has two parameters μ and σ that correspond to the expected value of the random variable and the standard deviation of the distribution respectively. Key aspects of the normal distribution are that it is symmetrically distributed about the mean, and drops off as it moves away from the mean. It follows the **68, 95, 99.7** rule, which states that 68% of results lie within one standard deviation of the mean, 95% of results lie within two standard deviations of the mean, and 99.7% of results lie within three standard deviation of the mean. The normal distribution is the natural choice for any continuous random variable that is equally likely to be above the mean as below the mean.

The normal distribution is also referred to as the bell curve, or Gaussian distribution

Exponential distribution: The *exponential distribution* is the distribution which represents the random time between consecutive random events in a process with no memory. This is strongly related to the Poisson distribution. The Poisson distribution fixes a time frame and examines how many random event occur in the time frame, the exponential distribution fixes the number of random events and examines the random time length between occurrences. The **pdf** for the exponential distribution only relies on one parameter, λ, and is defined as

$$f_\lambda(x) = \lambda e^{-\lambda x} \; x > 0.$$

The exponential distribution has one parameters λ that relates to both the expected value and the standard deviation of the distribution. The mean of the exponential distribution is equal to $\frac{1}{\lambda}$, and the variance is equal to $\frac{1}{\lambda^2}$. Due to its nature, the exponential distribution is only defined for nonnegative values.

3.4. Poisson and Exponential Distributions in Healthcare. The Poisson and exponential distributions are a natural choice for modelling arrival rates for several reasons. Foremost is the logical supposition that an individual visiting a doctor or specialist is an independently occurring event that happens at a constant average rate. That is, the event of one individual visiting a given specialist in a

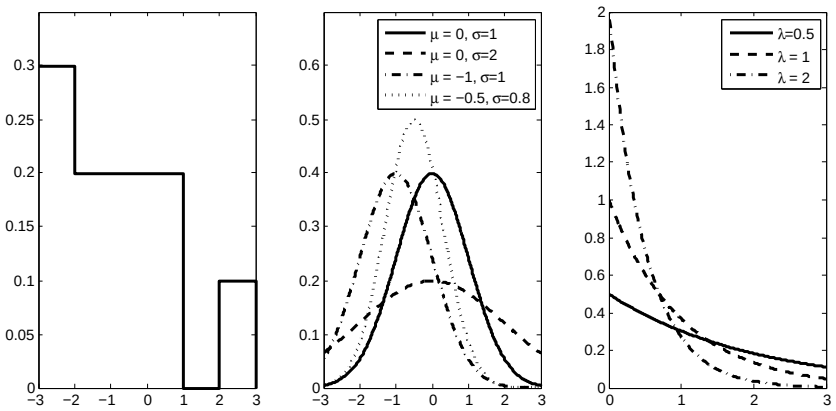

FIGURE 1. **Various probability distributions:** The probability distribution functions for a multinomial distribution (left), various normal distributions (middle), and various exponential distributions (right).

given month (hour, day, year, etc.) does not make a different individual more or less likely to visit the specialist in the same month (hour, day, year, etc.).

There are some reasons, however, why the Poisson distribution may not ideally describe healthcare utilization. For one, a restrictive feature of the Poisson distribution is that the mean and variance of a random variable following this distribution are equal. This notion is called *equidispersion*. Departures from equidispersion can occur if the variance is either greater than the mean (overdispersion) or less than the mean (underdispersion). Overdispersion, which can be caused by a large number of zero counts in the data set, is unfortunately common in healthcare data, as there will be some people who never utilize services.

There may also be reasons to question that events occur independently at a constant rate. Events in health utilization data are not always independent nor is the probability of event occurrence always constant (for example, multiple visits to a physician by the same patient are often related). Therefore, it seems likely that the Poisson assumptions are often violated in health utilization data. As a result, a more flexible generalization called the negative binomial distribution has been considered. (The **pdf** for the negative binomial distribution is nontrivial, and beyond the scope of this book.) For the negative binomial distribution, the variance and mean are unlinked, which may better model healthcare utilization counts. In addition, the negative binomial is not sensitive to event dependency and variable event probabilities, so it is often considered to be an attractive alternative to the Poisson distribution for modelling healthcare utilization data.

3.5. Confidence Intervals, and Statistical Significance. Once a probability distribution is selected, the mean and variance of a data set can be used to construct *confidence intervals* for the data. Determining a confidence interval begins by selecting a desired *degree of confidence* $1-\alpha$. For example, if one desires 95% confidence then $\alpha = 0.05$. After the degree of confidence is selected, the $1-\alpha$ confidence interval for a random value X is defined as the range of values $[x_0, x_1]$

such that X has a $(1-\alpha)\times 100\%$ chance of lying within. Mathematically this is,

$$\Pr(x_0 \leq X \leq x_1) = 1 - \alpha.$$

Determining a confidence interval is not a trivial task, but fortunately most statistical software is capable of performing it for the user. Therefore, instead of discussing how to compute confidence intervals, we discuss how to interpret the output of common statistical software.

Most statistical software will provide a mean μ, a standard error of the mean $\sigma/\sqrt{n}$, a *z-value* z, and an associated probability (either $\Pr(< z)$ or $\Pr(> z)$ depending on the software). If the associated probability is given in the form $\Pr(< z)$ then this is the value of $1-\alpha$, if it is given in the form $\Pr(> z)$ then this is the value of α. By combining these one can generate a $1-\alpha$ confidence interval:

$$\mu - z\frac{\sigma}{\sqrt{n}} < \mu_{\texttt{true}} < \mu + z\frac{\sigma}{\sqrt{n}}.$$

That is, there is a probability of $1-\alpha$ that the true mean of the distribution is between $\mu - z\frac{\sigma}{\sqrt{n}}$ and $\mu + z\frac{\sigma}{\sqrt{n}}$. Depending on the software the user may be able to input the desired α and the computer will return the appropriate z value. Clearly the desired result is for both the standard error of the mean, the z value, and the value of α to be small. Generally an $\alpha < 0.05$ is considered statistically significant, and an α greater than this value is considered insufficiently small to develop any conclusions. More discussion on the relevance of confidence intervals can be found in Example 4.2.

4. Examples

4.1. "9 out of 10 Doctors Agree" – Interpreting Descriptive Statistics. Let us pretend that a pharmaceutical company has decided to try and convince the public that its headache medicine is "better" than its competitors'. In order to distinguish the two, we call the first company Company A and their four major competitors Companies B, C, D and E. Company A begins by collecting some data on the cost and effectiveness for each product. The cost is easily obtained by going to the local drug store and asking the price of a package of 24 tablets for each type of headache medicine. To compare the effectiveness of each product, the company examines the number of milligrams of pain medicine per tablet for each medicine, and the number of tablets in a recommended dose. In addition to this, the company performs a case study of 500 people in which it asks each person to use a specific brand of headache medicine for one month, and then rate the medication as either "not effective (1)," "somewhat effective (3)," or "completely effective (5)." Participants who report not requiring headache medication during the month of the study are excluded. The findings of this research is found in Table 3.

To present their results to the public, Company A produces the pamphlet found in Figure 2. Let us critically examine this figure. The leftmost graph in the figure shows the mean value for the survey data they collected. At a glance it would appear that Brand A is significantly more effective (according to this survey) than any other brand, especially Brand E, which appears to be the least effective. Let us do a more detailed statistical analysis of the results, in particular comparing Brand A with Brand E. First notice that as the survey was not presented within the pamphlet, the public is left with the question of what is meant by a value of 5, 4, 3, 2, or 1. (Note that there was actually no value for 4 or 2, something the public

Company	Cost for 24 tablets	mg of medicine per tablet	tablets per dose	mg of medicine per dose	% rating effect 1	% rating effect 3	% rating effect 5
A	8.99	175	3	525	5	25	70
B	11.99	250	2	500	5	35	60
C	10.99	225	2	450	10	20	70
D	10.49	225	2	450	20	0	80
E	14.99	400	1	400	10	30	60

TABLE 3. Results of Company A's research into cost and effectiveness for various headache medications.

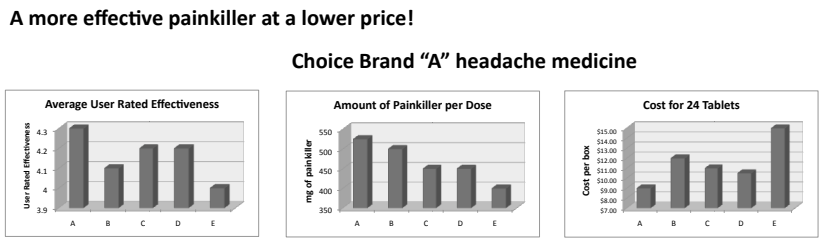

FIGURE 2. **An example of poor descriptive statistic:** "Descriptive Statistics" developed by Company A.

does not know.) Next, notice the scale on the y-axis does not start at 0. Indeed the effectiveness of Brand E may appear 4 times lower than the effectiveness of Brand A, but the actual mean values are $\mu_A = 4.3$ (for Brand A) and $\mu_E = 4.0$ (for Brand E). Moreover, the standard deviation in the data for Brand A is $\sigma_A = 1.15$, and for Brand E is $\sigma_E = 1.35$. Since 100 people were sampled for each group, this gives a standard error of the mean of $\sigma_A/\sqrt{100} = 0.115$ for Brand A, and $\sigma_E/\sqrt{100} = 0.135$ for Brand E. Looking up the z value for a normal distribution associated with a confidence of $1 - \alpha = 0.95$ we find $z = 1.645$ (as can be found in most statistics textbooks). Thus we have the confidence intervals

$$4.3 - 1.645(0.115) < \mu_{A,\texttt{true}} < 4.3 + 1.645(0.115) \Rightarrow 4.110825 < \mu_{A,\texttt{true}} < 4.489175,$$

and

$$4.0 - 1.645(0.135) < \mu_{E,\texttt{true}} < 4.0 + 1.645(0.135) \Rightarrow 3.777925 < \mu_{E,\texttt{true}} < 4.222075$$

for Brand A and Brand E respectively. The fact that these two intervals overlap implies that there is no statistically significant difference between the two means at the standard 95% confidence level.

Let us now turn our attention to the two graphs on the right, "Amount of Painkiller per dose" and "Cost for 24 Tablets." Together these charts make it appear that Brand A is providing significantly more painkiller per dose, and cost significantly less. Both of these statements are false. Notice that as before, neither chart begins at zero. Further more, the cost given is per 24 tablets, not per dose. Since Brand A requires 3 tablets per dose, a box of 24 tablets actually only contains

8 doses, while Brand E, which appears the most expensive, only uses one tablet per dose.

The conclusion of this example is that one must be wary of descriptive statistics. Although there is certainly a place for descriptive statistics in research, one should not attempt to draw any conclusions without being provided a more detailed analysis. By providing some select descriptive statistics it is not difficult to manipulate data to have it appear to support the conclusion(s) one desires.

4.2. Age, Health, and Confidence Intervals. The traditional definition of "old" or "senior citizen" usually refers to individuals of 65 years of age or older. This concept dates back to the post World War II era of 1955, when the life expectancy of a newborn in most industrialized countries (including Canada and the USA) was slightly over 65 years [**25**]. Today, a newborn American child has a life expectancy of just under 80 years, and a newborn Canadian has a life expectancy of just over 80 years[1]. Clearly then, the traditional definition of "old" is no longer appropriate. Indeed, over the past 160 years the life expectancy of the leading countries have risen **linearly** by three months per year [**25**]. (The United States and Canada have both followed this trend for at least the past 50 years.) Various scientists have repeatedly forecast a levelling of this trend, but this has yet to occur [**25**]. This suggests that any predictive model that involves population age distributions as a proxy for population health status will eventually become flawed.

In order to counteract this problem, some researchers are beginning to examine health status, and healthcare usage, in terms of distance from death, as opposed to distance from birth. A 2003 study by Yang, Norton, and Stearns, using American data from the *Medicare Current Beneficiary Survey Cost and Use Files* clearly demonstrates this idea [**230**]. In this example we quickly review this experiment, and highlight how confidence intervals play a very important role in the analysis of how health status interacts with age.

Yang, Norton, and Stearns' work begins by compiling the data into three distinct age groups: 65-74, 75-84, and 85+. Their results are nicely summarized by two figures, which we reproduce in Figure 3.

On the left side of Figure 3 we see Yang, Norton, and Stearns' figure regarding healthcare expenditure by age. One immediately notices that these expenditures rise with age. However, one should also notice that the 95% confidence intervals of these expenditures expand very rapidly. Both of these trends are explained by the pair of figures on the right side of Figure 3. This pair of figures shows that healthcare expenditure is very strongly correlated to proximity to death, and that the likelihood of death increases with age. In these figures we see that the error bounds remain much steadier.

Without examining the confidence intervals, the left side of Figure 3 would create a very convincing argument that healthcare expenditures are a function of age alone. However when confidence intervals are considered, the certainty of that relationship becomes less secure. A correlation still exists to be sure, but it is evident that it is not age alone, but age as a proxy to something else (nearness to death in this case) that carries most of the strength in the relationship. This means that if the connection between age and nearness to death changes, (as it has

[1]The CIA world factbook provides USA's life expectancy at birth in 2007 as 78.14 years, and Canada's life expectancy at birth in 2007 as 81.16 years. `https://www.cia.gov/library/publications/the-world-factbook/`

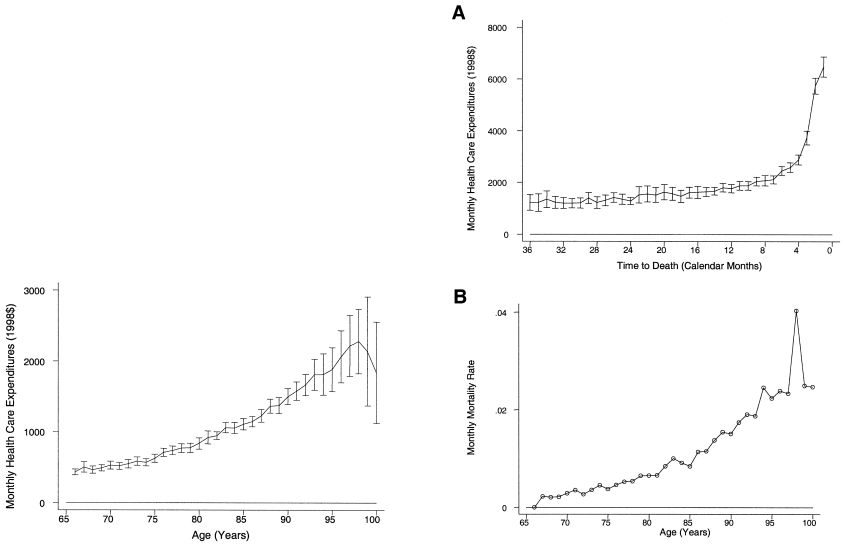

FIGURE 3. **Age versus proximity to death in healthcare expenditures:** Relating age and proximity to death to healthcare expenditures.
Left: Healthcare expenditures by age with 95% confidence intervals.
Right: (A) Healthcare expenditure by proximity to death with 95% confidence intervals and (B) monthly mortality rate by age.
Reproduced from [**230**, Fig. 1] (left) and [**230**, Fig. 2] (right).

and will likely continue to do over long periods of time) the use of age alone to predict costs will create an error. This demonstrates the importance of reporting more than just basic statistics in research.

4.3. Private Accommodation Environments in British Columbia. In British Columbia (BC), Canada, Home and Community Care (HCC) provide a range of healthcare and support services for acute, chronic, palliative or rehabilitative healthcare needs. Unlike most healthcare services in BC, HCC has a prominent private sector. To understand the drivers of HCC demand, it is valuable to understand the drivers of non-publicly funded HCC services. However, since non-publicly funded HCC facilities lie outside of the jurisdiction of the BC Ministry of Health Services, the Ministry knows relatively little about the quantity or quality of non-publicly funded accommodation environments in BC. To reduce this knowledge gap, in 2007 Dodd and Hare performed a province wide telephone survey of accommodation environments in BC. In this example we outline the method and some of the results of this survey.

The survey began by seeking names, telephone numbers, and addresses for non-publicly funded accommodation environments in BC. Facilities were cross listed with a list of publicly funded facilities provided by the BC Ministry of Health to determine which facilities received some public funding. Facilities that were

receiving no funding from the BC Ministry of Health were labelled **Type 1** facilities. Facilities that were receiving some funding from the BC Ministry of Health were labelled **Type 2** facilities. A telephone survey of all facilities was then conducted.

Each Type 1 facility was telephoned and asked a series of questions, including:

- How many people can your facility support?
- Is there a wait-list to enter your facility?
  If yes: How long do people usually have to wait?
  If no: How many more people could you accept into your facility?

As it is possible for a publicly funded facility to have some non-publicly funded beds, Type 2 facilities were also contacted. For these facilities the above list of questions was prefaced with the question: "Do you have any non-publicly funded beds in your facility?" If the answer was yes, the facility was asked the same questions as Type 1 facilities, but requested to answer only in regards to their non-publicly funded beds.

Analysis of the survey data began by determining if a facility was a reasonable approximation of a BC accommodation environment. To be considered a reasonable approximation, facilities had to provide residents with a bed, full meal service and a 24 hour emergency medical response system. Type 2 facilities that answered "no" to the preface question ("Do you have any non-publicly funded beds in your facility?") did not fulfill the requirements. A basic break down of facility numbers and bed counts is provided in Table 4.

	Type 1	Type 2	Total
Potential Facilities	368	419	773
Successful Contacts	212	199	411
Fulfill Requirements	133	46	179
Number of Counted Beds	9590	1963	11553

TABLE 4. **Survey of Non-publicly funded Accommodation Environments in BC:** Type 1 facilities receive no funding from the BC Ministry of Health Services, Type 2 facilities receive some funding from the BC Ministry of Health Services. Successfully contacted facilities represent facilities that were reached via telephone and agreed to participate in the survey. To "Fulfill Requirements" means that the facility provides a minimum of a bed, full meal service and a 24 hour emergency response system, and the facility receives some private funding.

As one can see from Table 4, there are **at least** 11,553 accommodation environment beds in BC that are not publicly funded. However, we also see that of the potential 773 facilities, approximately half (411) were successfully contacted. (A facility is considered successfully contacted if it was reached via telephone and agreed to participate in the survey.) Given the large number of facilities that were not successfully contacted, 11,553 is likely to be much lower than the actual bed count. A quick, realistic, expected bed count for the number of non-publicly funded accommodation environments in BC can be computed as follows. Since 133 of the 212 successfully contacted Type 1 facilities fulfilled the requirements, we estimate

that $133/212 = 62.7\%$ of the uncontacted Type 1 facilities will fulfill the requirements. As the average number of beds in successfully contacted Type 1 facilities that fulfilled the requirements is 9590/133, we estimate the total number of Type 1 facility beds in BC to be

$$9590 + 156 \times \frac{133}{212} \times \frac{9590}{133} \approx 16647. \tag{1}$$

Similarly, we estimate the total number of Type 2 facility beds in BC to be

$$1963 + 220 \times \frac{46}{199} \times \frac{1963}{46} \approx 4133.$$

This provides a total estimate of 20,780 beds.

Of course these estimates are very basic. In particular, they do not take into account that facilities in more urban areas are larger than facilities outside of urban centers. This can easily be addressed by breaking down our examination of beds into geographic areas. BC is separated into 5 Health Authorities (HA) and 16 Health Service Delivery Areas (HSDA). In order to gain a greater understanding of non-publicly funded accommodation environments in BC, Dodd and Hare break down the survey results into HSDA level estimates by performing the same basic estimate technique:

$$\begin{aligned} \textit{known beds} \quad + \quad & \textit{uncontacted facilities} \times \\ & \textit{likelihood of fulfilling requirements} \quad \times \\ & \textit{average bed count.} \end{aligned} \tag{2}$$

Now however, *likelihood of fulfilling requirements* is defined by

$$\textit{likelihood of fulfilling requirements} = \frac{\#\ \textit{fulfilling requirements in HA}}{\#\ \textit{potential facilities in HA}}.$$

Similarly, *average bed count* is now defined by

$$\textit{average bed count} = \frac{\#\ \textit{known beds in HA}}{\#\ \textit{fulfilling requirements in HA}}.$$

Table 5 reports the results of this technique.

It should be noted that, although estimated bed count is given at the HSDA level, the *likelihood of fulfilling requirements* and *average bed count* used in equation (2) are still calculated at the HA level. This is done, as several HSDAs lack a sufficent number of successfully contacted facilities to provide reasonable estimates values. For example, in the Northeast HSDA only one Type 1 facility was successfully contacted, and it did not fulfill the requirements. Therefore, if the HSDA values were used for the Northeast HSDA, we would return an estimate of 0 Type 1 facility beds. (This may be correct, but it seems less likely than the 12 reported in Table 5.)

Summing the Type 1 and Type 2 facility bed counts estimated in Table 5 we estimate that there are 21,220 non-publicly funded accommodation environment beds currently in BC. Considering that in March of 2007 the Ministry of Health Services funded 27,967 assisted living and residential care beds, we can see that non-publicly funded beds are filling a significant portion of total province wide services.

There are several potential sources of error with these preliminary results. First, telephone numbers were obtained via internet and phonebook searches. It is likely

HSDA name	Estimated Type 1 Bed Count	Estimated Type 2 Bed Count
East Kootenay	163	549
Kootenay Boundary	64	329
Okanagan	4062	419
Thompson Cariboo Shuswap	1238	330
Fraser Valley	1124	75
Simon Fraser	1818	468
South Fraser	1768	683
Richmond	174	131
Vancouver	1516	409
North Shore/Coast Garibaldi	713	211
South Vancouver Island	2287	447
Central Vancouver Island	1249	123
North Vancouver Island	347	38
Northwest	37	32
Northern Interior	238	134
Northeast	12	32
Total BC	16809	4411

TABLE 5. **Non-publicly funded Accommodation Environments in BC by HSDA:** Type 1 facilities receive no funding from the BC Ministry of Health Services, Type 2 facilities receive some funding from the BC Ministry of Health Services.

that some non-publicly funded facilities were missed during this information gathering stage. Second, data was only obtained for less than half of the facilities. Therefore we have a survey bias towards facilities that are more likely to have a full time secretarial staff. As it would seem likely that larger facilities would be more likely to have full time secretarial staff, this may have lead to an over-estimation of the number of beds.

Nonetheless, these results give a **clearer** picture of non-publicly funded accommodation environments in BC by providing a snapshot of the current private sector. The research of Dodd and Hare continued by analyzing the data using common demographic factors and linear regression (see Chapter 7). Further information can be found in reference [**101**]. Of note in this example, is that although the survey and statistics generated from the survey were simple descriptive statistics, the numbers provide a much clearer picture of non-publicly funded HCC services in BC. When developing models of healthcare demand, it is often useful to begin with straightforward surveys such as the one listed here, in order to gain an intuitive understanding of the issues involved.

5. Related Reading

Descriptive statistics are the basis of Regression Analysis (Chapter 6), as well as Epidemiology (Chapter 7). At some level all quantitative models should be based in reality. This is often done by comparing model output to Descriptive Statistics.

Reference [**230**] provides details for example 4.2. Reference [**148**] provides further evidence that proximity to death is a better estimator of healthcare expenditures than age.

Reference [**64**] discusses methods of statistically analyzing utilization data in healthcare. Reference [**41**] develops a microeconometric model of the demand for health insurance. Reference [**40**] discusses regression analysis. Reference [**116**] reviews some of the misuses of statistics as well as some of the techniques employed to distort results. References [**176**] and [**177**] discuss potential publication biases and data analysis errors, including publication bias in situ.

CHAPTER 6

Predictions and Responses

I don't try to describe the future. I try to prevent it. Ray Bradbury (1920-)

When men speak of the future, the gods laugh. Ancient Chinese proverb

Regression Analysis

1. Model Overview

In 2003, researchers from the University of California, Berkley, completed a four year study of how much data existed in the world[1]. The results showed that from 1999 to 2002 the amount of stored data in the world approximately doubled. This corresponds to a growth rate of approximately 25% per year (consider that the population of the earth is growing at a mere 1.14% per year[2]). Approximately one quarter of this data is in the form of electronically stored statistics. With this in mind, policy makers worldwide are left with the increasingly daunting task of sifting through this data in an attempt to make better decisions.

Fortunately for policy makers, the recent past has also seen great growth in statistical analysis techniques and software. For policy makers in healthcare, the interest in data analysis often lies in developing a quantitative relationship between a set of variables and a possible outcome. For this, researchers often turn to the field of *regression analysis*.

The word model *in statistical literature usually refers to an equation to which one tries to fit data via regression analysis.*

In mathematics, regression analysis refers to modelling techniques designed to uncover relationships between a dependent variable and one or more independent variables. When data is collected, one of the variables measured is often considered to be a response or outcome of interest. The other variables measured are explanatory or predictor variables. In healthcare, researchers generally attempt to develop a formula or equation that relates the *predictor variables* to the *response variable*. In literature the words *model, model function,* or *statistical model* are often used to refer to the formula that relates these two types of variables. When the model functions are based in economic theory, literature often refers to the results as *Econometrics* (meaning economic measurement). However, other approaches of developing model functions exist.

It should be noted that the terms *explanatory variable* and *predictor variable* are used interchangeably. We favour the term predictor variable to emphasize that

[1] "How Much Information? 2003" `www2.sims.berkeley.edu/research/projects/how-much-info-2003/`

[2] CIA world factbook: `www.cia.gov/cia/publications/factbook/`

when choosing them, it is important that these variables be measurable before the response variable is known. For example, suppose we are interested in determining an equation that will help provide an estimated cost for a patient undergoing chemotherapy to treat cancer. The response variable is the "cost of treatment." Explanatory variables might include the initial size of the cancer discovered, the age of the patient, and the number of chemotherapy sessions the patient uses. The first two of these variables are predictive, in the sense that they can be determined before treatment is complete. The last of these, the number of chemotherapy sessions, is not predictive as it cannot be determined until treatment is complete. Of course, at this point the total cost of chemotherapy is already known, so in a practical sense the third variable is not particularly useful for developing healthcare policies.

Three possible explanatory variables for determining the response variable "time required to recover from hip replacement surgery" could be:
1. Age of patient.
2. Body Mass Index of patient.
3. Number of post-surgery physiotherapy sessions.
Although the third provides a very high degree of explanation, it is not particularly useful as it cannot be determined until the response variable is already known. Therefore it is not a good predictor *variable.*

Once the response variable and predictor variables have been determined, statistical techniques can be used to understand and predict how the value of the response variable will change for different values of the predictor variables. This process begins with the creation of a hypothetical model under which the data is likely to fit. After creating a hypothetical model, a statistical tool called regression analysis is used to fit the data to the model. If a good fit can be created then this helps validate the model, if no fit can be found then the model is rejected and the process begun again.

The creation of the hypothetical model is essentially the creation of a mathematical equation which one feels describes the "shape" of the data. For example, we might expect the cost of running a hospital would increase in direct proportion with square footage of the hospital, so the model linking the two would be a straight line. Alternately, we generally expect body mass to increase with the square of a person's height, so the model linking these would be in the form of a quadratic. Other more complicated interactions (such as prevalence of HIV/AIDS in relation to a country's GNP, literacy rate, and prevalent religion) are likely to require more complicated models.

The next step is to associate a probability distribution function with the data. To explain, consider that even if two individuals have the exact same predictor variables, we would not expect the response variable to be exactly the same. Thus we assume that the response variable contains some degree of randomness. In associating a probability distribution with the data, we are quantifying how to predict the behaviour of this randomness. The choice of probability distribution function will depend on the nature of the data collected. Questions such as, "is the measured response discrete or continuous?" and "do we expect a skewing of the probabilities or that the measured responses will be evenly distributed about its average?" should be considered in the decision of which probability distribution to use.

Information on various probability distributions can be found in Chapter 5, Subsection 3.3.

Once the model and probability distribution have been created, the model is fit to data. This involves using information from the data to estimate any unknown parameters in the mathematical equation. If a collection of parameters can be found that create a good fit of the equation to the data, then the model is accepted and can be used to make predictions. Otherwise the researcher should return to the beginning, and create a different model to try to fit the data.

2. Common Uses

Regression is the collection of statistical tools that can be used to relate predictor variables to response variables. In healthcare, many of the questions approached by these techniques involve trying to understand what influence (positive or negative) various predictor variables have on long-term health. As such, these techniques are very useful in answering questions regarding the cost of healthcare and healthcare demand, such as:

- *How does age impact the expected cost of cancer treatment?*
- *What is the role of health insurance on the demand for health services?*
- *How does the physician to population ratio affect public access to healthcare?*

Often, the ultimate goal of regression analysis when applied to healthcare is to characterize the incentive structures underlying observed patterns, test the effects of incentive-altering policy, and to estimate future demand. Regression analysis can be used to help develop an equation that relates the risk factors to the outcome of interest. In these equations the input variables should be known (or testable) properties of the patient (e.g. gender, age, etc.) and the output variables should be the expected value of the health outcome. For example we might try to create equations relating:

- *Number of cigarettes smoked per week* to *likelihood of contracting cancer*,
- *Age and education* to *likelihood of attending an immunization clinic*, or
- *Initial sense of pain and age* to *recovery rates after knee surgery.*

3. Mathematical Details

Suppose that an experiment (or series of experiments) has been performed and a collection of data has been developed. From the raw data we seek to develop formulae which can be used to predict what impact a change in the information will have on the probability of an event. (From this point on we will call the outcome we are seeking to predict the *response variable*, and information which we are using to make the prediction the *predictor variables*.)

The response variable *is the outcome we are trying to predict. The* predictor variables *are the information we are using to make the prediction.*

Analysis begins by creating a model function under which we feel the data is likely to fit. When the models are based in economic theory, literature often refers to the results as *Econometrics* (meaning economic measurement). However, other approaches of developing model functions exist. After creating a hypothetical model, we use the statistical tool of regression analysis to fit the data to the model. If a good fit can be created then this helps validate the model, if no fit can be found then the model is rejected and the process begun again.

Let us consider the question of how to select a model function to attempt to fit the data to. If the data is simple enough, one of the best ways to begin this process is to plot the data points on a graph. To see this, let us examine a simple example.

A young student wishes to know how much her car is costing her to drive. In order to develop an answer to this, she records her mileage car related expenses (gas, insurance, maintenance, etc.) for several months. This hypothetical data is shown in Table 1. Examining Table 1 it is clear that the further she drives, the more car related expenses she incurs. Plotting the four data points on a graph she notices they look roughly like a straight line. With a little bit of playing she

month	Jan.	Feb.	Mar.	Apr.	May
car expenses ($)	230	210	220	250	250
mileage (km)	345	304	309	430	450

TABLE 1. Car related costs and mileage by month.

determines that the formula

$$cost/month = 140 + mileage \times 0.25$$

gives a reasonable approximation of these numbers (see Figure 1).

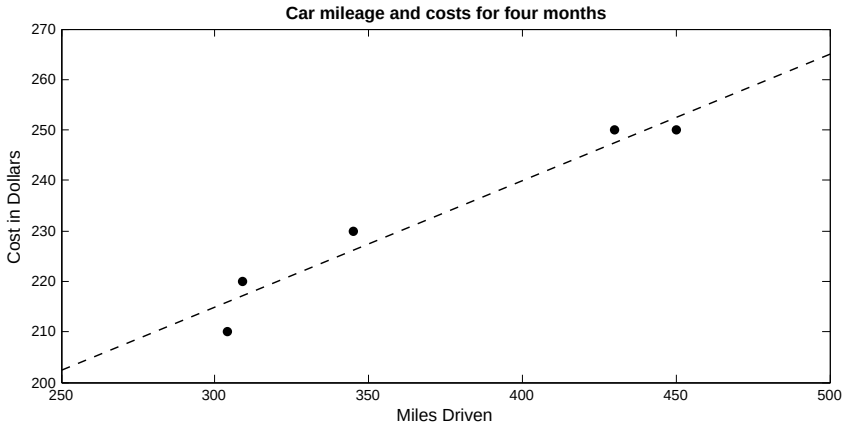

FIGURE 1. **Linear regression example:** Car related costs and mileage for four months. The dots are data points and the dashed line is the approximated linear regression.

Of course in most cases we cannot expect the relationship between two variables to be linear. For example, it is commonly accepted that the relationship between height and weight is nonlinear as weight increases with the square of an individual's height.

If the data is more complicated, or in particular if more than two variables are being considered, then it is likely that graphing the data to help guess at the relationship is impossible. In these cases, we can rely on economic theory to help select which style of equation might be best suited for the model. A complete survey of all possible model types is well beyond the scope of this book. Instead we present three important styles that are well suited to analysis in healthcare: normal linear regression, logistic regression, and generalized linear regression.

Normal Linear Regression: In *normal linear regression* we assume the relationship between the data is polynomial and that any errors in the data are distributed in a normal manner (i.e. along a bell curve). This is well suited for many physical phenomena such as the relationship of height to weight, or the relationship of the cost of maintenance to the size of a hospital. More details on normal linear regression are given in Subsection 3.1.

Logistic Regression: In *logistic regression* we assume that the relationship follows an 'S' shaped curve called the logistic curve. This implies that the response

variable is impacted less by changes in the predictor variables when the predictor variables are near the extreme ends of their range. This style of curve is well suited to many health related issues such as the relationship between recovered health and time since surgery. Also, whenever the data has a response variable that can only take on one of two possible values, the common consensus among statisticians is to use logistic regression. For example, logistic regression should be used if you are considering the relationship between age and the risk of a heart attack (since in the data each individual has either had a heart attack or has not). More details on logistic regression are given in Subsection 3.2.

Generalized Linear Regression: *Generalized linear regression* is a general style of regression analysis that includes both linear and logistic regression. The mathematics behind generalized linear regression is quite deep, so only a cursory overview of the subject is provided (Subsection 3.3).

3.1. Normal Linear Regression and Least Squares. In normal linear regression, the response variable is assumed to be a *polynomial* function of the set of predictor variables. That is, the relationship can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{3}$$

where X_i are the predictor variables, Y is the response variable, and β_i ($i = 1, 2, \ldots, n$) are fixed coefficients used to describe the linear relationship between X_i and Y. Despite appearances, it is important to note here that the term *linear* regression refers to the linearity in the coefficients β_i and not in the predictor variables X_i ($i = 1, 2, \ldots, n$). Indeed in many cases we do not wish to assume a linear relationship between the response variable and the predictor variables. (Recall for example, mass is generally considered proportional to the square of an individual's height.) How to deal with nonlinearity in the predictor variables is easily explained by an example. Suppose the response variable Y depends on predictor variables X_1 and X_2 in a relationship of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 / X_2^2.$$

Defining two new variables as $X_3 = X_1^2$ and $X_4 = X_1/X_2^2$, the above formula becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4,$$

which is linear in its predictor variables. One way to remember this is to remember that height, mass, and body mass index ($BMI = mass/height^2 = kg/m^2$) can all be predictor variables.

The word "linear" in linear regression refers to the coefficients β_i and not to the predictor variables X_i. A regression of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$, can easily be made linear in X_1 by writing $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where $X_2 = X_1^2$.

In the above example, the response variable was the cost of driving the car for a given month, and the predictor variable was the distance the car was driven in that month. In healthcare, the response variable may be the cost of treating a patient, the likelihood of an individual experiencing a given disease, the likelihood of an individual using the healthcare system, or any number of other things. Explanatory variables may include things like an individual's sex, race, body mass index, etc.

The difficultly in developing a linear regression largely reduces to how to determine the coefficients β_i ($i = 1, 2, \ldots, n$). To do this we begin by making multiple independent observations of the response variable. If the predictor variables are something we can control, this can be done in the form of experiments that ensure good distributions of the predictor variables and high accuracy in the response variable. However, in healthcare it is not common to have control over the predictor

variables, and so observations are often made via surveys which may not give a good distribution of the predictor variables or high accuracy in the response variable.

Suppose N surveys have been performed that provide us with N independent observations of the response variable and N distributions of the predictor variables:

$$\begin{array}{lll} \text{Survey 1} & \text{response} = Y_1 & \text{predictor variables} = (X_{1,1}, X_{2,1}, \ldots, X_{n,1}) \\ \text{Survey 2} & \text{response} = Y_2 & \text{predictor variables} = (X_{1,2}, X_{2,2}, \ldots, X_{n,2}) \\ \vdots & \vdots & \vdots \\ \text{Survey } N & \text{response} = Y_N & \text{predictor variables} = (X_{1,N}, X_{2,N}, \ldots, X_{n,N}). \end{array}$$

Since data collected naturally contains some randomness, we do not expect to be able to find coefficients β_i $(i = 1, 2, \ldots, n)$ such that

$$Y_j = \beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,j} + \cdots + \beta_n X_{n,j} \text{ for all } j = 1, 2, \ldots, N.$$

The normal distribution (also known as the Gaussian distribution) is the distribution associated with the classical "bell curve". Some of its key characteristics include, a symmetric unbounded distribution about its mean with a higher likelihood of being near the mean.

Instead we view each Y_j as a realization of a random variable Y that depends on the predictor variables $(X_1, X_2, \ldots, X_n)$ and a random factor ε. Our equation becomes

$$Y_j = \beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,j} + \cdots + \beta_n X_{n,j} + \varepsilon_j \text{ for all } j = 1, 2, \ldots, N.$$

and we attempt to find the coefficients β_i $(i = 1, 2, \ldots, n)$ that provide the best fit to the observed data (that is the smallest values for ε_j). For *normal linear regression*, we assume that the random variable Y takes on a *normal distribution* and then use ordinary least squares.

In the *ordinary least squares* estimation procedure, the unknown β coefficients are estimated by minimizing the sum of the squared differences between the observed responses Y_i and the potential linear approximation:

$$\begin{aligned} &\min_{\hat{\beta_1}, \hat{\beta_2}, \ldots, \hat{\beta_n}} \left\{ \sum_{j=1}^{N} \left(Y_j - [\hat{\beta}_1 X_{1,j} + \hat{\beta}_2 X_{2,j} + \cdots + \hat{\beta}_n X_{n,j}] \right)^2 \right\} \\ &= \min_{\hat{\beta_1}, \hat{\beta_2}, \ldots, \hat{\beta_n}} \sum_{j=1}^{N} \varepsilon_j^2. \end{aligned}$$

Taking squares of the differences between observed and fitted responses prevents positive and negative deviations from the line cancelling each other when summing the errors. An example of normal linear regression and least squares will be given in Subsection 4.1

The central limit theorem *states that that the sum of the random variables with finite variance tends towards a normal distribution as the random variables go to infinity.*

At this point it is worth noting an important result from statistical theory: the Gauss-Markov theorem. The Gauss-Markov theorem states that if the random variable is normally distributed then the best linear unbiased estimator for the coefficients is found via ordinary least-squares. That is, if a researcher is going to assume that the data fits a normal linear regression scheme, then ordinary least squares should be used to determine the coefficients β_i $(i = 1, 2, \ldots, n)$.

On a final note, the ordinary least square problem is a quadratic optimization problem (see Chapter 16). It, along with normal linear regression, can be accomplished by a number of optimization and statistical software packages (see Appendix A).

3.2. Logistic Regression. In logistic regression, the response variable is assumed to follow what is called the *logistic curve*. This curve is defined as

$$Y = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \tag{4}$$

where, as before, X_i are the predictor variables, Y is the response variable, and β_i $(i = 1, 2, \ldots, n)$ are fixed coefficients used to describe the linear relationship between X_i and Y. As in linear regression, it is important to note that we can easily incorporate non-linearity in the predictor variables (X_i) in this model. For example, if we desire the predictor variable X_1 to be cubed in the exponent, we can easily create a new variable $X_2 = X_1^3$:

$$\text{i.e. } Y = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_1^3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_1^3}} = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}.$$

In order to compute the coefficients β_i $(i = 1, 2, \ldots, n)$ for the logistic regression,

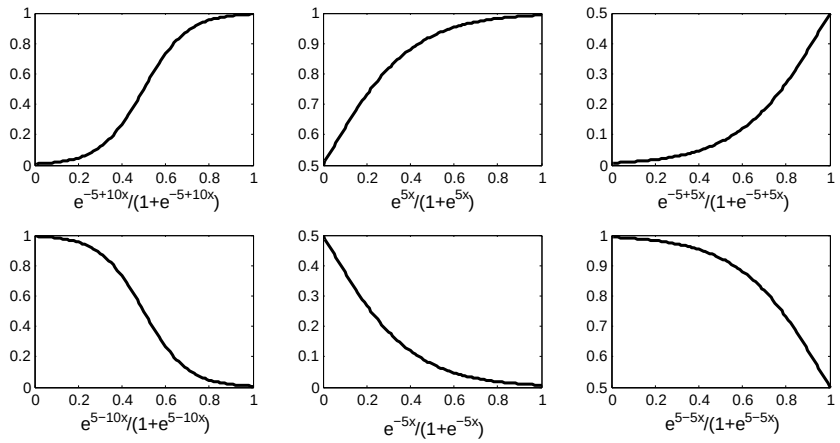

FIGURE 2. **Examples of logistic curves:** Logistic curves can take the shape of an 'S' *(top left)*, the shape of an 'inverted S' *(bottom left)*, or a portion of one of these shapes *(right)*

we begin by considering the function

$$P(x) = \frac{e^x}{1+e^x}.$$

Notice that

$$\begin{aligned}
\frac{P(x)}{1-P(x)} &= \left(\frac{e^x}{1+e^x}\right) \div \left(1 - \frac{e^x}{1+e^x}\right) \\
&= \left(\frac{e^x}{1+e^x}\right) \div \left(\frac{1+e^x}{1+e^x} - \frac{e^x}{1+e^x}\right) \\
&= \left(\frac{e^x}{1+e^x}\right) \div \left(\frac{1}{1+e^x}\right) \\
&= e^x.
\end{aligned}$$

Therefore $\log\left(\frac{P(x)}{1-P(x)}\right) = x$, and in particular, equation (4) tells us

$$\log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

At this point it appears that, if we replace the response variable Y with a new response variable $\widehat{Y} = \log\left(\frac{Y}{1-Y}\right)$ then we have effectively reduced the logistic regression to a linear regression. Unfortunately, this replacement ruins any chance we had of the error terms being normally distributed, and therefore we cannot apply ordinary least squares to determine the coefficients. Instead, something called *maximum likelihood estimation* must be used. The good news is most statistical software packages are capable of performing this estimation.

3.3. Generalized Linear Regression. In both the normal linear regression and logistic regression discussed above we make several assumptions on the relationship between the predictor variables and the response variable that may not be valid. Most importantly, in normal linear regression we assumed that errors in the data collected for the response variable are normally distributed. In many cases this assumption is unreasonable.

For example, suppose we are trying to predict people's family planning choices, specifically how many children families will have, as a function of income and various other socioeconomic indicators. The response variable (number of children) cannot be normally distributed, since it is bounded below (a family can never have less than 0 children) and there is a skewed likelihood towards a smaller number of children. In this case, it would be more reasonable to assume that the dependent variable follows a *Poisson distribution* (see Chapter 5, Subsection 3.3).

In order to deal with regression analysis for non-normal distributions, we usually turn to *generalized linear models*. The full mathematics of generalized linear models is beyond the scope of this book, but it is worth noting some of its features.

Most importantly, generalized linear models can be used when response variables follow distributions other than the normal distribution. More specifically, generalized linear models allow for regression analysis of response variables that follow any probability distribution in the exponential family of distributions. These include (but are not limited to) the normal, binomial, Poisson, and gamma distributions (see Chapter 5, Subsection 3.3).

More detailed information on various probability distributions can be found in Chapter 5.

In generalized linear models we replace equation (3) with

$$f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n. \tag{5}$$

Notice the only change is the replacement of the response variable Y with a function of the response variable $f(Y)$. This function (called the *link function*) is what allows for the great diversity in applying generalized linear models.

With the relaxation of the assumptions of a normally distributed response and homogeneous variance, the ordinary least squares estimation procedure is no longer appropriate. Instead, the *maximum likelihood estimation* as mentioned in the previous subsection, must be used. Most statistical software packages can perform this estimation.

An interesting final note is that if the link function f is the identity function (that is $f(Y) = Y$) then the generalized linear model reduces to normal linear regression, while if f is the logit function (i.e. $f(Y) = \text{logit}(Y) = \log(Y/(1-Y))$) then the generalized linear model reduces to logistic regression.

4. Examples

4.1. An Artificial Regression between Work Environment and Toe Stubbing. For the sake of example let us suppose that a researcher has developed a hypothesis that the number of stairs in an office has an impact on the frequency of office workers stubbing their toe[3]. In order to test this hypothesis he contacts 618 office workers and asks the employees to fill out a simple two question survey:

(1) How many stairs do you walk up/down to get to your office? (not staircases, but total stair count)
(2) Did you stub your toe last month?

He collects the data in Table 2.

# of stairs category	0	1-3	4-6	7-9	10-12	13-15	16+
# of people in category	193	85	72	113	71	63	21
# of toes stubbed in category	1	1	2	6	5	5	2
% of toes stubbed in category	0.52	1.18	2.78	5.31	7.04	7.93	9.52

TABLE 2. Artificial Data of Toes Stubbed in the Office.

Examining his table, he feels that the hypothesis was correct. In order to confirm this, and to apply his research, he decides to develop a model that would predict at what point (in regards to toe stubbing) an office should invest in an elevator.

He begins by plotting the midpoints of the number of stairs, against the percentage of toes stubbed, *i.e.* the points $[0, 0.52]$, $[2, 1.18]$, $[5, 2.78]$, $[8, 5.31]$, $[11, 7.04]$, and $[14, 7.93]$ on a graph (see Figure 3). Notice he omits the data regarding people with more than 15 stairs on their way to work. This is due to low data size, and the fact the interval is unbounded so no midpoint can be selected.

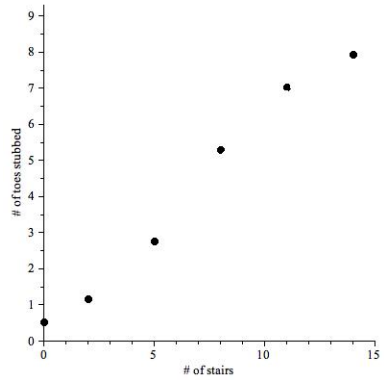

FIGURE 3. **Toes stubbed by number of stairs:** Percentage of toes stubbed by number of stairs to get to the office.

[3]All numbers for this example are made up. To the best of our knowledge no study has every compared work environment to toe stubbing.

Examining Figure 3 he hypothesizes that a good fit might be found via linear regression. Therefore he applies a linear regression model and proposes the line

$$y = \beta_0 + \beta_1 x \tag{6}$$

where y is the response variable representing the number of toes stubbed, x is the predictor variable of the number stairs, and β_0, β_1 are unknown coefficients. From here, he asks his favourite statistics software package to fit the coefficients of equation (6) to the points $[0, 1.04]$, $[2, 1.18]$, $[5, 2.78]$, $[8, 5.31]$, $[11, 7.04]$, and $[14, 7.93]$. The software package tells him that the $\beta_0 = 0.3$ and $\beta_1 = 0.6$. Thus he uses the line

$$y = 0.3 + 0.6x$$

to provide a goodness of fit to his data (see Figure 4).

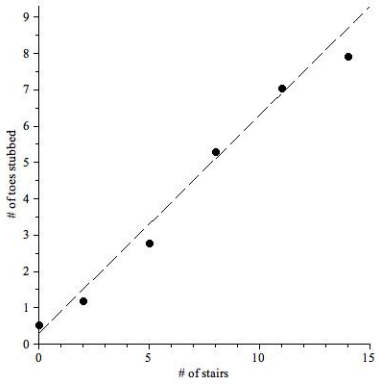

FIGURE 4. **Linear fit to toes stubbed by number of stairs:** Number of toes stubbed by number of stairs to get to the office fit with a line.

On further observation he notes that first, a linear regression is unlikely, as any linear regression would result in a 100% toe stubbing rate once the number of stairs became sufficiently high; second, the measured response variable ("Did you stub your toe") is binary in that it can only take one of two values. This means that applying a normal distribution function to the outcome is illogical.

Then the researcher proposes the logistic curve

$$P = A\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}\,\% \tag{7}$$

where P is the response variable representing the probability of stubbing your toe, x is the predictor variable of the number stairs, and β_0, β_1 are unknown coefficients. The coefficient A represents the maximal proportion of the population that will stub their toe in any given month. He asks his statistics software package to fit the coefficients of equation (7) to the data and gets $A = 9$, $\beta_0 = -3$ and $\beta_1 = 0.4$. These values provide a goodness of fit of over 90%, which pleases him immensely (details of goodness of fit are discussed in Chapter 5). He may now use the equation

$$P = 9\frac{e^{-3+0.4x}}{1 + e^{-3+0.4x}}\,\%$$

to predict the probability that an individual will stub their toe in a given month, given that the individual climbs/descends x stairs on the way to work. In particular his results suggest that there is never more than a 9% chance of stubbing your toe in any given month, regardless of the number of stairs in the office.

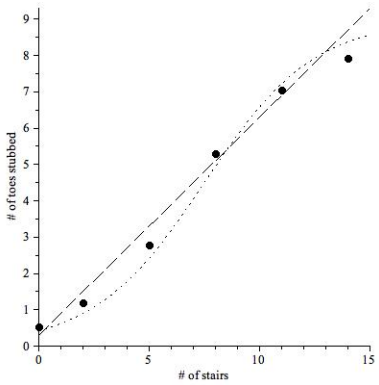

FIGURE 5. **Logistic fit to toes stubbed by number of stairs:** Percentage of toes stubbed by number of stairs to get to the office fit with a logistic curve.

4.2. Work, Wages, and Men's Health. The correlations between worktime, wages and health have been repeatedly addressed in literature. One example is due to Haveman, Wolfe, Kreider and Stone in 1994 [**103**]. In this example we review this study and discuss the results therein.

The data for the study was taken from the Michigan Panel Study of Income Dynamics, which consisted of following 613 white males from 1976 to 1983 and recording annually their job, income, number of hours they work each week, and various personal characteristics. In addition to this the participants were asked a series of questions regarding their health status. Specifically, each respondent was first asked, "Do you have a physical or nervous condition that limits the type of work or the amount of work you can do?" If the answer was "yes", then a followup question was asked: "Does it limit your work a lot, somewhat, or just a little." Based on these questions a score of 0 (for no condition) to 3 (for a condition which limits a lot) was given to each respondent for each year.

This data was fit to the system of linear equations jointly indexed in individual (i) and time (t) below (equations (8)).

$$\begin{aligned} H_{it} &= \beta_0 + \beta_1 W_{it} + \beta_2 P^H_{it} + \beta_3 O^H_{it} + e_{it} \\ W_{it} &= \alpha_0 + \alpha_1 H_{it-1} + \alpha_2 R_{it} + \alpha_3 P^W_{it} + \alpha_4 O^W_{it} + u_{it} \\ R_{it} &= \pi_0 + \pi_1 H_{it-1} + \pi_2 W_{it-1} + \pi_3 P^R_{it} + v_{it}, \end{aligned} \tag{8}$$

In these equations H is health status, W is hours worked, R is wages, P^H is personal characteristics that determine health, P^W is personal characteristics contributing to work time, and P^R is personal characteristics contributing to wages. O^H is job characteristics that determine health and O^W is job characteristics that

determine hours worked. Additionally, e, u, and v are error terms related to observed factors.

To fit the equations, a method called the generalized method of moments was used. This method is similar to ordinary least squares, but adjusts for having a system of equations where the error terms (e, u, and v) may have some correlation. This allows correlations between the quantities in the model to be estimated without introducing biases. (See [**100**] for more details.)

The conclusions of the study support the hypothesis that health limitations and age are positively correlated, while health limitations and education are negatively correlated. However, an expected positive correlation between prior work-time (that is, total number of years working) and health limitations was found to be absent. This is surprising, as the authors predicted that extended time in a hazardous work environment would lead to a greater risk of poor health. The study suggests that it is not how long you work in a given job, but the nature of the job that correlates with perceived health status.

4.3. Recovery Curves for Post Surgery Physiotherapy. As the population of Canada continues to age, Canada will see a growing demand for total hip arthroplasty (THA) and total knee arthroplasty (TKA). This will result in a growth in post-operation rehabilitation services. In order to meet this growth, it will be beneficial for Occupational Therapists and Physical Therapists to have an understanding of a patient's expected post-surgery recovery rate. Once developed these rates can be used to benchmark individual patient improvement, help estimate expected costs and lengths of therapy, and possibly help design optimal treatment session scheduling.

Recovery rates can be modelled as a function inputing days since surgery and outputting expected recovery status. The graph of such a function provide a visual explanation of recovery rates that can be used to discuss individual patient's recovery. We shall refer to such a graph as a *recovery curve*. In [**125**] [**126**] and [**102**], we find research discussing the form of recovery curves, and statistical analysis fitting the recovery curve form to real recovery data collected from various locations within Canada. This research provides a solid platform for research on recovery curves, and most importantly demonstrates that a *hierarchical linear model* can be used as a framework to develop recovery curves for post-THA and post-TKA patients. It is likely that such a model would fit recovery curves for other types of surgery. In this example we discuss this model, and analyze how regression analysis can be used to fit data to the model.

Research on patient recovery rates has suggested that recovery curves are most likely to take the form of a standard logistic curve,

$$E(d) = \frac{exp(\alpha + \beta d)}{1 + exp(\alpha + \beta d)} \tag{9}$$

where $E(d)$ represents the expected level of recovery d days after the surgery was performed and the coefficients α and β differ from patient to patient (as usual exp represent the exponential function $exp(x) = e^x \approx (2.71)^x$). The coefficient α represents the intercept (or starting point) of the recovery curve, while the coefficient β represents the growth rate of the recovery curve.

The Hierarchical portion of the model follows next. Since recovery rate varies from patient to patient, we hope to predict the coefficients α and β via a collection

of patient demographics. For example the coefficient α may be dependent on the patient's gender, age, and weight category. We shall let x represent a vector of patient demographics factors, and apply the assumption that α and β depend linearly on the vector x. That is

$$\begin{aligned} \alpha &= \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i \\ \beta &= \beta_0 + \sum_{i=1}^{N} \beta_i x_i, \end{aligned} \tag{10}$$

where x_i $(i = 1, 2, \ldots, N)$ are the predictive variables. This provides a two-level model for recovery, as $E(d)$ depends on d (level 1: post-surgery days), and on x (level 2: patient demographic factors).

Notice that under this model, the x_i predictive variables are all treated in the same manner. In particular, if a predictive variable is doubled in value, the effect on the model is doubled. For many predictive variables, such as age, this is not a concern. Basically, it enforces the assumption that as age increases, the impact of age on recovery time increases. However, for predictive variables which are categorical in nature, such as "type of walking aid used", this is a problem. For example, suppose we have three types of walking aids: no aid, walking cane, and wheel chair. If a single numerical variable is used to represent all three types of walking aids, then we would have to give each category a numerical representation. However, this introduces bias into the model, as such a numerical representation automatically assumes that one type of walking aid has a greater effect than another.

There are two common ways to correct this problem. The more mathematical method is to create a predictive variable for each category. For example, given our three types of walking aids, we create three predictive variables, say x_1, x_2 and x_3, representing

$$\begin{aligned} x_1 = 1 &\Rightarrow \text{no walking aid used,} \\ x_2 = 1 &\Rightarrow \text{walking cane used, and} \\ x_3 = 1 &\Rightarrow \text{wheel chair used.} \end{aligned}$$

Clearly, for any one patient exactly one of these variables is equal to 1 while the remaining two would be equal to 0.

The second, and somewhat easier method, is to use a statistical software package that allows the user to define categorical variables. Categorical variables are variables that have no intrinsic ordering between different variable values. Such variables are treated differently, with the software automatically applying the mathematical method above. If such software is available, then modellers should be careful to define variable types correctly.

The model created by equations (9) and (10) is what is referred to as a *hierarchical linear model*, or *multi-level model*. In this case the model consists of two levels, the first being the logistic regression curve for $E(d)$, and the second being the linear regression of α and β. In order to determine the best coefficients for this model we proceed by reducing it to a *single-level model*. To do this we construct N new predictive variables defined as $y_i = x_i \times d$ $(i = 1, 2, \ldots, N)$. Equation (9) can now be reduced to a single logistic regression defined by

$$E(d) = \frac{exp(\alpha_0 + \beta_0 d + \sum_{i=1}^{N} \alpha_i x_i + \sum_{i=1}^{N} \beta_i y_i)}{1 + exp(\alpha_0 + \beta_0 d + \sum_{i=1}^{N} \alpha_i x_i + \sum_{i=1}^{N} \beta_i y_i)}. \tag{11}$$

Thus the final logistic regression has up to $2N + 2$ undetermined coefficient. Standard statistical software can now be used to determine goodness-of-fit information for the single-level model, and determine which demographics factors are most important in patient recovery.

For the case of TKA and THA, specific data analysis can be found in [**125**] [**126**] and [**102**]. In particular, it was found that the single largest determining factor of recovery was time past surgery and the type of surgery performed. Gender and the availability of at-home physiotherapy played a small role in certain surgery types, but was not significant for most measures of recovery. In Figure 6 we display two sets of sample recovery curves. These examine the expected loss in post-surgery knee extension by post-operation date (POD). The dashed line (at 2 degree) is the aimed for level of recovery.

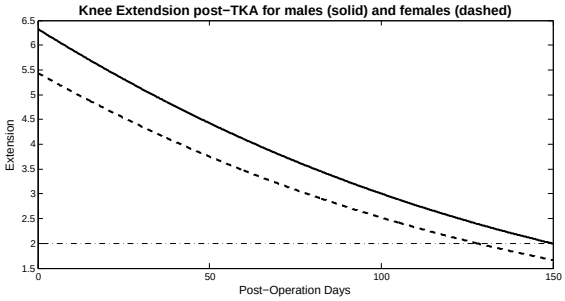

FIGURE 6. **Post-Surgery Knee Extension Recovery Curves:** The expected loss in post-surgery knee extension by post-operation date (POD). Notice that the expected recovery rate for males is slightly slower than females. However, the statistical significance of this result is minor.
Reproduced from [**102**].

5. Related Reading

Regression analysis has close connections to Descriptive Statistics (Chapter 5) as well as Epidemiological Risk Modelling (Chapter 7). Like most fields of statistics, regression analysis is useful in almost all forms of quantitative modelling.

Reference [**69**] discusses the generalized utility function, the indifference curve, and the contract curve; all of which are now used in economic theory. Reference [**32**] provides an overview of economic analysis, while reference [**66**] overviews applications of econometrics to healthcare. Reference [**207**] introduces single-equation linear regression analysis with an emphasis on real-world examples. References [**122**] and [**223**] examine the range of applied econometric work in the field of healthcare. References [**121**] and [**225**] provide an introduction to econometric techniques for use with different types of survey and count data.

Reference [**100**] studies estimators that make sample analogues of populations' orthogonality conditions close to zero. Reference [**105**] looks at minimizing the impact of assumptions in econometric models. Reference [**115**] presents some of the currently available and easily used methods for assessing the adequacy of a fitted logistic regression model.

Reference [**19**] examines a method of using econometrics to provide a unified framework for understanding human behaviour. Reference [**223**] provides a survey of British applied econometric work in the field of healthcare. Reference [**20**] uses econometrics to analyze social issues. Reference [**103**] investigates the complex relationships between worktime, wages and health using econometrics. Reference [**212**] reviews Australian research pertaining to socioeconomic health inequalities as well as Australia's research capacity for socioeconomic health inequalities and policies, along with interventions that have been suggested. Reference [**132**] provides an econometric analysis of veterans' healthcare utilization. Reference [**193**] provides an econometric analysis of healthcare utilization in Canada.

CHAPTER 7

Evaluating Detrimental Behaviour

The policy of being too cautious is the greatest risk of all. J. Nehru (1889 - 1964)

Life is a sexually transmitted disease and the mortality rate is one hundred percent. R. D. Laing (1927-1989)

Epidemiological Risk Modelling

1. Model Overview

A recent study by Hakes and Viscusi showed that the general public has a serious problem in evaluating risky behaviour. The study consisted of asking 493 adults a series of questions of the form "how many people do you think died in the U.S. due to _ in 1991?" The results were staggering. On average people believed that roughly the same number of deaths resulted from the measles as from accidental poisoning[1]. The authors suggested that the public's difficulty in distinguishing between differing magnitudes of risks has led to a similar spending for reducing each risk. The result is a gross misdistribution of healthcare budgets.

A similar effect results when examining the factors that create each risk. Every day healthcare policy makers must make decisions such as whether money would be better spent researching anticancer drugs or supporting antismoking campaigns. To make these decisions, policy makers rely heavily on the statistical methods of *epidemiological risk modelling*. (Before continuing we note that, *epidemiology* in general encompasses any research that examines factors affecting health. This includes many research directions outside the scope of this book. For example, biological studies of disease dynamics are considered epidemiological research.)

In epidemiological risk modelling one uses statistical methods to attempt to determine associations between a given factor (or factors) and a health outcome. Although the factor is generally referred to as a *risk factor*, it should not necessarily be viewed as negative. For example, the risk factor of "hand-washing" has been shown to reduce the spread of disease in hospitals. On the other hand, numerous studies have linked the risk factor of "smoking" to an increased chance of contracting lung cancer[2]. In the same manner it should be noted that, with a little creativity, we always phrase a given health outcome as either a positive or a negative outcome. For example, instead of examining what factors impact the chance of a wide spread bubonic plague (poor sewers, rat infestations, etc.), we could study what factors

[1] In 1991, the measles resulted in 5 deaths, while 5200 people died from accidental poisoning in the U.S.

[2] Perhaps the most famous study linking smoking to lung cancer is the 1964, the US Surgeon General report "Smoking and Health."

impact the chance of *avoiding* a wide spread bubonic plague (clean sewers, effective pest control, etc.). In general this confuses the issue with a large collection of double negatives, therefore, in order to make discussion easier, *we shall always assume that the health outcome is the undesirable outcome.* As such, we will refer to the health outcome as a *disease,* and call a risk factor *beneficial* if it reduces the likelihood of a given health outcome and *harmful* if it increases the likelihood of the outcome. It should be noted now that the word disease is used in the sense of a "harmful development," and therefore may be used to represent any negative health outcome. For example, under this terminology, we may view exercise as a beneficial risk factor for the disease of obesity.

In epidemiological risk modelling the term disease *is used to refer to any negative health. For example, drinking and driving can be viewed as a harmful risk factor for the "disease" of automobile accidents.*

To develop an epidemiological risk model, we begin collecting data regarding the risk factors and diseases to be examined.We then use the data to determine various statistical values that associate the risk factors to the disease. Some of the most popular of these values are the *relative risk*, *attributable risk*, *prevented fraction*, and *potential impact fraction.*

Before discussing these values, we will briefly turn our attention to another popular value called the *odds ratio.* We single out this popular value as a warning to policy makers. The odds ratio tells you the change in the odds of having a specified disease given that one has the risk factor. This is extremely similar to the relative risk (below) but very misleading in its implementation. The problem is, without knowing the original odds of having the disease, the change in odds is useless information. For example, telling someone that moving to Saskatchewan will triple their chance of being killed by a lightning strike[3] may sound impressive; however, since one's chance of being killed by lightning is less than 1 in 10,000,000 per year and the population of Saskatchewan is about 1,000,000, it still seems like a fairly safe place to live. As a general rule, anytime a researcher reports on odds ratio instead of one of the more standard measures of risk, the results should be strongly questioned.

Loosely speaking, the relative risk tells you the change in the probability of having the specified disease given that one has the risk factor. More specifically, relative risk provides the multiplicative factor relating the disease status of those with the risk factor to those without. For example, if the relative risk is 3, then an individual with the risk factor is three times more likely to experience the disease than an individual without the risk factor. Alternately, if the relative risk is 0.25 then an individual with the risk factor is four times less likely to experience the given disease than an individual without the risk factor. Thus, if the relative risk is greater than 1 then the risk factor is harmful, while if the relative risk is less than 1 then the risk factor is beneficial. Like the odds ratio, relative risk does not provide a base level of risk, so small changes to a low risk activity will result in the same relative risk as large changes in a high risk activity.

In some epidemiological literature relative risk is called the Risk Ratio. *However, the term risk ratio is more commonly used in a generic sense to refer to either relative risk or odds ratio.*

If the relative risk is greater than 1, one may wish to examine the attributable risk. Whenever the relative risk is greater than 1, the attributable risk lies somewhere between 0 and 1 and represents the proportion the given disease that could be (theoretically) reduced if the given risk factor were eliminated. For example, an attributable risk of 0.28 means that if the risk factor could be eliminated from the population, the prevalence of the disease would (theoretically) be reduced by 28%.

[3]This statement is made up for the sake of example, there is no statistical evidence that Saskatchewanites are in any greater danger of death by lightning.

Conversely, if the relative risk is less than 1, then we may wish to examine the prevented fraction. Whenever the relative risk is less than 1, the preventive fraction lies somewhere between 0 and 1 and represents the proportion the disease that could be (theoretically) reduced among the individuals that do not have the risk factor. For example, a prevented fraction of 0.37 means that 37% of individuals that do not currently have the risk factor would benefit from obtaining the risk factor.How to calculate these values will be explained in Section 3.

Regardless of whether the risk factor is beneficial or harmful, we may wish to examine the potential impact fraction. In the potential impact fraction, we assume that we are able to make an intervention that somehow changes the prevalence of the risk factor in the population. (If the risk factor is beneficial we attempt to increase its existence, if it is harmful we attempt to decrease it.) The potential impact fraction then measures the reduction that would result from a given change in prevalence. This is a considerably more complicated concept than that of relative risk, attributable risk, and prevented fraction, so we leave further discussion to Section 3 and refer the reader to Example 4.1 for a detailed example on these concepts.

On a final note, in simple risk models, the various statistical values that associate the risk factors to the disease may be computed analytically. In general, however, the risk factors and the impacts of intervention hold complicated interrelationships that make analytic calculations difficult. In these cases we often resort to computer based simulation to approximate the value of objects such as the potential impact fraction. An example of this is provided in Subsection 4.3 and more details on computer simulation can be found in Chapter 9.

2. Common Uses

The goal of epidemiological risk modelling in healthcare is to develop models of the relationship between risk factors and diseases, accidents or mortality. Therefore, in general, epidemiological risk modelling answers the question of "what (if any) is the relationship between risk factor X and health outcome Y?" For example, epidemiological risk modelling may be used to approach questions such as:

- *What is the relationship between smoking and lung cancer?*
- *What is the relationship between obesity and the likelihood of experiencing a heart attack?*
- *What is the relationship between childhood exercise programs and obesity?*

More advanced techniques in epidemiological risk modelling are also capable of theoretically examining the effect of interventions on the health of the population. This allows epidemiological risk modelling to approach questions such as:

- *What effect would a tax increase on tobacco products have on the amount of lung cancer in the population?*
- *Would a policy enforcing exercise programs in high-school have a significant impact on the number of heart attacks in young adults?*

3. Model Details

3.1. Relative Risk, Attributable Risk and the Prevented Fraction. We begin with the precise definitions of *relative risk*, *attributable risk*, and *prevented*

fraction. To do this, recall that the probability of an event E occurring is

$$\Pr(E) = \frac{\#\text{ of ways } E \text{ occurs}}{\#\text{ of possible outcomes}}.$$

We use the notation $\Pr(E|F)$ to represent the probablity that an event E occurs given factor F is present:

$$\Pr(E|F) = \frac{\#\text{ of ways } E \text{ occurs given } F \text{ is present}}{\#\text{ of possible outcomes where } F \text{ is present}} = \frac{|E \cap F|}{|F|}.$$

where $|F|$ is the number of elements contained in set F, and $|E \cap F|$ is the number of elements in **both** E and F.

Orally $\Pr(E|F)$ *is read as: "the probability of* E *given* F*," while* $\Pr(E|F^c)$ *is read as: "the probability of* E *given* F *complement."*

We use the notation F^c to represent the option that the factor F is not present (the c stands for complement, which is the mathematical word for "that which completes the set").

For example, suppose you at a party consisting of 40 males and 60 females. Suppose that 30 of the males are drinking beer, while 20 of the females are drinking beer. Then the probability that a randomly selected person is drinking beer is

$$\Pr(Beer) = \frac{\#\text{ of beer drinkers}}{\#\text{ of people}} = \frac{30+20}{40+60} = 0.5,$$

the probability of a randomly selected male drinking beer is

$$\Pr(Beer|Male) = \frac{|Beer \cap Male|}{|Male|} = \frac{\#\text{ of } \textbf{male} \text{ beer drinkers}}{\#\text{ of } \textbf{males}} = \frac{30}{40} = 0.75,$$

and the probability of a randomly selected female drinking beer is

$$\Pr(Beer|Male^c) = \frac{|Beer \cap Male^c|}{|Male^c|} = \frac{\#\text{ of } \textbf{not male} \text{ beer drinkers}}{\#\text{ of } \textbf{not males}} = \frac{20}{60} \approx 0.33.$$

We now formally define relative risk. The relative risk is the ratio of the probability of having the specified disease given the risk factor is present to the probability of having the specified disease without the risk factor. Mathematically we denote relative risk by RR and define it as

We usually denote the Disease *by* D *and the risk* Factor *by* F*.*

$$RR = \frac{\Pr(Disease|Risk\ Factor)}{\Pr(Disease|Risk\ Factor^c)} = \frac{\Pr(D|F)}{\Pr(D|F^c)}. \tag{12}$$

Since the probability of an event is always between 0 and 1, the relative risk is can be any number greater than zero. (If $\Pr(D|F) = 1$ i.e., $\Pr(D|F^c) = 0$, this means that the disease always occurs. If $\Pr(D|F) = 0$ then we never get the disease. Neither of these cases need modelling to be understood, so we assume that neither $\Pr(D|F)$ nor $\Pr(D|F^c)$ is 0).

To simplify discussion we assume that the researched health outcome is undesirable, and therefore refer to it as a disease*. As such,* harmful risk factors *increase the chance of the disease, and* beneficial risk factors *reduce the chance of the disease.*

It is not difficult to see that the risk factors of interest are those that result in a relative risk that is noticeably greater than or noticeably less than 1. If the relative risk is greater than 1 then the probability of the disease is increased in the presence of the risk factor, and therefore the risk factor is harmful. Conversely, if the relative risk is less than 1, then the probability of the disease is decreased in the presence of the risk factor, and therefore the risk factor is beneficial. As such, the relative risk provides a reference for the potential benefit or harm of a given risk factor. This helps dictate how future epidemiological risk modelling should proceed.

If the risk factor is harmful (i.e., $RR > 1$) then one would like to proceed by examining the proportion of the disease that is (theoretically) attributable to the risk

factor. This is generally done by calculating the *attributable risk*. Mathematically the attributable risk is denoted AR and defined by

$$AR = \frac{\Pr(D) - \Pr(D|F^c)}{\Pr(D)}. \tag{13}$$

In literature, relative risk is sometimes also referred to as the likelihood ratio.

A public health interpretation of attributable risk is that it is a measure of the extent that a disease is preventable, if the risk factor could be eliminated or reduced. As such, attributable risk should be thought of as a population based risk measure, whereas relative risk is more individual based. That is, attributable risk describes how the risk factor impacts the population as a whole, while relative risk describes how the risk factor alters an individuals chance of the given disease.

If the risk factor is beneficial (i.e., $RR < 1$) then one would like to examine the proportion the disease that could be (theoretically) reduced if the risk factor were universally present. This is done by calculating the *prevented fraction*. Mathematically the prevented fraction is denoted PF and defined by

$$PF = \frac{\Pr(D|F^c) - \Pr(D)}{\Pr(D|F^c)} \tag{14}$$

Like attributable risk, the prevented fraction should be thought of as a population based risk measure, instead of an individual based measure.

Other terms used for attributable risk include: excess risk, risk difference, etiological fraction, etiological proportion, attributable fraction (or proportion), and preventable fraction (or proportion).

To compare the attributable risk and prevented fraction notice that

$$1 - AR = \tfrac{\Pr(D|F^c)}{\Pr(D)} \quad \text{and} \quad 1 - PF = \tfrac{\Pr(D)}{\Pr(D|F^c)}, \tag{15}$$

provided $\Pr(D|F^c) \neq 0$. (Note that if $\Pr(D|F^c) \neq 0$ then $\Pr(D)$ cannot be 0, so $\frac{\Pr(D|F^c)}{\Pr(D)}$ is well defined.) Therefore the attributable risk and prevented fraction are related by the equation

$$(1 - AR)(1 - PF) = 1 \text{ whenever } \Pr(D|F^c) \neq 0. \tag{16}$$

This formula has several ramifications. To begin, it provides a means for statistical estimates of AR to be used to estimate PF, and vice versa. Secondly, since AR and PF are always less than one, it demonstrates that exactly one of these values will be between 0 and 1 while the other value will be negative. Reexamining the definitions of AR and PF we can see that if the relative risk is greater than 1 then $0 < AR < 1$ while if the relative risk is less than 1 then $0 < PF < 1$.

To explain this, let us consider the case where the relative risk is greater than 1. From the definition of relative risk, this implies that $\Pr(D|F) > \Pr(D|F^c)$. Also $\Pr(D)$ must lie between these two values: $\Pr(D|F) > \Pr(D) > \Pr(D|F^c)$. By subtracting $\Pr(D|F^c)$ from the rightmost inequality we see

$$\Pr(D) - \Pr(D|F^c) > 0.$$

But since $\Pr(D|F^c) > 0$ we have that

$$\Pr(D) > \Pr(D) - \Pr(D|F^c).$$

Therefore we can combine these two inequalities to see that

$$\Pr(D) > \Pr(D) - \Pr(D|F^c) > 0.$$

Dividing through by $\Pr(D)$ gives $0 < AR < 1$. Conversely, if the relative risk is less than 1 then $\Pr(D|F^c) > \Pr(D|F)$, and again $\Pr(D)$ must lie between these two values. A similar logic demonstrates

$$\Pr(D|F^c) > \Pr(D|F^c) - \Pr(D) > 0,$$

and therefore $0 < PF < 1$.

Relating the attributed risk to relative risk is more complicated, but also possible. Some classical formulae include

$$AR = \frac{\Pr(F)(RR-1)}{1+\Pr(F)(RR-1)} \text{ and } AR = \frac{\Pr(F|D)(RR-1)}{RR}.$$

Notice that these formulae for AR require not just knowledge of the relative risk, but also the probability of the risk factor occurring, $\Pr(F)$, or the probability of the risk factor occurring given that the health outcome occurred, $\Pr(F|D)$. Relating the prevented fraction to the relative risk can now be done via equation (16).

From the presentation given above it may appear that the calculation of relative risk, attributable risk, and the prevented fraction is straightforward. If the risk model developed only concerns itself with a single risk factor which is either present or not present, then estimating the key values $\Pr(D)$, $\Pr(D|F)$ and $\Pr(D|F^c)$ is a fairly straightforward task. However, in this case there is a high chance that there exists *confounding risk factors*, that is risk factors that are not accounted for in the model. On the other hand, if multiple risk factors are considered or the risk factor can take one of several levels then it becomes necessary to select a statistical model to estimate the key values $\Pr(D)$, $\Pr(D|F)$ and $\Pr(D|F^c)$. The choice of which statistical model to use is usually based on the type of data collected, and can result in biases in the end analysis. We refer readers to Chapter 5 for information on dealing with these issues.

3.2. The Potential Impact Fraction of an Intervention. Regardless of whether a risk factor is beneficial or harmful, one might be interested in what sort of impact various interventions might have on the disease. To provide a numeric comparison between various intervention strategies, the potential impact fraction for each intervention is often calculated. To do this, one assumes that the prevalence of the risk factor in the population is altered in some manner. This could result in either an increase or decrease in the prevalence of the risk factor, which further results in either an increase or decrease in the disease incidence. The *potential impact fraction* is then the change in the disease probability relative to the current disease probability. Mathematically the potential impact fraction is denoted by PIF and defined:

$$PIF = \frac{(\Pr(D) - \Pr^*(D))}{\Pr(D)} \tag{17}$$

where $\Pr^*(D)$ is the probability of the disease under the modified distribution of the risk factor. (The division by $\Pr(D)$ gives a sense in the size of the change relative to the how likely the event is to begin with.)

Another term used for the potential impact fraction is the generalized impact fraction.

Although the mathematical statement of the potential impact fraction may be simple, its calculation is not trivial. The main difficulty lies in the computation of $\Pr^*(D)$. Unlike $\Pr(D)$, the value $\Pr^*(D)$ is a prediction of how the change in the risk factor will effect the overall probability of the disease. To develop a practical manner of computing the PIF consider the assumption

$$\Pr(D|F) = \Pr^*(D|F) \text{ and } \Pr(D|F^c) = \Pr^*(D|F^c). \tag{18}$$

In words, assumption (18) states that the probability of experiencing the disease *given* that an individual has the risk factor, and the probability of experiencing the disease *given* that an individual lacks the risk factor are constant

regardless of the prevalence of the risk factor in the population. Noting that $\Pr(D) = \Pr(F)\Pr(D|F) + \Pr(F^c)\Pr(D|F^c)$ and applying Assumption 18) we find

$$\begin{aligned} PIF &= \tfrac{\Pr(D)-\Pr^*(D)}{\Pr(D)} \\ &= \tfrac{\Pr(F)\Pr(D|F)+\Pr(F^c)\Pr(D|F^c)}{\Pr(D)} - \tfrac{\Pr^*(F)\Pr^*(D|F)+\Pr^*(F^c)\Pr^*(D|F^c)}{\Pr(D)} \\ &= \tfrac{\Pr(F)\Pr(D|F)+\Pr(F^c)\Pr(D|F^c)}{\Pr(D)} - \tfrac{\Pr^*(F)\Pr(D|F)+\Pr^*(F^c)\Pr(D|F^c)}{\Pr(D)} \\ &= \tfrac{\Pr(D|F)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) + \tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F^c) - \Pr^*(F^c)\right) \\ &= \tfrac{\Pr(D|F^c)}{\Pr(D|F^c)}\tfrac{\Pr(D|F)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) + \tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F^c) - \Pr^*(F^c)\right). \end{aligned}$$

In the last step we multiplied the first term by $Pr(D|F^c)/Pr(D|F^c) = 1$, which did not alter the expression. Next we recall that the definition of relative risk (equation (12)) and the fact that $\Pr(F^c) = 1 - \Pr(F)$ to simplify this to

$$\begin{aligned} PIF &= RR\,\tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) + \tfrac{\Pr(D|F^c)}{\Pr(D)}\left((1 - \Pr(F)) - (1 - \Pr^*(F))\right) \\ &= RR\,\tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) + \tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr^*(F) - \Pr(F)\right) \\ &= RR\,\tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) - \tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right) \\ &= (RR - 1)\,\tfrac{\Pr(D|F^c)}{\Pr(D)}\left(\Pr(F) - \Pr^*(F)\right). \end{aligned}$$

Finally, applying Equation (15) we conclude (under Assumption (18))

$$PIF = (RR - 1)(1 - AR)\left(\Pr(F) - \Pr^*(F)\right). \tag{19}$$

A similar approach can be done to relate PIF to the prevented fraction, yielding the final equation

$$PIF = (1 - RR)\left(\tfrac{1}{1-PF}\right)\left(\Pr^*(F) - \Pr(F)\right). \tag{20}$$

On the surface Assumption (18) may seem very reasonable. After all, why would the prevalence of the risk factor in the population alter how the risk factor impacts the given disease. However, consider the risk factor of smoking and the health outcome of lung cancer. Should the impact of smoking on your chance of developing lung cancer change based on how much of the population smokes? On the surface the answer appears to be no, but if we look deeper we see that the answer is yes. The reason lies in the concept of secondhand smoke.

Consider a hypothetical town where 90% of the inhabitants smoke. Suppose now that an intervention is performed that (somewhat miraculously) changes the town dynamics so only 10% of the population smokes. Let $\Pr(L|S)$ be the probability before the intervention of an individual contracting lung cancer given that the individual smokes and $\Pr^*(L|S)$ be the same after the intervention. The question of whether $\Pr(L|S^c) = \Pr^*(L|S^c)$ is therefore:

> *Does a nonsmoker in a town of 90% smokers have the same probability of contracting lung cancer as an individual in a town of 10% smokers?*

Given this extreme example, the answer is clearly no: the non-smoker in a town of 90% smokers is assaulted with nine times the amount of secondhand smoke as the individual in a town of 10% smokers.

Our example here of smoking and lung cancer may seem a little contrived. However, the truth is that, in general, the risk factors and the impacts of intervention hold complicated interrelationships that make Assumption (18) false. In these cases it is often easiest to resort to computer based simulations to approximate the

value of $\Pr^*$ under given interventions. An example of this is given in Example 4.3 and more information on computer simulations can be found in Chapter 9.

4. Examples

4.1. An Artificial Comparison between Chocolate Consumption and Chickenpox. For the sake of example let us suppose that a researcher has developed a hypothesis that the consumption of chocolate helps speed recovery from the chickenpox[4]. To test this she successfully contacts 3429 parents of children who had the chickenpox within the last year. Each of these parents fills out a survey stating how many days their child took to recover from the disease, and whether their child ate any chocolate during that time. Letting D represent the "disease" of spending 4 or more days sick from the chickenpox and F represent the factor of eating chocolate, she compiles the following table:

		Consumed Chocolate	
		yes (F)	no (F^c)
Days spent sick	≥ 4 days (D)	749	463
	< 4 days (D^c)	1515	702

TABLE 1. Artificial data table relating chocolate consumption to chickenpox recovery time.

The above data suggests that

$$\begin{aligned} \Pr(D) &= \tfrac{749+463}{3429} &&\approx 0.3535, \\ \Pr(D|F) &= \tfrac{749}{749+1515} &&\approx 0.3308, \quad \text{and} \\ \Pr(D|F^c) &= \tfrac{463}{463+702} &&\approx 0.3974. \end{aligned}$$

Dividing $\Pr(D|F^c)$ by $\Pr(D|F)$ we see that the relative risk is $RR = 0.8324$ which is less than 1, therefore Chocolate appears to be a beneficial risk factor. Since the risk factor is beneficial we also compute the prevented fraction:

$$PF = \frac{\Pr(D|F^c) - \Pr(D)}{\Pr(D|F^c)} \approx \frac{0.3974 - 0.3535}{0.3974} \approx 0.1106.$$

This suggests that 11% of the non-chocolate-eaters would benefit from eating chocolate. To see where this number relates, consider the 1165 children who did not consume chocolate during their sickness. The survey stated that 463 of these children took four or more days to recover. If they had all consumed chocolate, we would expect only 33.08% of the above 1165 children to take four or more days to recover, this corresponds to 385 children. Thus, $463 - 385 = 78$ children would have benefitted from eating chocolate; this is an improvement of $\frac{78}{702}100\% = 11.1\%$.

Finally, the researcher explores the potential impact of making chocolate freely available to any child with the chickenpox. To do this she notes that currently only $\frac{749+1515}{3429}100\% = 66.03\%$ of children consumed chocolate while they had the chickenpox. If chocolate were freely available for children with the chickenpox,

[4]All numbers for this example are made up. To the best of our knowledge no study has ever compared chocolate consumption and chickenpox recovery rates.

she feels this would increase to 98%. Thus the potential impact fraction for this intervention would be

$$\begin{aligned} PIF &= (1 - RR)\left(\tfrac{1}{1-PF}\right)(\Pr^*(F) - \Pr(F)) \\ &\approx (1 - 0.8324)\left(\tfrac{1}{1-0.1105}\right)(0.9800 - 0.6603) \approx 0.0602. \end{aligned}$$

This suggests that if chocolate were freely available to children with the chickenpox, the number of children who require four or more recovery days would decrease by approximately 6%. Considering our example, if 98% of the 3429 children ate chocolate during their illness, our new data set would have $(0.98)3429 = 3360$ children who ate chocolate during their illness and 69 who did not. Assuming $\Pr(D|F)$ and $\Pr(D|F^c)$ do not change this would result in $3360\Pr(D|F) + 69\Pr(D|F^c) = 1139$ children taking four or more days to recover and 2290 children recovering in less than 4 days. Previously we had 1212 require four or more days to recover, thus we see a decrease of $\frac{1212-1139}{1212}100\% = 6.0\%$.

Of course, the above study is too naive to be persuasive. For example, the confounding factor of family income is ignored (a higher family income is likely to impact both recovery time and the amount of chocolate consumed). Also, the analysis appears to arbitrarily select 4 days as the cutpoint for recovery time. This suggests some level of data analysis bias has been employed in the creation of the results (see Example 4.2). It would be much more logical to group people into more than two categories with regards to chocolate consumption, and consider recovery time as a continuous variable. Finally, the data is completely fabricated, so no results are valid anyways. Nonetheless, the above example does provide a simple demonstration of the prevented fraction, the potential impact factor, and their interpretations.

4.2. Publication Bias In Situ, and Overfitting the Model. Recent work of Phillips suggests that in reporting the statistic results of data collection and the probabilities they imply, healthcare research suffers from a variety of problems that are not understood by policy makers, individuals trying to make health decisions, or in some cases epidemiological researchers [**176**] [**177**]. Two examples of this are *publication bias* and *publication bias in situ*. *Publication bias* is a well-understood concept that usually refers to the fact that studies that produce "interesting" results (i.e. show a greater association, are statistically significant, and are in the "right" direction) are more likely to generate journal articles, while those that do not are more likely to end up in a file drawer. *Publication bias in situ* is a less understood concept, that results from researchers allowing the data to unduly influence the model design [**177**]. In this example we discuss publication bias in situ, and provide a small demonstration of how easy it is for a researcher to fall into the trap of allowing data to unduly influence the model design.

Given that every study and data set is a bit different, there is no standard way to analyze epidemiologic data, and therefore there are many choices to be made about the statistical model. In this example, we suppose that a data set has been constructed that consists of a collection of patients, a level of *exposure* to some risk factor, and a disease outcome (the patient either tests positive for the disease or negative for the disease). One method to analyze such a data set would be to *dichotomize* the exposure variable. That is, select a *cutpoint* such that everyone whose exposure level is above the cutpoint is considered "exposed" and everyone

whose exposure level is below the cutpoint is considered "unexposed". The choice of the cutpoint will clearly influence the estimated effects of exposure that are reported as the result of the study. For example, the cutpoint for an exposure might be far lower than the exposure level that actually affects health, so the *exposed* category actually includes a lot of people who have no elevated risk, reducing the apparent risk. Similarly, a cutpoint that is set too high will include many people in the *unexposed* category who are experiencing the effects of the exposure, elevating the baseline risk and thus reducing the apparent relative risk.

This suggests that a modeller might allow the data to select the cutpoint, by locating the cutpoint which provides the highest apparent risk from the data set. Although this may seem benign, it runs the risk that it will exaggerate what the data actually shows. This practice is sometimes known by the technical term *overfitting the model*.

Suppose that a data sample consisting of N participants has been collected. We label these participants $1, 2, \ldots, i, \ldots, N$. Further suppose that for each participant the data consists of a level of exposure to a given risk factor $e_i \in (0, 1)$ $(i = 1, 2, \ldots, N)$, and a marker indicating whether the patient tested positive for a specific disease $d_i \in \{0, 1\}$ $(i = 1, 2, \ldots, N)$. Immediately we categorize each patient i into two bins

$$D = \{i : d_i = 1\} \text{ and } D^c = \{i : d_i = 0\}$$

which categorize whether the patient tested positive for the disease. Next, for a given *cutpoint* k we separate the population into two other bins: exposed and unexposed. Specifically, we shall say participant i is exposed if $e_i > k$, and unexposed otherwise, which creates the two sets:

$$E = \{i : e_i > k\} \text{ and } E^c = \{i : e_i \leq k\}.$$

Our interest now lies in how these four bins interact. Specifically notice that we have created four categories of people:

- $E \cap D$: *E*xposed to the risk and contracted the *D*isease $(e_i > k, d_i = 1)$,
- $E^c \cap D$: Not *E*xposed to the risk and contracted the *D*isease $(e_i \leq k, d_i = 1)$,
- $E \cap D^c$: *E*xposed to the risk and did not contract the *D*isease $(e_i > k, d_i = 0)$,
- $E^c \cap D^c$: Not *E*xposed to the risk and did not contract the *D*isease $(e_i \leq k, d_i = 0)$.

By determining the number of individuals in each category we can calculate the risk ratio associated with the risk factor and disease. Specifically, we define the *risk ratio* via

$$RR = \frac{\Pr(D|E)}{\Pr(D|E^c)} = \frac{|E \cap D|/|E|}{|E^c \cap D|/|E^c|}$$

where $\Pr(X|Y)$ is the probability of X given Y, and $|X|$ is the number of elements contained in set X.

Examining the definitions above, it should be clear that RR is not a fixed number, but is in fact a function dependent on the cutpoint k. Above we argued that the "logical" course of action was to select k to maximize RR, so as to best demonstrate the relationship between the risk factor and the disease. To show the error in this logic, we perform a simple experiment.

We begin by creating a series of fake data sets with a known exposure-risk relationships. Specifically, we generate data that follows the rule: participants with exposure level below 0.7 have 0.1 chance of acquiring the disease, while those with exposure level above 0.7 experience 0.15 chance of acquiring the disease. This is

accomplished by a two stage subroutine: for each patient we first randomly generate an exposure level between 0 and 1, then we randomly generate whether the patient has the disease (using the probability based on their exposure level). For such a data set the "correct" risk ratio should be 1.5 and should arise when the cut-point $k = 0.7$ is selected.

Having generated 100 such data sets, each with 1000 elements, we next analyze each data set to determine $RR_{\max}$ the maximum value of RR by computing RR values using all cutpoints between 0.1 and 0.9 (we avoid the cutpoints too close to 0 or 1 as they are likely to be criticized in published literature). Since, for this data set, the correct risk ratio should arise when the cut-point $k = 0.7$ is selected, we also compute RR for this point, labeling it $RR_{0.7}$. The results are summarized in Table 2.

RR_{max} value	occurrences	RR_{max} value	occurrences
$RR_{max} < 1$	0 times	$RR_{0.7} < 1$	3 times
$1 \leq RR_{max} < 1.3$	0 times	$1 \leq RR_{0.7} < 1.3$	15 times
$1.3 \leq RR_{max} < 1.8$	58 times	$1.3 \leq RR_{0.7} < 1.8$	71 times
$1.8 \leq RR_{max} < 2.0$	17 times	$1.8 \leq RR_{0.7} < 2.0$	7 times
$2.0 \leq RR_{max}$	25 times	$2.0 \leq RR_{0.7}$	4 times

TABLE 2. **Demonstration of Publication Bias In Situ:** Maximal RR occurring by testing cutpoints k between 0.1 and 0.9 and RR occurring at $k = 0.7$ on 100 randomly generated data sets. Correct RR is 1.5 and should occur at $k = 0.7$.

Table 2 is divided into five categories, $RR < 1$, $1 \leq RR < 1.3$, $1.3 \leq RR < 1.8$, $1.8 \leq RR < 2$, and $2 \leq RR$. These could be though of as "RR very low (unpublished)", "RR low (probably unpublished)", "RR correct (published)", and "RR high (published)", and "RR very high (published)" respectively. It is clear from Table 2 that RR_{max} has a much higher tendency to exaggerate the risk ratio than $RR_{0.7}$.

A more extreme example can be made by generating random data sets where exposure has no influence on disease prevalence. That is, for each patient we randomly generate an exposure level, then we randomly generate whether the patient has the disease using the fixed probability of 0.1. The result of the such an experiment are summarized in Table 3.

RR_{max} value	occurrences	RR_{max} value	occurrences
$RR_{max} < 1$	6 times	$RR_{0.7} < 1$	53 times
$1 \leq RR_{max} < 1.3$	22 times	$1 \leq RR_{0.7} < 1.3$	37 times
$1.3 \leq RR_{max} < 1.8$	57 times	$1.3 \leq RR_{0.7} < 1.8$	10 times
$1.8 \leq RR_{max} < 2.0$	7 times	$1.8 \leq RR_{0.7} < 2.0$	0 times
$2.0 \leq RR_{max}$	8 times	$2.0 \leq RR_{0.7}$	0 times

TABLE 3. **2nd Demonstration of Publication Bias In Situ:** Maximal RR occurring by testing cutpoints k between 0.1 and 0.9 and RR occurring at $k = 0.7$ on 100 randomly generated data sets. Correct RR is 1.

On a final note, it should be mentioned that we are not suggesting that picking RR_{max} is standard practice. The above example should be simply considered an example demonstrating that allowing data an undue amount of influence on model selection can lead to extreme biases in the results.

4.3. PREVENT Model: A Computer Simulation for Computing Potential Impact Fractions. In many cases the risk factors and impact of intervention hold complicated interrelationships that make it difficult (or impossible) to compute the potential impact fraction of an intervention. In these cases, one often resorts to developing a computer simulation in order to estimate the potential impact fraction. One such simulation is the PREVENT model developed in the late 1980s [**95**]. In this example we attempt to provide a broad stroke outline of the PREVENT model.

The PREVENT model is an epidemiological approach to predicting the effect on mortality resulting from a health condition after an intervention on known risk factors. In order to capture realistic disease dynamics in a population, it extends the usual epidemiological methodologies to take into account two important facts:

(1) The relationship between risk factors and diseases typically involves many risk factors and many diseases. A single risk factor may play a role in multiple diseases and a single disease may involve multiple risk factors. These two phenomena occur simultaneously, leading to a complicated dynamical relationship between risk factors and diseases.
(2) There may be considerable latency between exposure to a risk factor and the incidence of a disease. These time lags must be incorporated into a dynamical risk model.

In the PREVENT model, the potential impact fraction is time-dependent and defined by

$$PIF_t = \frac{\Pr_t(D) - \Pr_t^*(D)}{\Pr_t(D)},$$

where $\Pr_t(D)$ is the time-dependent probability of disease in the reference population and $\Pr_t^*(D)$ is the time-dependent probability of disease in the intervention population. The model also introduces a new quantity, the trend impact fraction, which is defined by

$$TIF_t = \frac{\Pr_0(D) - \Pr_t(D)}{\Pr_0(D)}.$$

This measures the number of disease cases in the reference population that are prevented (or caused) by a time evolution of the risk factor prevalence. In Figure 1 we see how these values interrelate.

In 1989, Gunning and Schepers evaluated the PREVENT model on Dutch population health data collected from a variety of sources [**95**]. The risk factors examined included cigarette smoking, hypertension, hyperlipidemia, occupational hazards, obesity, diet, and alcohol use. The diseases examined included Ischaemic heart disease, cerebrovascular disease, lung cancer, breast cancer, colon cancer, and stomach cancer. They also examined the link between traffic accidents and risk factors such as years of driving experience, alcohol use, and preventive regulations. This resulted in extremely complicated interrelationships between their risk factors and their health outcomes. Ultimately the PREVENT model was successfully used to study the impact of various interventions on these diseases.

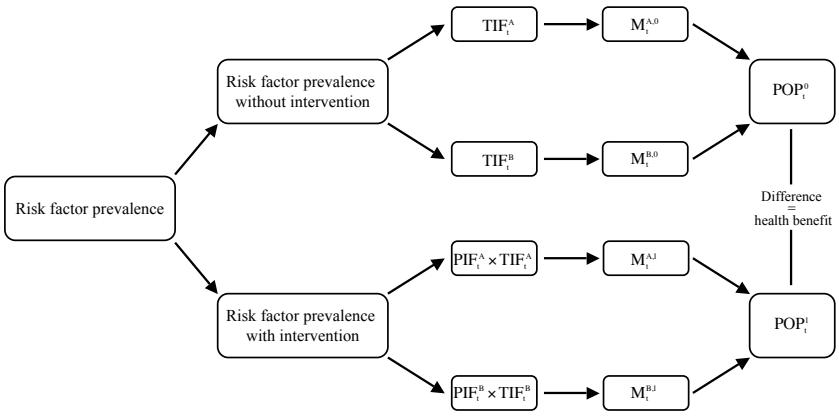

FIGURE 1. **Simulation flow in the PREVENT Model:** This flow represents one time interval of a simplified version of the PREVENT algorithm. Two demographic classes, A and B, are considered with different risk ratios. The superscripts 0 and 1 denote the reference and the intervention populations, respectively. PIF_t is the time-dependent potential impact function, TIF_t is the time-dependent trend impact function. The mortality rates in the reference and intervention populations are M_t^0 and M_t^1, respectively. The time-dependent outcome states of the reference and intervention populations are POP_t^0 and POP_t^1, respectively. The difference between these two states is the health benefit of the intervention.

In 2006, Bronnum and Hansen evaluated the accuracy of the PREVENT model by applying it to synthetic population data generated by simulation [**35**]. They concluded that the model is generally quite accurate, but it does tend to slightly overestimate the health benefits of the intervention. Nonetheless, in realistic scenarios this error would be minor and estimates obtained from the model are accurate enough for use in health policy planning.

In summary, PREVENT is a sophisticated risk simulation model with the ability to use population health data to handle complex disease dynamics. The basic concepts are also flexible and a variety of extensions to the model are possible. Furthermore, the model has considerable scope for applications to healthcare policy development and evaluation.

5. Related Reading

Epidemiological Modelling has close associations with Descriptive Statistics (Chapter 5) and Regression Analysis.The question of how to explain epidemiological results to the public is discussed in Chapter 8.

Reference [**96**] examines public biases in risk beliefs, and is discussed in the opening paragraph of this chapter. Reference [**224**] uses an epidemiological approach to examine the benefits and drawbacks of replacing soap hand washing with an alcohol hand rub in hospitals.

Reference [**152**] discusses the epidemiological parameter of the "etiologic fraction" in comparison to the "preventable fraction." Reference [**185**] gives a commentary on the computational and conceptual issues faced by epidemiologists when dealing with population attributable fraction estimations.

References [**95**], [**123**], and [**35**], are examples of implementation of the PREVENT model.

CHAPTER 8

Adjusting Risky Behaviour

The best cure for hypochondria is to forget about your own body and get interested in someone else's. Goodman Ace (1899-1992)

The best way to stop smoking is to carry wet matches. Anonymous

Psychosocial Risk Modelling

1. Model Overview

Chapter 7 was concerned with the question of how to measure the link between a risk factor and a disease. This provides a measure of healthcare demand based on disease prevalence in the population and a methodology for modelling the impact of public health interventions. In this chapter, we turn our attention to the question of how to develop an intervention strategy. That is, the psychological and social aspects of applying risk modelling: *Psychosocial Risk Modelling.*

The term disease refers to any negative health outcome.

In many cases, once a link between a negative health outcome and a risk factor is found, policy makers attempt to change the prevalence of the risk factor in society by altering the policies governing society. Perhaps the best example of this is the effect of the US Surgeon General's 1964 announcement that smoking causes lung cancer. In Chapter 4 of the US Surgeon General's report "Smoking and Health," it states "Cigarette smoking is causally related to lung cancer in men ... The data for women, though less extensive, point in the same direction" (page 37). Almost immediately the United States enacted the "Federal Cigarette Labeling and Advertising Act", which compelled cigarette companies to issue warnings on all cigarette packs. Shortly thereafter, Britain banned television advertisements for cigarettes, and in 1970 the United States followed suit. Since then, most first world countries have enacted laws regarding warning labels on cigarette packages, restricting (or eliminating) the advertisement of cigarettes, and creating smoke-free zones in many public places. Recently, in July of 2007, England has banned smoking in *all enclosed public places*; this includes every pub, club and bar, and some outside shelters.

Risk factors can be beneficial*, if they decrease the likelihood of a disease, or* harmful*, if they increase the likelihood of a disease. (see Chapter 7 for more details)*

An attempt to adjust the prevalence of a risk factor in society via a new law or policy is a *policy-based* intervention. The word policy is used instead of law since policy-based approaches include examples such as the distribution of free condoms and clean needles in areas with a high prevalence of HIV.

Policy-based interventions can have a profound impact on the prevalence of a risk factor in a population, as in the case with smoking[1]. However, in other circumstances enacting new policies is not effective or not practical. For example, clear links have been uncovered between unprotected sex with multiple partners and the spread of HIV/AIDS. However, enacting anti-adultery laws is ineffective, and policies regarding condom usage can infringe upon religious ideals. Other problems arise when considering more complicated risk factors, such as the recent research linking the consumption of moderate amounts of red wine and individual health. These results suggest that moderate amounts of red wine increase overall health, while large amounts decrease overall health. Clearly, enacting a policy enforcing adults to drink the appropriate amount would be nearly impossible.

When dealing with issues where policy-based approaches are ineffective, many researchers often focus on the idea of improving the individual's knowledge of the matter at hand. In this regard, psychosocial risk modelling attempts to adjust the individual's *perceived risk* and *perceived efficacy* regarding the risk factor and the particular disease. We refer to these approaches as *education-based* interventions.

When researching psychosocial models of risk it is important to remember that it is not the actual risk nor the individual's actual efficacy, but the individual's perceived *risk and* perceived *efficacy that affect how they behave.*

In psychosocial modelling the phrase *perceived risk* is used to refer to an individual's belief on how detrimental (or beneficial) a given course of action is to their health. In this context, perceived risk can be thought of as a balancing of the answers to "what is the worst that could happen?" and "how likely is that to happen?" In adjusting the public's perceived risk, one tries to help them gain a better grasp of the correct answers to these two questions.

The phrase *perceived efficacy* refers to an individual's belief on how much control they have over a given situation. With regards to healthcare, perceived efficacy is not just an individual's knowledge about what actions they can take, but also their knowledge on their ability to perform these actions. For example, a drug user may know that sharing needles is dangerous and increases the risk of HIV, but unless the same drug user knows where to obtain clean needles, they will probably disregard this knowledge. In many cases altering an individual's perceived efficacy involves breaking down certain cultural barriers and beliefs the individual has developed. For example, in many cultures women do not feel they have the right to demand that their sexual partners wear a condom, changing this belief can have very positive effects on the control of spread of sexually transmitted diseases.

As with policy-based interventions, education-based interventions are not always effective in combating risky behaviour. For example, many people choose to take up smoking despite being aware of the health risks and nicotine's addictive properties. Moreover, there is no clear cut answer to the question of when to use a policy-based approach and when to use an education-based approach to adjust risky behaviour. Even in hindsight it is often unclear which intervention has made the larger impact. It could be argued that the decrease in the prevalence of smoking in the United States is not due to policy changes but instead due to spreading the knowledge that smoking is hazardous to one's health. Alternately, we could point out that, despite this knowledge, every year over one million teenagers take

[1]According to the National Center for Chronic Disease Prevention and Health Promotion the adult smoking rate in the United States dropped from 42.4% in 1965 to 22.5% in 2002.

up smoking[2], and argue that without laws restricting the purchase of cigarettes to minors this number would be even higher.

2. Common Uses

The goal of psychosocial risk modelling in healthcare is to develop models explaining how the general public may be swayed into better health behaviour. In general, this question can be reduced to "how can we increase/decrease the prevalence of the risk factor X in the population?" For example:

- *How can we increase the use of clean needles among drug users?*
- *How can we decrease the prevalence of smoking in the population?*
- *How can we improve the eating habits of the general population?*

Psychosocial risk modelling often approaches these questions through education. This alters the question to "how can we better educate the public on the dangers/benefits of X?" For example:

- *How can we better educate drug users on the importance of using clean needles?*
- *How can we better educate the public on the importance of a healthy diet?*

In other cases, psychosocial risk modelling approaches these questions through social policy making or stricter laws. For example:

- *Should smoking be banned from public places?*
- *Should flu-shots be mandatory for certain professions?*
- *Should public school cafeterias adopt a policy of no longer serving high fat food?*

3. Model Details

The two most common approaches to adjusting risky behaviour lie in eduction and policy making. In education-based interventions, we attempt to increase the public's awareness of the risk factor and what they can do to reduce it. In policy-based interventions, we attempt to adjust the prevalence of the risk factor by developing laws or policies that reduce harmful risky behaviour or reinforce beneficial risk factors. Regardless of which approach one selects, in order to alter the health behaviour of an individual it is imperative to understand the underlying factors that impact health behaviour. Many models have been proposed to explain human behaviour, several of which are discussed in other chapters of this book. For now we satisfy ourselves with a brief overview of two of the most important factors in health behaviour: the concepts of *perceived risk* and *perceived efficacy*. (Further descriptions of models that explain human behaviour can be found in Chapters 11 and 10.)

3.1. Perceived Risk and Perceived Efficacy. *Perceived risk* refers to an individual's perception of the harm or benefit of a given course of action. For example, consider the action of maintaining a healthy diet. An individual who sees little benefit in eating healthy would have a low perceived risk while an individual who sees harm in poor eating habits would have a high perceived risk. For this example, an education-based approach to changing an individual's perceived risk

[2]Data according to the National Center For Chronic Disease Prevention and Health Promotion.

might be to educate them on the negative side effects of unhealthy eating habits. Alternately, a more policy-based intervention might involve enacting laws forcing restaurants to provide nutritional information about each menu option.

The term *perceived efficacy* refers to an individual's perception of whether they can achieve a given course of action, and how much effort it would require. Again consider the example of maintaining a healthy diet. Even people with a high perceived risk may eat poorly based on a low perceived efficacy. For example, somebody in a corporate sales position may feel that their job requires them to dine out a lot with clients, which lowers their ability to maintain a healthy diet. One may attempt to change an individual's perceived efficacy by educating them on how to take better control over a given course of action or developing policies that make a given course of action more achievable. In the example of a healthy diet, this might be achieved by teaching people how to select healthy affordable choices from restaurant menus or developing corporate policies that encourage eating at healthier restaurants.

It is worth emphasizing here that when researching approaches to altering health behaviour, it is not the actual risk and actual efficacy, but the *perceived* risk and *perceived* efficacy, that drive an individual's actions. In many cases an individual's perception of risk and efficacy are very different than their actual risk and efficacy.

3.2. Education-based versus Policy-based interventions. Unfortunately there is often no clear cut answer regarding when one should use an education-based intervention and when one should use a policy-based intervention. In fact, in many cases it is unclear whether a given intervention is education-based or policy-based. Consider for example the "Health Canada: Challenge to Youth Media Contest." This contest was organized and funded by Health Canada, and asked youth from across Canada to create a 20 second anti-smoking commercial. The winning entries (the top 20 out of over 10,000 entries received nationwide) were given the opportunity to see their advertisements produced by a professional agency and aired in cinemas across Canada. As a government organized and funded contest, it could be argued that this represents a policy-based intervention. However, the result of the contest was 20 anti-smoking advertisements and the chance for schools nationwide to educate their students on the dangers of smoking. Thus, the contest could very easily be seen as an education-based intervention.

Although there are no strict rules to determining what intervention will be most effective in a given situation, there are some ways to guide our thought process. To begin, it is always a good idea to examine previous interventions. If a past intervention was particularly effective, repeat it or refine it; if a past intervention was a failure, learn from its mistakes.

Another good question to ask is how good is the public's knowledge of the given risk factor and how can they control it (i.e. what is the public's perceived risk and perceived efficacy)? If public knowledge is high, then further education-based interventions may not be effective, while if public knowledge is low, education-based interventions are more likely to have an impact. If public efficacy is low then interventions designed to increase the public ability to act in a positive manner will probably be effective, while if public efficacy is high then such interventions will probably not have an impact. Information regarding public knowledge and efficacy levels can often be obtained via public surveys and forums.

A final question to consider is what are the current policies and laws regarding the risk factor? In some cases, such as drug abuse, the current laws are extremely strict, and creating further laws might not be effective. In these cases more creative interventions are needed.

There are many arguments for implementing both policy-based and education-based interventions to adjust health behaviour. For policy-based interventions the arguments are often along the lines of: without intervention, perceived risk and perceived efficacy are too low. New policies can increase perceived risk by enforcing extra penalties for poor health behaviour (for example, laws governing the use of seat-belts in automobiles) and increase perceived efficacy by forcing health options to be available (for example, providing free clean needles to drug users). The arguments for education-based interventions generally reduce to: inaccurate perceptions of risk and perceptions of efficacy lead to incorrect decisions regarding risky behaviour. Therefore, when developing an education-based intervention strategy, it is important to consider how to accurately communicate information about risks to the public. Measures such as risk ratio and attributable risk (see Chapter 7) may be useful to policy-makers, but for the general public these measures are confusing and difficult to interpret. (Indeed, even trained epidemiologists make errors when interpreting these measures; we discuss this further in Example 4.2.) Similarly, journals such as the *American Journal of Epidemiology* are excellent sources of information for healthcare professionals, but are seldom read by the general public. The lessons from these examples are simple; information should be communicated to the public in a manner they can comprehend and in a location they frequently access. However, effective implementation of these lessons is surprisingly difficult.

4. Examples

4.1. Tanzanian soap-operas and HIV: a success story. The country of Tanzania and its 37 million citizens lies on the East coast of Africa. It is a poor country, with 3 television stations, 150,000 telephones lines, and 1.6 million people living with HIV/AIDS[3]. This HIV infection rate is among the highest in the world.

It has been found that the epidemic is maintained primarily through extra-marital sex, with about 97% of cases occurring through heterosexual intercourse. Thus, modification of sexual behavior appears an important means of controlling the epidemic. In 2000, the results of a 4 year education-based intervention study were published [**173**]. In this example we outline the study, and highlight some of its successes.

A soap opera is an ongoing, episodic work of fiction, usually broadcast on television or radio. They are characterized by their open ended plots and complicated inter-character relationships.

The study began with the proposal of an educational soap opera entitled "Twende na Wakati" (Let's Go with the Times) designed to adjust the levels of perceived risk and perceived efficacy regarding HIV/AIDS. Although it is widely accepted that entertainment-education programs are effective in influencing audience behaviour in various communities, this is the first case of a national-level intervention using entertainment-education as a tool to combat the HIV epidemic.

Much of the theory underlying the entertainment-education asserts that social-cognitive dimensions have a greater impact in effecting behavioural change than merely providing information. That is, providing the information on HIV/AIDS is necessary, but without impacting social beliefs the intervention is unlikely to be

[3]Data from the 2007 CIA world fact book: `https://www.cia.gov/cia/publications/factbook`

effective. As such, Twende na wakati transmitted its message by using characters in the show as negative, transitional and positive role models. Each broadcast was also followed by a 30 minute educational epilogue providing more direct education on HIV/AIDS.

Twende na wakati was broadcast via radio in Swahili twice a week for 30 minutes each. The choice of radio as an intervention media was highly appropriate, as radio is the most popular form of electronic entertainment in Tanzania and an important source of information on HIV/AIDS. Since the intervention strategy aimed at reducing the number of sexual partners and increasing condom use in the broadcast area, the soap opera focused on sharing the following ideas:

- all STDs should be medically treated,
- condoms are effective in preventing HIV infection,
- AIDS is an incurable disease spread by sexual contact, and
- various rumors about HIV/AIDS are false. [4]

In order to assess the impact of the intervention, one region of the country was denoted a control region, and the radio show was not broadcast in this region for the first two years (1993-1995). Due to the positiveness of the preliminary results, from 1995 to 1997 the soap opera was broadcast nationwide.

Data was collected through personal interview surveys that were carried out five times at 1 year intervals, starting just prior to the first broadcast. Males (aged 15 to 60) and females (aged 15 to 49), selected on a sampling grid, were asked to provide information on personal characteristics, exposure to Twende na wakati, other sources of information on HIV/AIDS, and personal attitudes and preventive behaviour practices regarding HIV/AIDS.

Statistical analyses of the Twende na wakati project used ANOVA and logit loglinear models to compare the control and treatment areas. Possible geographic effects were also tested by stepwise multiple linear regression models. (See Chapter 6 for descriptions of these models.)

During the first two years of the study, exposure to Twende na wakati was only 2% in the comparison area and 47% in the treatment area. When broadcasting was extended nationwide, listenership increased to about 60% and 75% in the treatment and comparison areas, respectively.

In the treatment area (the non-control region before 1995, and nationwide after 1995), survey respondents showed a modest but significant increase in knowledge regarding HIV/AIDS. This increase was not present in the control region, suggesting that the soap opera was directly responsible. Although no significant change in attitude toward having sexual partners prior to marriage was seen, personal perception of risk among respondents has increased and significant changes in preventive behaviour was found. Specifically, survey respondents showed a significant increases in condom use and decline in the total number of sexual partners. This behaviour was shown to be influenced through several intervening variables. Most notably, respondents showed an increased perception of risk of contracting HIV/AIDS, an increased self-efficacy with respect to preventing HIV/AIDS, and an identification with the primary characters in the soap opera.

In conclusion, this study is an excellent example of the indirect nature of cognitive processes involved in assimilating information and producing behavioural change. It represents an empirical test of cognitive theories of behaviour change, and shows that entertainment-education can be a highly effective intervention impacting both perceived risk and perceived efficacy.

[4]Some common, and false, rumors about HIV/AIDS in Tanzania include: condom lubricant contains HIV, you can visually identify if people are infected with HIV, and it is harder for fat people to contract AIDS.

4.2. Help the Public Interpret AR and RR. The issue of communicating information in a form the general public can make sense of has been addressed by various researchers ([**178**] and citations therein). As mentioned, there is a common consensus that measures such as risk ratio and attributable risk are difficult to interpret for the general public.

To help avoid public misunderstanding of risk, some researchers have suggested other forms of risk measures to help the public digest what level of risk various activities entail. For example, the impact of risky behaviour could be communicated to the public in terms of "average number of life years lost/saved" [**178**]. Measures such as these are much easier to understand, and allow the public to better gauge what level of risk each risk factor represents. However, the cost of this communication is that it is often difficult to estimate these values from epidemiological quantities such the attributable risk. One possible manner of estimating these values is to resort to a computer simulation involving the risk factors and diseases of interest.

For the case of the impact of moderate alcohol consumption on coronary heart disease, Phillips and Zeckhauser [**178**] developed a simulation model using mortality data from the Framingham Heart Study. Their simulation predicts that the potential life years gained from moderate alcohol consumption[5] would be 0.75 years for men and 0.63 years for females. A number of approximations were made in the simulation, which may call into question the accuracy of their results. Nonetheless, the research of [**178**] provides an excellent example of how combining risk analysis with simulation provides a useful tool for generating information about health choices in a form that may be easily communicated to the public.

In Chapter 7 the concepts of relative risk (RR) and attributable risk (AR) are developed. Relative risk is a measure of the change in the probability of having the specified disease given that one has the risk factor. Attributable risk represents the proportion of the given disease that could be (theoretically) reduced if the given risk factor were eliminated. Relative risk is defined in equation (12) as $RR = \frac{\Pr(D|F)}{\Pr(D|F^c)}$, *while attributable risk is defined in equation (13) as* $AR = \frac{\Pr(D)-\Pr(D|F^c)}{\Pr(D)}$, *where* $\Pr(D|F^{(c)})$ *is the probability of experiencing disease D given risk factor F is present (*^{c} *- not present).*

4.3. Potential Effects of a Slow Reduction of Sodium in the Canadian Diet. The World Health Organization suggests that the average adult should consume a maximum of 2000 mg per day of sodium (or lower, depending on the country) [**172**], and a recent report by the American Academy of Science recommends that 1200 mg/day to 1500 mg/day is an adequate intake for optimal health [**170**]. However, the average Canadian consumes approximately 3500 mg sodium per day. Dietary research suggests that reducing the national sodium intake of Canada would have profound effects on the overall health of the country. For example, researchers have found links between high sodium consumption, increased blood pressure, increased risk of cardiovascular diseases, congestive heart failure, stroke and myocardial infarctions [**104**]. The projected health benefits of reducing the average sodium intake to 1800 mg per day include, a 30.3% reduction in the prevalence of hypertension, resulting in a potential direct savings of $430,000,000 to the Canadian healthcare system [**157**].

Despite the clear health benefits, motivated individuals find it difficult to reduce sodium intake reasonably because roughly 80% of daily intake of sodium comes from processed and restaurant foods in most of the Western world. Moreover, much of the processed food industry is reluctant to reduce the sodium in their products, fearing loss of sales due to a sudden reduction in flavour. As it is not economically defensible to advocate for a sudden dramatic reduction in sodium intake, Joffres and Alimadad have begun research into the effect of a slow annual reduction in sodium levels in processed food [**120**].

[5] 3 to 12 drinks per week, spread out evenly over the course of a week

Since experiments with taste-tasters have shown that most recipes do not begin to show a change in flavour until about a 10% reduction in sodium levels, analysis considers estimates of reductions in cardiovascular diseases in Canada following a 5% and 10% yearly reduction in sodium intake at the population level, and potential savings in hospital costs. In their model, sodium intake begins at 3600 mg per day, and decreases annually. A 5% annual reduction will require 22 years to reduce sodium intake levels to 1226 mg per day, while a 10% annual reduction will require 11 years to reduce sodium intake levels to 1250 mg per day. Nonetheless, the impact of such a reduction would be noticed almost immediately in the healthcare system. In Table 1 we see the predicted number of avoided ischemic heart disease (IHD), stroke and congestive heart failure (CHF) events each year for the 5% and 10% reduction scenarios, along with the expected savings to the healthcare system.

Scenario I: 10% Annual Sodium Intake Reduction

Year	IHD	Stroke	CHF	Hospital savings ($1,000,000s)
1	2238	928	1645	$ 45
2	1968	815	1445	$ 39
3	1726	714	1265	$ 34
4	1508	623	1102	$ 30
5	1311	541	956	$ 26
10	587	238	419	$ 11
11	483	195	341	$ 9
Total	13465	5549	9810	$271

Scenario II: 5% Annual Sodium Intake Reduction

Year	IHD	Stroke	CHF	Hospital savings ($1,000,000s)
1	1262	518	913	$ 25
2	1195	490	865	$ 24
3	1132	464	819	$ 22
4	1071	439	775	$ 21
5	1014	416	734	$ 20
10	767	315	556	$ 15
15	576	237	419	$ 11
20	428	177	313	$ 8
22	379	157	277	$7
Total	16423	6749	11918	$330

TABLE 1. **Potential Effect of a Slow Reduction in Sodium Intake:** Predicted number of avoided ischemic heart disease (IHD), stroke and congestive heart failure (CHF) events each year, along with expected healthcare savings, resulting from an annual 5% or 10% decrease in sodium intake.

Overall the research argues that a small steady sodium reduction plan would increase the level of acceptance by industry and the general population, yet still provide significant improvement to overall Canadian health status. The total predicted hospital cost savings would be about $270 million over 11 years in the 10% model, and $330 million over 22 years in the 5% model.

It should be noted that the model uses the unlikely generalization that all Canadians consume 3600 mg per day of sodium intake in the initial year. This assumption is useful for generating initial estimates, but limits the accuracy of the model. In particular, individuals with very high sodium intake will have significantly better health gains from a 5% reduction in sodium than individuals with low sodium intakes. This may result in even higher savings than those predicted in Table 1. Future work of the researchers would be to find more realistic ways to model the national population.

5. Related Reading

Psychosocial Risk Modelling has close connections with Epidemiological Risk Modelling (Chapter 7). Psychosocial models in general examine individual's behaviour and how it can be altered. Other forms of psychosocial modelling are discussed in Chapter 10 (Psychosocial Modelling), Chapter 11 (Game Theory and Human Capital Models), and Chapter 12 (Network Models and Graph Theory).

A complete copy of the US Surgeon General's report "Smoking and Health" can be found online at `http://www.cdc.gov/tobacco/sgr/sgr_1964/sgr64.htm`. Reference [**118**] uses survey data from 1974 to 1985 to project the smoking prevalence rate in 2000. Information on the Health Canada: Challenge to Youth Media Contest can be found at `http://www.schoolfile.com/cash/HealthCanadaYouth.htm`. Reference [**96**] looks at how biases in risk beliefs vary across a population.

Reference [**173**] examines the effect of an entertainment-education radio soap opera on HIV prevention behaviours in Tanzania, and is discussed in Example 4.1.

Reference [**178**] examines effective ways to communicate information to patients regarding moderate alcohol consumption and the reduction in heart attacks, and is discussed in Example 4.2. Reference [**185**] provides a commentary on conceptual and computational issues faced by epidemiologists when dealing with attributable fraction estimations. Reference [**79**] is an ongoing comprehensive study looking at the general causes of heart disease and stroke.

Part 3

Model Design and Interpretation

CHAPTER 9

Issues in Mathematical Modelling

It's tough to make predictions, especially about the future. Yogi Berra (1925-)

We have a lot of people revolutionizing the world because they've never had to present a working model. Charles F. Kettering (1876-1958)

Model Selection, Development, and Implementation

1. Overview

In many cases, the problems and issues arising in healthcare can be answered (at least as a first approximation) via the statistical techniques discussed in Part 2 of this book. For example, if one wishes to measure whether a particular drug is effective in combating a particular illness, then a double blind test followed by the epidemiology techniques discussed in Chapter 7 is perfectly satisfactory. Alternately, if one wishes examine the basic relationships between various societal factors and health, then the descriptive statistics of Chapter 5, and the regression analysis and econometrics of Chapter 6 provide most of the tools necessary for the project.

However, in many cases, statistical techniques are not sufficient to determine how one factor impacts another. In such cases, one often builds a model of the system and uses it to determine what equation is likely to describe the relationships. In other cases, researchers and policy-makers are more interested in "what-if's" than what is. That is, policy-makers are happy to see that a drug is effective, but are actually interested in questions such as "if we made this drug freely available to the public, how would the global prevalence of the illness change?" To answer questions such as this, more advanced models must be employed. Part 3 of this book describes some of the modelling techniques that have been used to answer these harder problems.

Before turning our attention to these techniques it is prudent to lay down a key observation to keep in mind while using these models. Namely, it must be strongly noted that, just because these techniques use mathematics other than statistical analysis, this does not mean statistical analysis is no longer useful. Indeed, regardless of the modelling technique we use, at some level the model should be rooted in real life, which often means tuning the model using real statistical data. Although, the ideas and mathematics discussed in Part 2 of this book are not necessary to understand how various types of models are constructed, at some point in the modelling process it usually becomes necessary to apply some of the statistical methods presented in Part 2.

Regardless of the modelling technique used, at some level all models must be grounded in reality. Often this means the model must be tuned using real data and statistical analyses (discussed in Part 2 of this book).

With this in mind, we note that the remainder of Part 3 will largely ignore the statistical analysis required to tune models. Instead we focus on providing the reader with a high level understanding of what each modelling technique is and what it may be able to accomplish. Our collection of modelling techniques is in no way exhaustive, but hopefully provides a broad background for any researcher or policy-maker interested in modelling in healthcare.

In the remainder of this chapter we discuss some of the issues that arise in selecting a modelling technique, developing the model, and implementing the results.

2. Selecting a Modelling Technique

Unlike what is commonly taught in grade school, for most problems (especially mathematics problems) there is more than one way to arrive at a solution. This is particularly evident in modelling, as any given question can be approached by several different modelling techniques. This makes selecting the "best" technique for the job a nearly impossible task. In fact, the "best" technique might be to approach the question via several different modelling techniques and compare the answer each model provides.

Selecting a modelling technique to employ is largely a matter of experience and luck. When in doubt, the best course of action is probably to use multiple techniques and compare the results from each.

Broadly speaking, models fall into two categories: qualitative and quantitative. These categories are discussed in some detail in Chapter 3, Section 1, so we do not rewrite these details here. Instead, we simply remind the reader that *qualitative models* are models which avoid the use of numbers. Instead qualitative models are designed to provide insight about why a given situation exists and what its driving factors are. Conversely, *quantitative models* are models that use mathematical variables and equations to describe the behaviour of a system. Such models are designed to make numerical predictions about how a system will evolve over time and how interventions will impact this evolution.

Quantitative models can be sub-divided further into the categories: stochastic or deterministic, static or dynamic, and discrete or continuous. These are discussed in Chapter 3 Section 1, so we say no more here. Instead we turn our attention to the idea of a feedback loop, which is an important concept in both qualitative and quantitative modelling.

2.1. Feedback Loops. A *feedback loop* occurs when the state of one variable in the model impacts how the state of that same variable progresses over time. The classic example is impact of population size on population size. Consider Figure 1 (this figure reappears in Figure 1 of Chapter 10).

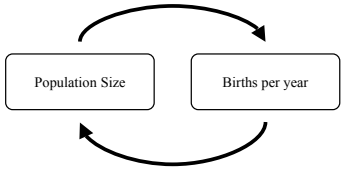

FIGURE 1. **A simple feedback loop.**

In Figure 1 we see a diagram representing how population size and the number of births per year interact. The first arrow, pointing from "population" to "births,"

represents the fact that the population has an impact on the number of births per year. The second arrow, pointing from "births" to "population," represents the fact that the number of births has an impact on the population. Simply put, the first arrow states that the more people there are the more babies are born, while the second arrow states that the more babies are born the more people there are. Therefore, the more people there are the faster the population will increase.

From this simple example it may appear that feedback loops are trivial in their construction and interpretation. However, as models become more complicated, feedback loops can become very difficult to analyze. In fact, in many cases the state of a variable may both positively and negatively impact how that variable changes over time.

Most advanced models in healthcare will contain at least one feedback loop. This reflects real life where, for example, the current health of an individual impacts their future health, the current efficiency of a hospital impacts the future efficiency of the same hospital, and the current length of a surgical waitlist impacts how quickly that waitlist will grow. Feedback loops are often so important and so complicated that sometimes the entire goal of a qualitative model is simply to describe the feedback loops in a system.

When selecting a modelling technique to employ, we should ask if the problem in question has feedback loops, and if so, how important it is to the model to understand them.

3. Developing the Model

After a modelling technique is selected, the next stage a researcher proceeds through is the development of the model. Since this is discussed in some detail in Chapter 3 we do not elaborate here. Instead we focus on some traps that modellers may fall into during the development process.

The model does not answer the stated question: One of the most common problems in a researcher/policy-maker relationship is communication. As a result, a researcher will sometimes spend months developing a brilliant model which in no way tells the policy-maker what they want to know. To reduce the likelihood of this occurring, it is important that the policy-maker clearly state the exact question they wish answered, and that the researcher clearly state how the model will help answer that question. This process should be repeated at regular intervals to reduce the likelihood of losing focus partway through the project.

The theoretical model is incomprehensible: One of the guiding principles of modelling is transparency. Models that are too confusing to understand may be correct, but are seldom employed in the long run. With transparency comes the ability for experts in the field to examine the model and determine if they believe the underlying assumptions are correct.

The model is not believable: It is important to seek out experts in the field we wish to model and include them in the process of model design. Note that experts in the field of study does not mean experts at modelling the field of study. For example, if one is to model disease spread through a hospital, one should discuss the model with doctors who have seen how the hospital system works and how diseases can be spread within that

environment. If the model is clear and believed by them, then there is a significantly higher chance that the model will be successful in the long run.

The model does not fit the data: After developing a model, it is important to test the model against real data. If the model does not fit real data, then it is best to assume that the model is wrong. It should be noted that this is not a total loss, the modeller can now learn something new about the system by examining the model's assumptions and determining which ones are flawed.

In all of the above problems, there is one major trap that modellers often fall into. Specifically, if the model is not working, modellers will often attempt to tweak the model in some manner in order to fix the problem. In some cases this works, especially if the margin of error for the model is small. In other cases, this can lead to a long series of "model corrections" that result in an incomprehensible model that works only for the very specific situation for which it was designed, or that is riddled with untestable assumptions. In the end, the tweaked model will probably be discarded, and instead of losing months of work the researcher ends up losing years of work. To avoid this, modellers (and researchers in general) must be willing to admit that the method they have attempted did not work, and to restart the process beginning with a different modelling technique. This is extremely easy to say, and very difficult to achieve.

4. Implementation of Models

Supposing that a question has been specified, a model has been proposed and designed, data has been collected and the model has been tuned to the data, it is now the task of the modeller to "solve" the model. That is, the model must be examined in a manner which allows the user to answer the questions which were specified. In many cases this means developing equations which describe the state of the model at a given time, given a certain initial state. This is referred to as *implementing* the model, and in general can be accomplished in three manners: *mathematical analysis*, *numerical analysis*, and *simulation*. No single one of these techniques is better than the rest. In fact, like the rest of the modelling process, results are most convincing when multiple techniques are employed. We now discuss each technique in turn.

4.1. Mathematical Analysis. The oldest and most robust manner of implementing a model is through the careful use of a pen and paper. After developing and tuning the model, the modeller may be able to describe the model as a collection of equations. Sometimes these equations can be studied without the aid of a computer, and many properties of the model can be determined. For example, finding *equilibrium points* for the model involves examining for which initial conditions the model does not evolve over time. Answers to questions such as these may provide insight into the model and the system modelled that computer aided techniques do not.

The tools used for mathematical analysis of a model vary from model to model, and a comprehensive review of analytical methods in mathematical modelling is far beyond the scope of this book (in fact it would essentially amount to a survey of the entire field of mathematics). Instead, we will give a brief introduction to a few

of the branches of mathematics that are most often applied to modelling. In each of the remaining chapters of Part 3 we will discuss the tools needed to analyze specific models.

Statistical Analysis: Statistics is the mathematics beyond the collection, analysis, interpretation, and presentation of data. It is a huge field of study and most schools require at least a first or second year statistics course in order to complete a Bachelor of Science degree. As mentioned above, and illustrated in Part 2 of this book, the field of statistics plays a pivotal role in the development of models.

Calculus: Calculus is the study of mathematics enhanced by the concept of a limit. Like statistics it is a huge field of study and most schools require at least a first year calculus course in order to complete a Bachelor of Science degree. From a modelling perspective the most important concepts from calculus are the derivative and integral of a function. In this book we denote these by $\frac{d}{dt}f(t)$ and $\int_a^b f(t)dt$ respectively. In the world of modelling the derivative of a function most commonly represents the rate of change of that function with respect to time (it some cases it may be with respect to another variable). The integral of a function plays the role of an anti-derivative. In modelling the integral often represent a sort of continuous version of the sum.

Linear Algebra: Linear algebra is the branch of mathematics concerned with the study of vectors, and their operators. Most schools teach linear algebra as a first or second year course and make it mandatory for a large variety of degrees. The most common appearance of linear algebra in modelling is in the form of matrix multiplication and eigenvalue analysis. Matrix multiplication generalizes the notion of a linear function ($y = mx + b$) to multi-dimensional spaces. Eigenvalue analysis is used to determine fixed points of such functions, and to understand how "stable" the functions are with respect to errors in the data.

Multivariate Calculus: Calculus in more than one dimension is usually referred to as multivariate calculus. Most schools teach multivariate calculus as a second or third year course and only make it mandatory for certain degrees. Although the principles of the subject are the same, many of the definitions must be reworked to make sense in a multi-dimensional light. In particular, derivatives become gradients ($\frac{d}{dt}f$ becomes ∇f) and integrals are taken over sets instead of intervals ($\int_a^b f$ becomes $\int_S f$). In modelling, multivariate calculus arises when the model has multiple interacting variables.

Differential and Integral Equations: A system of differential equations is a set of equations that describe the infinitesimal interaction between different components of the system. They are characterized by equations that involve both the function and its derivative (or its integral). Most schools teach differential equations as a third or fourth year course and only make it mandatory for certain degrees. Differential equations arise often in modelling, particularly when the model evolves over time in a continuous manner. If the model contains feedback loops then the differential equations become more complicated and more difficult to solve analytically.

Two important concepts in differential equations are *attractors* and *equilibrium states*. An attractor is a state towards which the system evolves over time, regardless of initial conditions (or at least for a wide variety of initial conditions). Equilibrium states occur when differential equations attain states from which they do not leave. All attractors are equilibrium states, but not all equilibriums states are attractors. Understanding how differential equations behave near equilibrium states is useful in understanding the model as a whole.

Graph Theory: Graph theory is the study of mathematical structures representing connections between objects. Graph theory is an advanced subject in mathematics and many schools do not teach it at an undergraduate level. In modelling, graph theory is most strongly applicable when examining network models. Such models are discussed in Chapter 12.

Optimization: Optimization is the study of minimizing or maximizing a function. Basic optimization is touched on in most first year calculus classes, but advanced optimization is not often taught at an undergraduate level. In modelling, optimization is usually applied at an end stage when policy-makers are interested in what policy changes might improve the systems behaviour. In order to be effective, optimization requires a clearly defined objective function. That is, questions like "how can a hospital be run best?" cannot be approached through optimization as they are too vague. However, specific questions such as "what nursing schedule minimizes the number of staff hours while maintaining a given level of service?" can be approached through optimization.

4.2. Numerical Analysis. As models become larger and more complicated, it becomes increasingly difficult to solve them via analytic means. Fortunately, many of the mathematical tools of analysis have been automated to various degrees in mathematical programming languages. Some, but far from all, of these languages are outlined in Appendix A of this text. Here we discuss in general terms what these languages are capable of.

If the complication in implementing the model is a result of model size (as opposed to model complexity), it can often be difficult to solve simply because there is so much room for error when using a pen and paper. A prime example of this is statistical analysis for large data sets. In this case, several of the mathematical programming languages are capable of solving complicated systems of equations (or differential equations) in a symbolic manner. The final equation may be complicated, but computer aid can be further enlisted to produce graphs of the equation. By changing parameters, resolving, and regraphing, researchers can often gain a good deal of understanding about a model in very little time. As an added bonus, computers do not tend to get upset about tedious grunt work.

If the complication in implementing the model is a result of model complexity, mathematical programming languages can be used to produce numeric (approximate) solutions to the equations of the model. Good examples of this are the solving of large systems of equations via fix point methods, or the solving of difficult differential equations via Euler's method or the higher order Runge-Kutta method. Numerical methods are also extremely useful for solving difficult optimization problems.

Fortunately, it is not necessary for the user to fully understand how a method works in order to use it. However, it is best if the user has a decent understanding of the methods employed so that they know whether these methods are appropriate in their situation. As an analogy, consider that we do not need to know how the combustion engine works to drive a car, but every good driver should know that the tires should be properly inflated before going anywhere.

4.3. Simulation. Sometimes a model is complicated enough that even writing down the equations that represent the model becomes challenging. In cases such as this, simulation software is often used to try and create a fully computerized version of the model. Due to the ease of understanding simulations, simulation software has become extremely popular in recent years. In Appendix A we list some of the simulation software packages currently available. Here we discuss in general terms what simulation is, and what it is capable of.

As a simulation is a method of implementing a model, it is not surprising that it can be stochastic or deterministic. In a stochastic simulation, events are triggered according to a probability distribution. For example, patients may arrive at the emergency room according to a Poisson process, with a given expected arrival rate. Clearly, this is more realistic than a deterministic simulation in which it is assumed that patients arrive at a specified fixed interval. Whether the additional complexity and computational overhead of a stochastic model are necessary depends on the level of detail of the problem that the model is addressing.

In the past, simulations have been termed either *discrete event simulation* or *continuous simulations*, however current practice has put discrete event simulations at the forefront. In discrete event simulation, the time flow follows a sequence of discrete steps, whereas in continuous simulation the time flow is continuous. Historically, continuous simulations were done using analogue computers, essentially electrical circuits custom built to emulate the system to be modelled. Although analogue computers are still used in a few highly specialized modelling problems, the majority of simulations now use digital computers. By necessity, all simulations on a digital computer must use discrete time steps, thus the term continuous simulation has largely fallen out of practical use. When using the term simulation, we shall always be referring to discrete event simulation.

In discrete event simulation, the model progresses through a series of events as defined by the simulation algorithm. For example, a patient entering a hospital emergency department might follow the following sequence of events:

(1) enter emergency
(2) initial assessment
(3) treatment
(4) release

The simulation algorithm is based on an understanding of the processes being modelled. In general, the simulation would not consist of a simple sequence of events as above, but would include conditional branching and iterations.

Since static models do not need to be simulated, all simulations are dynamic. Choosing the correct time steps for the simulation is crucial to its success. The *time steps* in a simulation correspond to a time scale in the physical system. For example, in a simulation of an emergency department, each iteration of the simulation may correspond to an hour of time in the physical system, or if the model is intended

to focus on more detailed dynamical behaviour it may correspond to a minute of physical time. The model is then run through the required number of iterations which correspond to the desired physical time span.

It is often the case that during the simulation we may wish to vary the input statistical distribution or model parameters over time. For example the rate of patient arrivals to an emergency department will typically depend on the time of day. It is well known that the emergency department is busier during the afternoon and evening than in the early hours of the morning. In this case, we may use a *moving Poisson process*, in which the expected arrival rate varies slowly with time.

Typically, discrete event simulations must be run for a "warm up" period, before they realistically model the physical system. Consider again our example of a hospital emergency department. When the simulation starts, there are no patients in the system. However, this is certainly not the case with the actual emergency department. It has existed, serving patients, since the hospital was constructed.

The level of detail within the simulation algorithm is critical. There is no simple rule about the degree of detail to incorporate. Like all models, simulations must tread a fine line between being detailed enough to solve the problem for which they are designed and simple enough to allow it to be clearly understood. Aspects that are extraneous to the question being investigated should be excluded from the model. It is best to begin with a simple model using aggregated data and add complexity only where needed, validating at each step.

5. Related Reading

Reference [**9**] examines how the operation of the healthcare industry differs from typical manufacturing industrie. Reference [**13**] presents guidelines for the verification, validation, and accreditation of simulation models. Reference [**3**] discusses the need for interdisciplinary research teams when modelling in healthcare. Reference [**222**] discusses the many applications of modelling in healthcare, including some useful advice on collaboration between modellers and policy-makers. Reference [**8**] looks at the potential uses of complexity theory and chaos theory in the understanding of healthcare organizations. Reference [**65**] discusses multilevel analysis and reviews the rationale for using it in public health research. Reference [**144**] describes and compares the development of strategies to manage demand in healthcare in the U.K. and the U.S. Reference [**12**] describes a methodology for carrying out a sensitivity analysis in healthcare models. Reference [**62**] gives a description of the area and role of Agent Based Social Simulation. Reference [**14**] presents a method for assessing the quality of large-scale complex modelling and simulation applications. Reference [**61**] reviews the location set covering model, maximal covering model and P-median model that form the core of the location planning in healthcare. Reference [**145**] illustrates the utility of models in assisting managers in optimizing healthcare delivery in military medical centers. Reference [**153**] looks at the background of system dynamics modelling and its uses in modelling public health.

CHAPTER 10

Explaining Irrational Behaviour

Man is a rational animal. Aristotle (384–322 BC)

Man is a rational animal who always loses his temper when he is called upon to act in accordance with the dictates of reason. Oscar Wilde (1854–1900)

Psychosocial Modelling

1. Model Overview

For modelling purposes, it would be considerably easier if everybody acted in an isolated and rational manner. However, a brief examination of almost anyone's life shows that this untrue. Indeed, if rationality was the norm, the fifty five billion dollar casino gambling industry would be in serious jeopardy[1], attendance at local vaccination clinics would be considerably higher[2], and over 150,000 psychologists would suddenly be out of work[3]. Therefore, whenever one attempts to develop models of human behaviour, it is important to examine the role irrationality plays in this behaviour.

On an individual level, modelling irrational behaviour is, of course, impossible. However, on a group level, it is possible to examine how social conditions affect the general behaviour of group members. To this end, the field of psychosocial modelling has been developed to examine the impact of social conditions on behaviour. In healthcare, psychosocial models develop psychological frameworks that examine how social conditions affect how people make decisions about their health.

Psychosocial: relating social conditions to mental health.

So far, psychosocial modelling in healthcare has enjoyed a fairly long and successful history. In the early 1950s, the *Health Belief Model* began the examination of why individuals were reluctant to accept disease preventive measures, such as vaccination and screening tests. In the 1960s, the *Behavioural Model of Healthcare* was developed to study the more general question of when and why *families* access healthcare. Both of these models have blossomed into large fields of research, which include examinations of how our use of the healthcare system is impacted by our social circle. More recently, advances in computers and mathematical data analysis have allowed these theoretical models to be examined and verified.

---

[1]The *2006-07 Indian Gaming Industry Report* totals 2005 casino gambling revenue in the USA as $55.3 billion.

[2]A 2003 survey of 1330 Canadian adults showed that 79.4% of subjects held positive views towards vaccines, but only 45.4% of subjects had or intended to have an influenza vaccine (Rivto, et. al., *Journal of Immune Based Therapies*, **1**:3).

[3]According to the US Bureau of Labor Statistics there were approximately 179,000 psychologists employed in the USA in 2005.

The Health Belief Model suggests that six elements affect an individual's decision to access healthcare: the perceived benefits, perceived susceptibility, perceived severity, perceived barriers, cues to action, *and* self-efficacy.

The Health Belief Model is based upon the psychological theory that human behaviour depends on the value placed by the individual upon a particular goal and the individual's estimation of the likelihood of achieving that goal. Essentially the health belief model suggests that an individual's decision to access healthcare is affected by six elements: *the perceived benefits of accessing healthcare, perceived susceptibility to requiring the benefits, perceived severity of not acquiring the benefits, perceived barriers of accessing healthcare, cues to action,* and *self-efficacy* (see Table 1, page 98). The goal in researching the Health Belief Model is to determine how these elements interrelate and what factors affect them. This is described in some detail in Subsection 3.1.

The Behavioural Model of Healthcare considers that an individual's behaviour is not only affected by their environment, but also contributes to the development of their environment (see Figure 2).

The Behavioural Model of Healthcare is based on the concept that an individual's behaviour is not only a product of their environment, but also a contributing factor to the development of that environment. For example, if a particular area is lacking in trained medical personnel , people in that area may seek alternate forms of healthcare, thus reducing the need for medical resources in the area. Such an effect is called a *feedback loop* and a system that describes these feedback loops is called a *demand-access-utilization chain* or *influence diagram.* The goal in researching the Behavioural Model of Healthcare is to examine the interrelationships between the *physical environment, social environment, health behaviour,* and *health outcome.* Details of this are discussed in Subsection 3.2, and the influence diagram of the Behavioural Model for Healthcare is given in Figure 2.

2. Common Uses

It should be noted immediately that the goal of psychosocial models is to provide a psychological framework to help understand patient behaviour, not to produce numerical predictions of future patient behaviour. Nonetheless, psychosocial models can help answer many questions posed in the healthcare industry.

Perhaps the most common use is the examination of when and why people make use of healthcare. In this regard , psychosocial models are used to approach the general question: "what prompts people to visit a trained medical practitioner?" More practically one might use psychosocial modelling to answer questions such as:

- *How can we improve attendance at immunization clinics?*
- *How can we improve patient compliance to medical instruction (such as proper use of antibiotics)?*
- *How can we increase the usage of mammography examinations to diagnose breast cancer?*

Of course this is far from the only use of psychosocial models in healthcare. In fact, it is reasonable to say that anytime we attempt to model general patient behaviour on a group level, we should incorporate Psychosocial models to some degree. For example, psychosocial models should be included when approaching questions such as:

- *How can we improve the lunch time eating habits of high school students?*
- *What factors impact smoking rates in teenage girls?*
- *Who should we educate to reduce unwanted pregnancies?*

3. Model Details

Two of the most popular psychosocial models in healthcare are the *Health Belief Model* and *Behaviour Model of Healthcare*. We will approach each of these in turn. However, before doing this, it is worth noting that psychosocial models are qualitative models, and not quantitative models. That is, the goal of psychosocial models is to provide a psychological framework to help understand patient behaviour, not to produce numerical predictions of future patient behaviour.

Qualitative versus quantitative modelling is discussed in Chapters 1 and 9.

3.1. The Health Belief Model. The Health Belief Model (henceforth referred to as HBM) is a conceptual framework developed in the early 1950s by psychologists attempting to understand the widespread failure of people to accept preventative measures in healthcare (such as vaccinations and screening tests). The model is based on the psychological theory that human behaviour is largely driven by the perceived value of a given goal, and the perceived likelihood of achieving that goal. In the field of healthcare, the model states that an individual's likelihood of accessing healthcare is based on six elements: *perceived benefits, perceived susceptibility, perceived severity, perceived barriers, cues to action,* and *self-efficacy* (see summary in Table 1).

The original HBM *of the 1950s only considered the elements of* perceived benefits, perceived susceptibility, perceived severity, *and* perceived barriers. *The elements of* cues to action *and* self-efficacy *were not added until the late 1970s.*

The element of *perceived benefits* captures the positive outcomes that an individual feels may occur by accessing healthcare. Generally this captures an improvement in life quality or life span. For example, one might visit a doctor in the hope of relieving a persistent cough, take an influenza vaccine in the hope of avoiding a nasty flu bout, or begin chemotherapy in the hope of prolonging one's life.

The element of *perceived susceptibility* captures the individual's feeling on whether or not the perceived benefits will be required. For example, a teacher may feel highly susceptible to catching a nasty flu during the course of a year, and therefore be likely to take an influenza vaccine. Conversely, a young university student may feel they are the epitome of health and therefore not perceive themselves susceptible to influenza. Similarly, a 25 year old woman will perceive a lower susceptibility to breast cancer than a woman of 55 years old, and therefore be less likely to have a mammogram.

Sometimes the elements of perceived susceptibility and perceived severity are combined into one element: perceived threat.

The element of *perceived severity* captures how an individual feels about the result of requiring the benefits but not acquiring them. For example, an individual educated on the effects of cancer may perceive a higher severity in the illness than someone who has not been educated. Perceived severity also captures possible social consequences such as how a disease may affect an individual's work or family life. For example, influenza may be perceived as more severe to someone who cannot abide the possibility of missing work and HIV/AIDS may be perceived as a greater stigma to devout individuals.

The element of *perceived barriers* captures the potential difficulties in accessing the particular aspect of healthcare in question. For example, the distance to the local clinic and the difficulty of getting the necessary time off work would be perceived barriers to getting regular mammograms. Perceived barriers also include any potential negative aspects of accessing healthcare. For example, chemotherapy generally has a negative impact on quality of life, and many vaccines are given via a potentially painful needle.

The element of *cues to action* captures both bodily and environment events that motivate individuals to act. For example, a disease that causes a constant

Element	Summary	Examples
Perceived Benefits	Perceived positive outcomes of accessing healthcare.	• Relief from pain • Reduced likelihood of getting sick • Increase in life span
Perceived Susceptibility	Perceived likelihood of requiring the perceived benefits.	• Vaccination and profession • Age and cancer screening
Perceived Severity	Perceived seriousness of requiring but not receiving the perceived benefits.	• Education and cancer screening • Sickness and loss of work
Perceived Barriers	Potential difficulties and negative aspects of accessing healthcare.	• Distance required to travel • Loss of time (personal or work) • Pain of receiving treatment
Cues to Action	Bodily and environmental motivation to seek healthcare.	• Current symptoms • Media coverage
Self-Efficacy	An individual's confidence in their ability to overcome the perceived barriers to accessing healthcare.	• Positive or negative reinforcement

TABLE 1. **Elements in the Health Belief Model:** Elements affecting an individual's access to healthcare according to the HBM.

ache is more likely to prompt an individual to seek help than a disease that does not . Another major contributor to action is media coverage.

Efficacy: the power to produce an effect.

The element of *self-efficacy* or *perceived efficacy* captures the individual's belief in their own personal ability to overcome the perceived barriers to accessing healthcare. This element differs from perceived barriers in that it is focused more on the individual confidence than the individual perception of potential difficulties. For example, an individual's self-efficacy can be greatly influenced by the positive and negative reinforcement they receive in the various social circles in their life.

Since its inception, the HBM has undergone considerable scrutiny and analysis. As a purely qualitative model, the vast majority of the research on the HBM has been in the form of psychological experiments. Interviews and survey data has been collected to examine the plausibility of the HBM in practice. Not surprisingly, the HBM model has been largely supported by the published data.

As a psychosocial model, the HBM is primarily designed to explain patient behaviour in a psychological framework. As such, one of the primary goals in researching the HBM is to determine strategies to alter each element of the HBM. Most of these strategies can be summarized under the heading "improving patient education."

Another goal of research in the HBM is to determine which element has the greatest impact on a given problem. For example, if we want to improve attendance at an immunization clinic, we would like to know whether to increase the population's perceived susceptibility (perhaps through media coverage), decrease the population's perceived barriers (perhaps by providing free public transit to and

from the clinic), or provide a cue to action (perhaps through flyers reminding people of the clinic date and time). One example of this type of research is given in Example 4.1.

Previous research in the HBM has left it with several drawbacks. First and foremost, as a Psychosocial model it is incapable of forming predictions on the effect of a given policy change. Second, past research using the HBM has only focused on a single element of the model, and therefore the usefulness of the model as a whole has never been confirmed. And finally, attempting to change an element in the HBM model, even for a given individual, is seldom as easy as it appears. Nonetheless, the HBM provides an important and potentially powerful model for understanding individual behaviour towards healthcare.

3.2. The Behavioural Model of Healthcare. The *Behavioural Model of Healthcare* first emerged in the late 1960s as an approach to understanding when and why families access professional healthcare. The primary premise of the model is that an individual's behaviour is not only a product of their environment, but also a contributing factor to the development of that environment. Such a relationship is called a *feedback loop* since changes in the system "feedback" into the system, causing further changes in the system.

Before discussing the case of healthcare, consider a simple example of population change. In 1900, there were approximately 76 million people living in the United States of America[4]. By 1950, this number had increased to approximately 151 million, an increase of 75 million. By 2000, the population had increased by another 145 million to approximately 296 million. The reason for this increase in population growth can be simply explained: more babies were born from 1950 to 2000 than from 1900 to 1950, because there were more people to have babies. Diagrammatically we capture this idea in what is called a *demand-access-utilization chain* or *influence diagram*. The influence diagram for this example can be found in Figure 1.

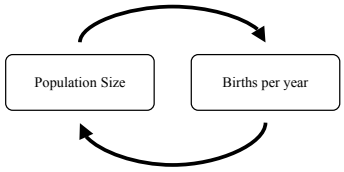

FIGURE 1. **Feedback loops in a simple influence diagram.**

Figure 1 is an influence diagram with just two boxes and two arrows. The first arrow, pointing from "population" to "births," represents the fact that the population has an impact on the number of births per year. The second arrow, pointing from "births" to "population," represents the fact that the number of births has an impact on the population.

Returning to healthcare, the Behavioural Model of Healthcare can be captured in a similar influence diagram. In Figure 2, we provide a version of the influence diagram typically used in the Behavioural Model of Healthcare . As before, each

[4]All statistics from the US census bureau.

arrow represents that the box from which it leaves has an impact on the box to which it points.

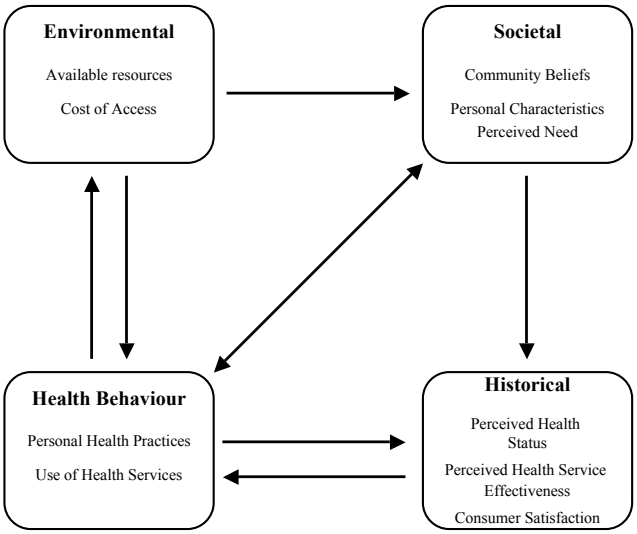

FIGURE 2. **Feedback loops for the Behavioural Model for Healthcare.**

The Behavioural Model of Healthcare considers three main classifications of factors that impact the use of health services: environmental, societal, and historical. The environmental factors largely consist of the availability of health services, and the cost of accessing these services. The societal factors include personal, family, and community beliefs about healthcare. These beliefs are weighted against the perceived need for healthcare. The historical factors include consumer satisfaction and the perceived effectiveness of previous uses of health services.

The category titles given in Figure 2 differ in various research papers. Some common alternatives include "population characteristics" *instead of societal factors and* "past outcomes" *instead of historical factors.*

In the Behavioural Model of Healthcare , the decision to access or not access healthcare is contained in the health behaviour of the individual. The health behaviour also includes personal health practices (such as eating habits) that are not directly involved with the use of health services.

In Figure 2, we see that (according to the Behavioural Model of Healthcare Utilization) health behaviour is impacted by all three of these factors. We also see that health behaviour has an impact on historical factors, which in turn have an impact on the societal factors and health behaviour.

Displayed in Figure 2 we have shown one version of the Behavioural Model of Healthcare. Figure 2 is actually a fairly simple version, as more modern versions include a variety of other factors. For example, research has suggested that factors such as age, gender, personal values, knowledge on healthcare, physician to population ratios, and health insurance may also influence the use of health services. Other research has evolved to consider the importance of perceived need for care and the perceived health status as a function of socioeconomic status. These ideas can be incorporated into the influence diagram either by adding new boxes, or by splitting previous boxes into various parts.

Like the Health Belief Model (Subsection 3.1) the Behavioural Model of Healthcare has several drawbacks. As with the Health Belief Model, the Behavioural Model of Healthcare is a psychosocial model. As such, it is incapable of forming predictions on the effect of a given policy change. In fact, due to the inherent feedback loops , one conclusion of the Behavioural Model of Healthcare is that a single policy change will cause changes at all levels of the system. The strength of this model lies in understanding why these feedbacks occur, and how to use them to our advantage.

4. Examples

4.1. The Impact of Self-Efficacy. In 2004, Albert Bandura published a work focused on health promotion and disease prevention using a variant of the Health Belief Model (HBM)[**15**]. His variant proposed that health behaviour was based on the following core determinants:

(1) personal health goals,
(2) personal outcome expectations,
(3) personal knowledge of health risks and benefits,
(4) perceived self-efficacy,
(5) perceived facilitators, and
(6) perceived social and structural impediments to the health goal sought.

Let us begin by examining each of these with regard to the HBM.

Items 1 and 2, personal health goals and personal outcome expectations, can easily be seen as perceived benefit of action. Item 3, personal knowledge of health risks and benefits, is a regrouping of perceived susceptibility, perceived severity and perceived benefits of the HBM. Item 4, perceived self-efficacy, clearly falls into the self-efficacy element of the HBM. Items 5 and 6, perceived facilitators and perceived social and structural impediments to the health goal sought, can be viewed as a regrouping of cues to action and perceived barriers of the HBM. Thus, Bandura's model can be simply seen as a new categorization of the HBM.

As the concept of self-efficacy was first introduced in 1977 by Bandura, it is no surprise he believes it to be a focal determinant in health behaviour.

Bandura's research led him to believe that an individual's quality of health is heavily influenced by that individual's lifestyle habits. As such, people are able to exercise some control over their health by managing their lifestyle habits. It is his belief that "self-efficacy is a focal determinant because it affects health behaviour both directly and by its influence on the other determinants"[**15**]. In order to support this claim he discusses the impact of self-efficacy on several previous studies into health behaviour.

For example, in 1987 Meyerowitz and Chaiken [**149**] tested which style of pamphlet would have the greatest effect on breast self-examination. They exposed several groups of college-aged females to four different pamphlets regarding breast cancer. The first group was shown a pamphlet focused on providing information on breast cancer. The second group received a pamphlet designed to raise the perceived severity of breast cancer. The third pamphlet was designed to raise one's perceived susceptibility, and the fourth to raise one's self-efficacy in regards to breast cancer. A final control group was not provided with any information on breast cancer. Meyerowitz and Chaiken interviewed the participants both immediately after the intervention and four months later. The results of the study conclude that only

measures of perceived self-efficacy were differentially affected by the various pamphlets. That is, any change in behaviour was due more to a change in self-efficacy than any other *tested* element of the HBM.

Bandura's work provides several other examples along these lines. [5] In total, Bandura's examples lend solid support to his claim that self-efficacy is a strong determinant in health behaviour. As such, Bandura suggests that future attempts to regulate health behaviour should include practices that are designed to raise an individual's self-efficacy.

4.2. Assessing Factors Influencing Mammography Visits. Breast cancer is the second most common type of cancer and it is among the top five causes of cancer mortality [**171**]. Like other forms of cancer, many lifestyle factors increase the risk of breast cancer. For example, diet, physical activity and alcohol consumption are considered important risk factors with respect to breast cancer [**81**]. Also like other forms of cancer, early detection is often crucial in improving the chances of successful treatment and recovery. For breast cancer, mammography is considered to be the most reliable method for detecting breast cancer, and testing frequency can be viewed as a modifiable lifestyle factor that can impact cancer risk. In one direct assessment of screening programs in two counties in Sweden, a 63% reduction in mortality was observed among women who underwent screening when compared with women who did not [**208**]. Yet, utilization of mammography services routinely falls short of targets set out by leading health agencies. This problem is even more pronounced among African-American women, who are less likely to get breast cancer but more likely to die from it due to later diagnosis. In 1991, Stein, Fox, and Murata found that health beliefs were an important factor in mammography compliance in general [**202**]. Therefore, the disparity seen between African-American and Caucasian women raised the question whether differences in mammography utilization may be explained by health belief differences.

In 2004, Vadaparampil et al. published a study of women accessing mammography services that used the Health Belief Model to examine differences in frequency of visits between African-American and Caucasian women [**216**]. This cross-sectional study conducted in the USA, assessed the Health Belief Model using a modelling approach known as *structural equation modelling (SEM)* [**216**]. In this example we overview SEM, and outline the insights it provided regarding the Health Belief Model and mammography test frequency.

SEM is a statistical method for testing the causal structure of the relationships among a group of variables. The causal structure is usually constructed as a conceptual model based on qualitative data in the form of path diagrams (see Figure 3 for example). Hypotheses are tested by examining the variances and covariances of the variables.

A key task in path analysis is the formulation of the conceptual model, which is depicted as a flow chart. Equations are represented by boxes and ovals, with arrows pointing from independent to dependent variables, also referred to as *exogenous* and *endogenous* variables. Exogenous variables are not explained by the model and may include factors such as age or socioeconomic class. Fluctuations in endogenous variables are explained by the model as variables influencing them

[5]Some of Bandura's more interesting examples include: the effects of serial dramas on AIDS prevention in Tanzania, a role-playing video game for diabetic children, and a self-management program for pain control in arthritis.

are specified in the model. SEM can include both observed and *latent variables*. Latent variables represent theoretical constructs about phenomena that cannot be observed directly. Self-concept, self-efficacy, perceived benefit, fear or motivation are examples of latent variables. Latent variables are operationalized in terms of observed variables using a measurement instrument. For example, questionnaires may be administered using a defined scale for self-efficacy. Scores on this scale represent values for the observed variable which in turn imply the latent variable of self-efficacy [**38**].

The main steps in structural equation modelling include: model specification, parameter estimation, assessment of fit, modification of the model, and interpretation of the results. Model specification refers to the formulation of the conceptual model using a path diagram. Parameter estimation and assessment of fit, uses statistical methods to determine the most likely correlations within the model. Modification of the model is used to improve fit through changing the model structure. Finally interpretation of the model involves communication of results implied by the model of best fit.

SEM has become popular in recent years in testing the Health Belief Model. This approach has provided a means of exploring the role of health belief parameters in choices made to prevent disease or promote health. SEM opened the possibility to study complex phenomenon that cannot be evaluated experimentally. Nonetheless, this method has some major weaknesses. Relationships among variables are not likely to be linear. Furthermore, a model that fits the observed data well does not imply a specific causal model. There may be alternative models that may also explain the data just as well. With multiple variables, the complexity may quickly become overwhelming and may cloud basic underlying phenomena. With these limitations in mind, SEM can be used effectively to evaluate potential directions and relationships, and exclude unlikely options.

Returning to our mammography example, the question to be examined was whether health beliefs contributed to African-American women accessing mammography services less often than Caucasian women did. The study began by interviewing 1045 African-American and Caucasian women over 50 years of age in Indiana and Missouri. Participants had no history of breast cancer and had no mammograms in the previous 15 months. Measurement instruments for breast health, mammography and cancer in general were applied to assess perceived susceptibility, perceived benefits, perceived bariers, self-efficacy, fear and fatalism. Local demographic information and data on mammography testing was also collected. The conceptual model for relationships among health belief variables is shown in Figure 3.

Structural equation modelling was carried out using the LISREL software and tested African-American and Caucasian women separately to examine differences between the groups. A number of differences were found. The Health Belief Model explained 13% of the variance for Caucasian women but only 9% for African-American women. Perceived benefits, perceived barriers, and self-efficacy were important variables for Caucasian women, while income, perceived benefits, perceived barriers, and fear were significant for African-American women. These results led the authors to conclude that while health beliefs are a good starting point in addressing inequitable use of mammography services, other differences such as socioeconomic issues must also be taken into consideration.

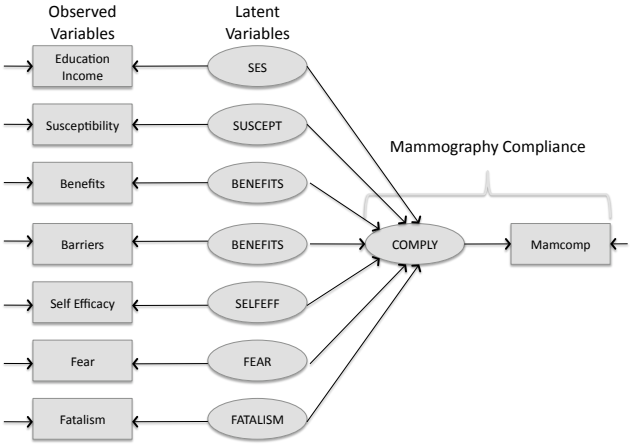

FIGURE 3. **Examining the Health Belief Model with respect to mammography visits:** A conceptual structural equation model (SEM) for relationships among health belief variables for evaluating differences between African-American and Caucasian women with respect to frequency of mammography visits. Based on the work of [**216**].

4.3. Healthcare Utilization Among HIV$^+$ Injection Drug Users. HIV infection among injection drug users (IDU) is a huge health concern on a global scale [**213**]. Nearly a fifth of all diagnoses of new HIV infections are attributed to IDU in Canada [**168**], and 25% of all people diagnosed with AIDS in the USA have been among IDU [**73**]. Moreover, IDU populations are often marginalized, and burdened with multiple economic, social and health problems that generate barriers to optimal healthcare utilization. Late diagnoses of infection, use of emergency rooms as the primary source of health services, and other similar factors lead to elevated healthcare costs related to HIV infection among IDU.

In 2006, Mizuno et al. used Andersen's Behavioural Model of Healthcare to analyze baseline data collected in a large US-based multi-site behavioural intervention study with the goal of identifying barriers and facilitators of healthcare utilization by IDU [**154**]. IN this example we review some of their findings.

Based on the Behavioural Model of Healthcare conceptual framework, Mizuno et al. examined *predisposing, enabling* and *need factors* that were specific to HIV positive IDU. For example, receiving case management and participating in methadone maintenance treatment were included as enabling factors. Other less specific enabling factors, such as social support and perception of the quality of engagement with care providers were also included. Need factors included self-percieved health status and having depressive symptoms. Also included were need factors specific to the study group, including CD4 cell count (a clinical measure indicative of the stage of HIV infection) and intensity of injection drug use.

The goal of the investigation was to identify associations between healthcare utilization measures and variables quantifying each of the three domains defined in the Behavioral Model of Healthcare. Interview data from 1161 HIV positive

injection drug users who participated in the *Intervention for Seropositive Injectors - Research and Evaluation (INSPIRE)* study was analyzed. Poor healthcare utilization was assessed in terms of two outcome measures: (1) reporting fewer than 2 outpatient visits in the past six months and (2) reported to the ER as the usual place of care. Variables with statistically significant associations to healthcare utilization were then included in multivariate logistic regression models predicting poor healthcare utilization.

The investigators found that enabling factors were more important predictors of healthcare utilization among injection drug users than predisposing and need factors. In general, enabling factors are considered more modifiable and better targets for interventions. In this study, having health insurance, having seen a case manager and better engagement with healthcare providers emerged as the most important enabling factors. Of these factors, provisions of health insurance and case management services were identified as the most important potential intervention targets, because these rendered non-signifcant, other, hard-to-modify factors, such as lower education and unstable housing, as barriers to healthcare utilization.

5. Related Reading

Psychosocial models have close connections to Psychosocial Risk Modelling (Chapter 8) and Systems Thinking (Chapter 14). Much of the quantitative side of Psychosocial Modelling is captured in the Human Capital (Chapter 11) and Network Models (Chapter 12).

Reference [**135**] uses multiple regression analyses to assess the ability of the Health Belief Model to account for observed variation in various preventative health behaviours. References [**119**] provides a review of 46 Health Belief Model studies and their findings published up until 1984. Reference [**149**] reviews an experiment that tested which style of pamphlet is most persuasive in increasing the number of breast self-examinations in college-age females. Reference [**167**] investigates some of the ways that economists can incorporate recent research insights in combining psychology and economics to understand risky behaviour by adolescents. Reference [**179**] examines effective ways to communicate information about moderate alcohol consumption and reduction in myocardial infarction risk to patients and physicians. Reference [**204**] investigates the possible role of lifestyles, knowledge about health, emotional well-being and perceptions of control and their effect on health in Europe. Reference [**15**] examines health promotion and disease prevention from the perspective of the Health Belief Model.

Reference [**5**] is one of the first papers developing the Behavioural Model of Healthcare. Reference [**6**] discusses the Behavioural Model of Healthcare, and contains another version of the influence diagram for the Behavioural Model of Healthcare. Reference [**70**] investigates the behavioural response of individuals to changes in costs of healthcare. Reference [**124**] contains an empirical study that investigates the correlation between schooling and good health. Reference [**216**] contains further details regarding Example 4.2. References [**38**] and [**201**] provide more information on structural equation modelling.

Reference [**154**] contains further details related to Example 4.3.

CHAPTER 11

Modelling Optimal Behaviour

Economics is haunted by more fallacies than any other study known to man. This is no accident. The inherent difficulties of the subject would be great enough in any case, but they are multiplied a thousandfold by a factor that is insignificant in, say, physics, mathematics, or medicine – the special pleading of selfish interests. Henry Hazlitt (1894-1993)

Insanity in individuals is something rare – but in groups, parties, nations and epochs, it is the rule. Friedrich Nietzsche (1844 - 1900)

Game Theory and Human Capital Models

1. Model Overview

The question of why people behave as they do is one of the oldest and hardest questions to answer worldwide. Usually this question is examined by psychologists via a battery of creative psychological experiments. Recently, however, mathematicians have begun their own assault on the question of human behaviour via logic and optimization. In order to approach the question of human behaviour in a tractable manner, mathematicians ask the question "how would people behave if each decision was made logically and focused on maximizing personal gain?"In such models each *player*'s decision impacts the gain of the other players in the game. (Here players refer to the individuals being modelled.) Therefore players seek to maximize their personal gain, subject to the worst possible (or most likely) decisions of the other players. These ideas are the basis of the rapidly expanding field of *game theory* and the development of *human capital models*.

In game theory*, decisions are associated with* payoff functions *(also called* utility functions*), and individuals are modelled to make decisions based on maximizing their individual payoff.*

In order to quantify personal gain, each decision is associated with a *payoff function*. Of course, quantifying the idea of personal gain in a payoff function is more difficult in some fields than others. If personal gain is measured in terms of monetary gain these concepts become easy to define, so it is not surprising that much of the early development of game theory was based in economics. However more recently, researchers have considered short term and long term health as quantifiable factors, and used these to develop what is referred to as the human capital model.

Before discussing human capital, it is prudent to discuss the *Nash equilibrium*. Introduced by John Nash in his Ph.D. thesis, the Nash equilibrium captures the idea that (in a multi-player strategic game) the optimal action for any given player depends on the actions of the other players. The Nash equilibrium is based on the assumption that each player will select their strategy based on their assessment

of the action to be taken by the other players. All players are assumed to follow this logic, and to correctly assess the strategy of other players. Equilibrium occurs when no player has anything to gain by changing only their strategy, thus all future rounds of the game will result in the same collection of decisions.

In the human capital model*, long-term health is considered a stock that can be bought or sold for other assets.*

Returning to healthcare, game theory has a clear use when examining how policy changes regarding financial incentives will affect doctor and patient decisions. To move beyond this limited usage we must examine more creative payoff functions. In particular, payoff functions in healthcare will often consider health as an asset that can be bought into or traded off for other gains. In such cases, the payoff functions are often renamed *utility functions*. Such functions are the keystone of *human capital models*. The main idea in human capital models is that people may increase their "value" through investing in education, training, or health.

Although human capital models are not generally considered game theory models, much of the theory is the same. Like in game theory, players (in this case people and their employers) make decisions regarding their training and health. These decisions are based on their assessment of how improving these factors will increase their value at the cost of time and/or money. Also, like in game theory, decisions are complicated by a player's perception of how the other players will react to a given decision. For example, will providing training to employees encourage them to seek better paying jobs? These ideas are further complicated by concepts such as quality of life, trust, and other random factors.

Both game theory and the human capital model have been criticized by various quarters. Opponents of game theory generally argue that decisions are seldom made in the purely logical manner that game theory uses. Proponents of game theory respond to this by stating that one of the strongest uses of game theory is to determine where humans are irrational thereby providing focus for future study.

The critics of human capital models generally run along the same lines. Critics state that most human capital models are too simplistic, and human capital models that are not overly simplified are impossible to work with. Furthermore, human capital models are founded on the *representative agent* approximation for economic systems. For many simple economic systems, the representative agent approximation has been shown to be valid. However, this is not true in general. In particular, in economic models in which consumers have limited information and economic models in which there are interactions between agents, the representative agent approximation has been shown to be flawed. This is the case in healthcare where the information asymmetry between patients and providers is typically significant, and patients interact with other patients to determine trust factors for given physicians. Nonetheless, human capital models may still be useful for understanding some components of healthcare demand and other health behaviour.

2. Common Uses

Game theory is a branch of applied mathematics that studies strategic situations where players choose different actions in an attempt to maximize their returns. The theory provides models of rational decision-making in strategic interactions. As such, it may be used to understand how people interact with the healthcare system, addressing questions such as:

- *How does trust affect doctor-patient cooperation and quality of care?*
- *How do incentive structures affect physician decisions?*

- *How might user fees affect patient waiting times?*

Human capital models seek to understand decisions regarding health from an economics perspective. Like game theory, the human capital model creates a payoff function that captures the idea that health is a *stock* that can be bought or sold for other commodities. These ideas make human capital models useful for examining questions such as,

- *How do user fees and insurance impact the demand for healthcare?*
- *What is the role of family in the demand for healthcare?*
- *How can we better understand drug addiction?*

3. Mathematical Details

Fortunately many game theory models often do not require advanced mathematics to understand. Unfortunately, they often require a long and carefully thought-out series of logical steps that can be difficult to validate. As such, it is probably easiest to approach the mathematics of game theory via an example. We begin with the classic Prisoner's Dilemma.

3.1. The Prisoner's Dilemma, an Introduction to Game Theory. Possibly the most famous example of a strategic game in mathematics is the *Prisoner's Dilemma*. The game considers the following situation:

> Two partners in crime are arrested by the police. Although the police have sufficient evidence for a minor charge, they have insufficient evidence for a major conviction, so they separate the prisoners and ask them to testify against the other. In return the police offer the following deal,
>
> (1) if neither partner testifies then they will both receive one year sentences on the minor charge,
> (2) if one testifies against the other and the other does not testify, the one who testifies will receive a full pardon but the one who does not testify will receive a 10 year sentence,
> (3) if they both testify then they will both receive 6 year sentences.
>
> Given that neither prisoner knows how the other will behave, how should the prisoners act?

Each prisoner has a choice of two strategies: testify or not testify. Let us denote these strategies by T (for testify) and N (for not testify). The result of each prisoner's choice is usually called the *payoff* and captured in a *payoff table*. To demonstrate, in Table 1 we provide the payoff table for each combination of strategies. The table represents the jail time each prisoner will incur given their own and their partner's strategy.

Next we consider the Nash equilibrium for the Prisoner's Dilemma. The *Nash equilibrium* is defined as a combination of strategies in which no player has anything to gain by changing only his or her own strategy unilaterally. From the table it is clear that the minimum total jail time occurs if neither prisoner testifies, and the maximum total jail time occurs when both prisoners testify. Therefore one might conclude that the best strategy is for neither prisoner to testify. However, if prisoner 1 does not testify, then it is in prisoner 2's best interest to testify, while if prisoner 1 does testify it is still in prisoner 2's best interest to testify.

		Prisoner 2	
		N	T
Prisoner 1	n	(1 year, 1 year)	(10 years, 0 years)
	t	(0 years, 10 years)	(6 years, 6 years)

TABLE 1. **The payoff table for the Prisoner's Dilemma problem.** The bracket value is the amount of jail time prisoner 1 and 2 will receive, respectively. For example, if prisoner 1 choses strategy t and prisoner 2 choses strategy N then prisoner 1 will receive no jail time and prisoner 2 will receive 10 years jail time.

Therefore it is always in prisoner 2's best interest to testify. Reversing 1 and 2, the same argument shows prisoner 1's best course of action is also to testify. As a result the Nash equilibrium is for both prisoners to testify. Hence, without some extra determinants of behaviour (such as partner loyalty) both prisoners will testify, resulting in the maximum total jail time.

Although the Prisoner's Dilemma may appear contrived, it provides a mathematical example of why people may not always work towards the greatest good. In particular, the Prisoner's Dilemma demonstrates how a lack of trust can result in the overall worst solution instead of the overall best solution. Interestingly, the Prisoner's Dilemma has been used to model many "irrational" behaviours in real world situations and healthcare. For example, the logic behind the Prisoner's Dilemma can be easily transformed into an explanation of the stock piling of nuclear weapons ("if they have them and we don't..."), the lack of concern over pollution ("if we slow production but they don't..."), and road congestion ("if I drive well, but he doesn't..."). In healthcare the Prisoner's Dilemma has been used to model doctor-patient cooperation, the demand for pharmaceuticals, and the rising cost of hospital nursing staff. In all of these cases the overall optimal solution is offset by the fact that an individual can gain (or at least not lose as much) by not playing to the communal good. Example 4.1 provides details on how the prisoner's dilemma can model doctor-patient cooperation.

In this book, the horizontal rows in the payoff table will represent the strategies for player 1, and the vertical columns represent the strategies for player 2.

3.2. Zero-sum Games and the Mini-max Criterion Solution. In the Prisoner's Dilemma, the participants of the game were called prisoner 1 and prisoner 2. Since game theory is easier discussed in a general form, we will use the word *players* to refer to the participants of a game. If a game is played once then the players must each select a *strategy* and the *payoff table* is consulted to determine the outcome. In this book, the horizontal rows in the payoff table will represent the strategies for player 1, and the vertical columns represent the strategies for player 2. If a game is played multiple times, we refer to each playing of the game as a *round.* In this subsection we will discuss more advanced notions in game theory, in particular zero-sum games.

An alternate definition of a zero-sum game *is a game in which wealth is never created or destroyed.*

A *zero-sum* game is a game in which the sum of the gains and losses between all players (with losses taken as negatives) is zero. In particular, in two player zero-sum games, whenever one player wins the other must lose. In zero-sum games with more than two players there may be multiple winners, but each value won must be lost elsewhere. Hence in zero-sum games, whenever there is a winner, there is a loser.

It is easy to provide examples of zero-sum games, as most forms of sports and gambling are zero-sum games. Simple zero-sum games, such as baseball or chess, result in each player either winning or losing. More complicated zero-sum games, such as poker, may have different levels of victory depending on how the game plays out. (It is worth remarking that, when poker is played in a casino it is no longer a zero-sum game as wealth is destroyed by the casino taking a cut of each pot.)

Solving zero-sum games (that is, finding the strategy that rational players should follow) can be accomplished by a series of techniques. The two most common are *dominance* and the *mini-max criterion*.

The idea of dominance is based on eliminating strategies that are dominated by other strategies. More precisely, if strategy x always provides a better payoff that strategy y, regardless of the other players action, then one should never select strategy y, so it can be removed from the payoff table. An example of a game successfully solved by dominance is provided in Table 2.

	A	B	C
x	(5, -5)	(10, -10)	(10, -10)
y	(0, 0)	(-25, 25)	(-10, 10)

$\Downarrow$

	A	B	C
x	(5, -5)	(10, -10)	(10, -10)

$\Downarrow$

	A
x	(5, -5)

TABLE 2. **A payoff table solved by dominance.** (player 1 payoff, player 2 payoff) Player 1 may select either strategy x or y, and player 2 may select strategy A, B, or C. Regardless of player 2's action, player 1 will always reap the most profit if strategy x is played, therefore player 1 should always select strategy x. In the resulting table, strategy A dominates player 2's payoff, so player 2 should always select strategy A. The end result is player 1 gaining 5 per round, while player 2 loses the same amount.

Dominance solutions rely on the payoff table containing strategies that are clearly superior for certain players. This is seldom the case, after all who would agree to play such a game? In cases where dominance does not result in a complete solution, one turns to the mini-max criterion (a.k.a. maxi-min criterion) for further guidance. For a two-person zero-sum game it is rational for each player to choose strategies that maximize the minimal expected payoff. However, a player must be careful not to be predictable or the other player will use this to their advantage.

To illustrate consider the following payoff table

	A	B
x	(5, -5)	(-20, 20)
y	(-10, 10)	(15,-15)

Player 1 may look at the table and see larger gains by playing strategy y (15 instead of 5) and smaller losses (-10 instead of -20) therefore lean towards using

that strategy. However, player 2 will probably notice this, especially if player 1 uses strategy y every time, and therefore lean towards using strategy A. This leads player 1 to use strategy x, which causes player 2 to use strategy B and so on. To break free from this tailspin of circular logic, game theory proposes that both players select their strategies randomly with some probability distributions.

The terms mini-max and maxi-min are used interchangeably in mathematics. This is not surprising as the first is short for minimize-maximize the second is short for maximize-minimize.

Let p_x be the probability that player 1 selects strategy x, p_y be the probability that player 1 selects strategy y, q_A be the probability that player 2 selects strategy A, and q_B be the probability that player 2 selects strategy B. In order to make sure everybody plays exactly one strategy per round we must have

$$p_x + p_y = 1, p_x \geq 0, p_y \geq 0 \text{ and } q_A + q_B = 1, q_A \geq 0, q_B \geq 0.$$

We can therefore reduce our variables to $p = p_x$ and $q = q_A$, and solve $p_y = 1 - p$ and $q_B = 1 - q$ later. Given these probabilities the expected value for player 1 in any given round is

$$E(p, q) = 5pq - 20p(1 - q) - 10(1 - p)q + 15(1 - p)(1 - q).$$

For this example, the expected value for player 2 is the negative of this value. Player 1 seeks to maximize this value and controls p, while player 2 seeks to minimize this value and controls q. Therefore we have the optimization problem

$$\min_q \max_p \{5pq - 20p(1 - q) - 10(1 - p)q + 15(1 - p)(1 - q) \ : \ 0 \leq p \leq 1, 0 \leq q \leq 1\}.$$

Notice that if one fixes q then this is a linear problem in p, and if one fixes p then this is a linear problem in q. Such problems are called *Linear Mini-max Problems*. In 1944, von Neumann showed that the linear mini-max problems that arise from two-person zero-sum games are always solvable. Moreover, he provided a technique for finding the solution. Simply put, he noticed that at the solution the expected function must be flat and therefore has a gradient of zero [1]. In our case this yields,

Recall that the gradient function is the multi-dimensional version of the derivative; it measures the slope of a multi-dimensional function.
$\nabla f = [\frac{d}{dx_1} f, \cdots \frac{d}{dx_N} f]$.

$$\begin{aligned} 0 = & \ \nabla E(p, q) \\ = & \ \left(\frac{d}{dp} E(p, q), \frac{d}{dq} E(p, q)\right) \\ = & \ (5q - 20(1 - q) + 10q - 15(1 - q), 5p + 20p - 10(1 - p) - 15(1 - p)) \\ = & \ (50q - 35, 50p - 25)\,. \end{aligned}$$

Which implies

$$p = 0.5, \text{ and } q = 0.7.$$

Therefore player 1 should use strategy x 50% of the time and strategy y 50% of the time, while player 2 should use strategy A 70% of the time and strategy B 30% of the time. The expected payoff following such strategies is $5(0.5)(0.7) - 20(0.5)(0.3) - 10(0.5)(0.7) + 15(0.5)(0.3) = -2.5$. On average, in each round, player 1 should expect to lose 2.5, while player 2 should expect to win this amount. (Therefore, player 1 should refuse to play this game.)

[1] Proving this observation, and proving that a solution will always occur, is considerably more difficult that stating it. All of these things were done in [**156**].

3.3. Human Capital Models. Although the human capital model was not developed under the guise of game theory, the ideas within bear a close resemblance with game theory. In particular, the *human capital model* is based on using long term health as a type of payoff function and modelling individuals as players who seek to maximize this payoff. In the human capital model the payoff function is usually referred to as a *utility function*.

The term human capital comes from economic literature, where human capital refers to the stock of productive skills and technical knowledge embodied in labor. The term became popular due to the economist Arthur Cecil Pigou who stated "There is such a thing as investment in human capital as well as investment in material capital."

Like game theory, human capital models in healthcare are largely based on logic and optimization techniques. To see this, let us begin by letting H_t represent the *health stock* of an individual at time t. That is, H_t represents a quantification of an individual's perceived health at a given time. Human capital theory supposes an individual will seek to maximize this stock over time.

The health stock at time $t+1$ is related to the health stock at time t by

$$H_{t+1} = H_t + I_t - \delta_t H_t \tag{21}$$

where I_t is the gross investment in health and δ_t is the depreciation rate of health. Human capital models suggest various methods to maximize this stock at various points in time subject to various *wealth constraints*. Wealth constraints state that a person's wealth can never become negative in order to buy more health stock.

One example of a wealth constraint is the *full wealth constraint*:

$$\sum_{t=0}^{n} \frac{P_t M_t + Q_t X_t + W_t(\Omega - \tau)}{(1+r)^t} = \sum_{t=0}^{n} \frac{W_t \Omega}{(1+r)^t} + A_0.$$

In the full wealth constraint n is the total number of time intervals t, Ω is the total amount of time available, τ is the time spent working, W_t is the wage rate (at time interval t), M_t is a vector representing all goods purchased that contribute to health (at time interval t), P_t is the price vector for goods contributing to health (at time interval t), X_t is the vector of other goods purchased (at time interval t), Q_t is the price vector for these other goods (at time interval t), r is the market rate of interest, and A_0 an individual's initial wealth. Essentially, the full wealth constraint states that the amount of money earned over a life time will exactly equal the amount spent. This may seem unrealistic unless one considers any money left over at the end of life as money spent on providing inheritance.

Most applications of human capital theory use a relatively simple form for the dependence of the utility function on the variables in the model. For such models, various well studied methods in mathematical optimization can be applied to determine maximal solutions to the model. For example, in many cases the model can be solved via the method of *Lagrange multipliers* . More realistic forms for the payoff function are likely to have numerous local maximums and therefore more complicated global analysis techniques are required to study them. In either case, the mathematics behind the optimization of human capital models are beyond the scope of this book.

4. Examples

4.1. Doctor-Patient Relationships as a Prisoner's Dilemma. Although, it is a highly simplistic view of the doctor-patient relationship, the Prisoner's Dilemma may be formulated in a medical context. In a medical consultation, it is possible for the physician either to recommend treatment that is in the patient's

best interest or (whether through error, misjudgement, lack of skills, or conflicting goals) to recommend treatment that is not in the best interest of the patient. In any given consultation, the patient has to decide whether to follow the physician's prescribed course of treatment or ignore the doctors advice and seek another physician. Let us label these "strategies" as follows:

The Doctor (Player 1)	The Patient (Player 2)
• Strategy G – provide Good advice	• Strategy F – Follow the advice
• Strategy B – provide Bad advice	• Strategy I – Ignore the advice

There are four possible outcomes:

(G, F): physician provides good advice; patient follows the treatment plan,
(G, I): physician provides good advice; patient does not follow the treatment plan,
(B, F): physician provides poor advice; patient follows the treatment plan, and
(B, I): physician provides poor advice; patient does not follow the treatment plan.

To create a payoff table for these options consider the following arguments. First, from the physician's perspective producing bad advice requires no effort, and therefore we set the physician's payoff for bad advice to 0. Next, producing good advice costs the physician some effort, but is rewarded if it is followed; therefore we set the physician's payoff for good advice as +1 if it is followed and -1 if it is not. From the patient's point of view, following good advice is rewarding (payoff = +1) but following bad advice is detrimental (payoff = -1). Finally, from the patient's perspective, ignoring the physician's advice causes no change in health status, so we set the payoff to 0. This produces the following payoff table: This table has two

Player 1

		F	I
Player 2	G	(+1, +1)	(-1, 0)
	B	(0, -1)	(0, 0)

Nash equilibriums, (G, F) and (B, I). Note, neither player benefits from departing from these strategies. However, the (G, F) equilibrium is unstable in the sense that if one player changes their strategy the other player stands to lose. Conversely, the (B, I) equilibrium is stable in the sense that if one player changes their strategy then only that player stands to lose. Therefore, game theory would argue that the (B, I) outcome is the rational outcome for this game. Thankfully, in the real world this is not the usual case.

The missing ingredient in this Prisoner's Dilemma model of the doctor-patient interaction is that the doctor-patient encounter is not an isolated event, but is a series of interactions over which a sense of trust develops. Therefore, this should be modelled as a repeated game that allows some concept of trust and communication. Once these elements are added to the game the (G, F) equilibrium becomes stable.

4.2. Incoporating Family and Health Decline into Human Capital Models. Two criticisms of using human capital models to describe health are that they do not take into account the role of relationships in health and they do not explain why people sometimes give up on their health. Works by Bolin, Jacobson, and Lindgren [**30**] and by Gjerde, Grepperud, and Kverndokk [**85**] have discussed these two issues and developed distinct models to examine them.

In 2002, Bolin, Jacobson, and Lindgren developed a human capital model based on the Grossman Model and game theory that incorporates the role of family into the health decision process. In this extended model, a family consists of a husband, a wife, and a single child. Spouses interact strategically both in the production of their own health and in the production of health in other family members.

The strategic aspect of health investment is that the more the wife invests in the health of her husband, even as he also invests in his own health, the more likely he will be to invest in her health. Conversely, the more the wife invests in her own health, even while the husband also invests in her health, the less likely he will be to invest further in the her health. The same strategic rules also hold with the husband and wife interchanged. With this in mind, situations are considered in which none of the individuals can be sure that the other individual will honour a co-operative agreement concerning how to allocate joint resources. Also, the incentives for husband and wife to invest in their child's health may be altered by changes in government policies and regulation (such as child allowance and custody rules).

One possible application of this model is that it may be used to understand the health effects of divorce. The model predicts that members of families that are divorced are less healthy than other individuals. However, as noted in [**30**], this prediction is not entirely borne out by the empirical evidence (although some recent studies do support this conclusion). It is possible that this discrepancy reflects the somewhat simplistic form of family dynamics in the model.

Another interesting aspect of health is that people tend to adapt to their state of health over time. For example, if someone has had a long illness or a history of chronic disease, then after an improvement they might rate their health status as good, even though their absolute health might be lower than someone who is normally healthy. This phenomenon is addressed in [**131**] and [**85**] where the Grossman Model is modified to incorporate adaptation.

One method to incorporate adaptation into a human capital model is to assume that the payoff function depends not on the objective health status, but on a definition of subjective health status. This could be done by defining the objective health function as

$$K(t) = \frac{H_0}{1+\beta} + (1+\beta) \int_0^t e^{-\beta(t-s)}\, H'(s)\, ds, \tag{22}$$

where $H_0 > 0$ is the initial health endowment, H' is the derivative of the objective health status, and the parameter $\beta \geq 0$ determines the importance associated with present health status as opposed to past health status. The extent to which the subjective health takes into account changes in the health status in the past is determined by the parameter β. The larger β is, the more the individual focuses only on recent changes in health. If β is smaller, then the individual takes into account a greater time period when determining their subjective health status.

In their model, Gjerde, Grepperud, and Kverndokk, treat lifetime as an endogenous uncertain variable by using a probabilistic *hazard function* to model the occurrence of death [**85**]. This allows them to calculate an *expected lifetime utility function*, without assuming a fixed lifetime. The model is then solved by optimizing the expected lifetime utility.

The primary conclusion from this model is that adaptation leads to a lower optimal health stock over time as the individuals adapt to declining health with age. Furthermore, the rate of return to health services also decreases with time.

This leads to a decline in health service demand as more resources are devoted to consumption.

4.3. Rational Addiction. The theory of rational addiction, first developed in [**22**], uses a human capital model to model addiction as a consistent plan to maximize a utility function over time. This model may be applied to harmful addiction, such as to alcohol, cocaine, and cigarettes, as well as to more benign addictions such as work, eating, or television. Note that maximizing utility does not necessarily mean maximizing beneficial utility. Rather, the claim is that the addictive behaviour of the consumer is forward-looking with strong, stable preferences and that this behaviour can be captured through a utility function.

A consumer is potentially addicted to a good, if an increase in his current consumption of this good increases his future consumption of it. Thus, in the rational addiction theory, someone is addicted to a good only when past consumption of the good raises the *marginal utility* of present consumption. This model of addiction implies that strong addictions may be ended only by going "cold turkey". From a public health policy perspective, this model for addiction may be useful for evaluating the impact of interventions on addictions that are harmful to public health, such as smoking, alcohol use, harmful drug use, or obesity.

Cold Turkey *is a slang phrase used to describe individuals who attempt to quit an addiction all at once. There are many ideas about the origin of the phrase, including the idea of keeping a cold turkey in the fridge and eating it whenever a craving arises.*

In this model, a permanent change in the price of addictive goods may have only a small initial effect on demand, but the effect grows over time until a new steady state is reached. This aspect of the model may be used to quantify the effect of "sin taxes", which have long been used to control the consumption of alcohol and cigarettes.

The rational addiction model of Becker and Murphy has been critiqued in detail. Critiques point out that the Becker-Murphy rational addiction model assumes that consumers become addicted while having perfect foresight of the consequences of the addiction. This is contrary to most behavioural studies of addiction. As noted above, a key prediction of the Becker-Murphy model is that expected higher future prices result in lower consumption today. Although this conclusion is supported by empirical studies, they are other possible explanations for this behaviour.

A detailed solution of the optimal control problem posed by the Becker-Murphy model is given in [**44**]. This provides the opportunity to examine detailed predictions of the Becker-Murphy model and compare it to other models of addiction. Further work along these lines may lead to models that are useful for evaluating public health policy towards addiction.

5. Related Reading

Game theory models have close connections to psychosocial risk models (Chapter 8), Psychosocial Models (Chapter 10), and Optimization (Chaprter 16).

Much of the initial early development of game theory can be found in reference [**156**] and Nash's Ph.D. thesis [**160**]. The latter includes the initial development of what came to be known as the Nash equilibrium. A non-technical introduction to game theory can be found in reference [**146**].

Uses of the Prisoner's Dilemma regarding the demand for pharmaceuticals can be found at `http://abcnews.go.com/Technology/WhosCounting/story?id=98179`. Uses of the Prisoner's Dilemma regarding the rising cost of hospital nursing staff can be found at `http://www.gametheory.net/News/Items/059.html`. Reference [**72**] uses game theory to examine wait times under various user fee scenarios.

The development and theory of human capital models can be found in references [**18**], [**23**], [**93**], [**19**], [**47**], [**20**], [**21**], [**131**] and [**85**]. Reference [**18**] is the original application of human capital models to health and references [**23**] and [**21**] expand the idea of human capital. Reference [**93**] constructs a model of the demand for the commodity "good health" and looks at the "shadow price" of health (variables other than the price of medical care). This model has come to be known as the *Grossman Model*. Reference [**19**] examines the possibility of using an economic approach to provide a unified framework for understanding human behaviour. Reference [**47**] applies human capital to pediatric healthcare. Reference [**20**] uses economic approaches to analyze social issues. References [**131**] and [**85**] extend the Grossman Model.

Some criticisms of human capital models include [**74**], [**128**], [**82**], [**129**] and [**83**]. References [**128**], [**129**], and [**74**] demonstrate that there are flaws in the representative agent approximation in economic models where there are interacting agents. References [**82**] and [**83**] expand on these flaws and suggest an approach to economic analysis that regards the economy as a complex system of interactions between agents.

Bolin, Jacobson, and Lindgren's work (Example 4.2) can be found in reference [**30**] and is largely based on work by Grossman [**93**]. Also discussed in Example 4.2 is the work of Gjerde, Grepperud, and Kverndokk [**85**]. Also of interest on this topic is reference [**29**].

Example 4.3 largely examines references [**22**] and [**44**]. Reference [**22**] develops a theory of rational addiction to provide insights into addictive behaviour. Their approach is critiqued in reference [**44**].

CHAPTER 12

Modelling Social Interaction

Christianity teaches us to love our neighbour as ourselves; modern society acknowledges no neighbour. Benjamin Disraeli (1804-1881)

It is your business when the wall next door catches fire. Horace (65 BC-8 BC)

Network Models and Graph Theory

1. Model Overview

In 1967, a prominent psychologist by the name of Stanley Milgram performed what became an extremely influential experiment in both pop culture and the scientific community. The experiment began with Milgram contacting a random individual in the city of Omaha and asking them to forward a letter to an individual in Boston[1]. If the Omahan knew the Bostonian on a first name basis, then they could mail the letter directly. Otherwise, the Omahan was asked to mail the letter to someone they knew on a first name basis whom they thought might know the Bostonian. Each person to receive the letter was given the same instructions: if you know the Bostonian on a first name basis mail the letter to him, otherwise mail the letter to someone you think might know the Bostonian on a first name basis.

Not surprisingly, a significant proportion of the people refused to participate, but eventually 64 of the original 296 letters reached the Bostonian. The remarkable result of this experiment was that of the 64 letters to arrive, the average number of times that the letter was passed on to reach its final destination was just 5.5. That is, on average 5.5 stamps were required for the letters to reach their final destination. This prompted the now famous *six degrees of separation* hypothesis, or the *small-world property*: if we define a person as one step away from each person they know, two steps away from each person who is known by one of the people they know, then everyone is no more than six steps away from each person on Earth. This hypothesis has resulted in a Broadway play[2], a film[3], a board game[4], and a growth in the use of modelling techniques based on *network theory*.

Before discussing network theory, it is worth noting that there are several critiques to the six degrees of separation experiment. For example:

[1] Omaha is situated near the center of the United States of America, 1400 miles (2250 km) west of Boston. Boston lies on the East coast of the United States of America.

[2] "Six Degrees of Separation" by John Guare, 1990

[3] "Six Degrees of Separation" directed by Fred Schepisi, 1993

[4] "Six Degrees of Kevin Bacon" by Craig Fass, Brian Turtle and Mike Ginelli, 1994

(1) Only 22% of letters successfully reached the target, an extremely low success rate.
(2) It is possible that the letters that did not reach the target were unconnected, or connected via very long chains.
(3) Participants mailed letters based on their best guess of a shortest path, these guesses could be very wrong.
(4) It is unlikely that the entire human population is acquainted within six degrees of separation because of the existence of certain populations that have had little or no contact with people outside their own culture.

The reason we note these is that one of the primary goals of network models is to develop mathematical models that display the small-world property, and to explore how diseases travel along these models. Therefore, if we disbelieve the six degrees of separation hypothesis then network models may not be a good tool to use during the modelling process.

Network theory explores the mathematical concept of *graphs*. Think of a mathematical graph as a collection of dots connected by a series of lines. Not all dots need to be connected, but to make things interesting there should be at least two dots and one line. Mathematically, the dots are called *nodes* or *vertices* and the lines are called *edges*. A *path* is a way of getting from one node to another node by traveling along the edges. Two nodes are *connected* if a path exists between the two nodes. Finally, the *distance* between two connected nodes is the length of the shortest path connecting the two nodes (if the nodes are unconnected the concept of distance becomes confusing).

Network models *use graphs to describe social or physical contacts between people.*

With this set up we can mathematically chart the social structure of the world. For each individual we create a node and for each relationship we create an edge. That is, if Jane knows Frank on a first name basis we draw a line connecting Jane and Frank. The small-world property states that on this graph of the world, any two nodes can be connected via a path of length six. Of course creating such a graph for the entire world, or even for a small city, would be an intractable task. Instead we rely on network theory to create examples of graphs describing social or physical contact between people. A *network model* is a model that uses such a graph as its underlying structure.

The application of network models to healthcare is a relatively new phenomenon. However, a good deal of research has already been done exploring how the small-world property might impact the spread of disease. Other applications have been slower to arise.

2. Common Uses

Network models are models that describe social or physical interactions between individuals in a society. Once these interactions are described, the main application in healthcare is in describing the spread of disease. This leads to questions like,

- *How do we expect disease to spread through the network?*
- *How can we adjust the network to better control the spread of disease?*
- *Can we determine who to vaccinate to create the maximum effect?*

Other applications include,

- *examining how social networks impact the demand for healthcare,*
- *exploring various strategies for the distribution of healthcare information,*

- *identifying key immunization targets within a community.*

3. Mathematical Details

To understand network theory it is first necessary to discuss the mathematical concept of a graph. (We should note here that due to the new nature of network theory some of the concepts may be more difficult than in previous chapters.) *Graphs* are mathematical objects that are composed of *nodes* (also known as *vertices*) and *edges* connecting them. In healthcare, it is easiest to think of the nodes as individuals and the edges representing connections between individuals, however many other interpretations are possible.

The edges may be given a *weight* that represents the strength of this connection. For example an edge connecting a father and his young daughter might be given a weight of 1 as they see each other every day. Conversely, the same daughter and her classmates might be given a weight of 5/7 since they only see each other five out of every seven days. (In Section 1 we simplified our discussion of graphs by assuming all weights were equal to 1.)

A *path* is a way of getting from one node to another node by traveling along the edges. Two nodes are *connected* if a path exists between these two nodes. If all nodes in the graph are connected then the graph is usually referred to as a *network*. Finally, the *distance* between two connected nodes is the length of the shortest path connecting the two nodes (if the nodes are unconnected the concept of distance is undefined).

A graph *is a collection of* nodes *(also known as* vertices*), some of which are connected by* edges*. If one can trace a path along the edges from any node to any other node, then the graph is called a* network *(or a* fully connected graph*).*

The concepts of paths and distances can be further complicated by the introduction of *directed edges*. A directed edge is an edge that connects nodes in one direction (the easiest analogy is a one way street.) For *directed graphs*, the concepts of connected nodes and distance may become difficult as the distance from node 1 to node 2 may differ from the distance from node 2 to node 1. However, directed graphs can be very useful in certain circumstances. For example, genetic diseases can generally only be passed toward descendants. Directed graphs can also be used as a mathematical representation of System Dynamics models (see Chapter 14).

Given a network (or graph), one can assign a degree to each node in the network. The *degree of a node* is defined as the number of edges exiting the node. In the analysis of networks, often one of the first things done is to bin (i.e. organize) the nodes by their degree and attempt to determine the probability distribution for a node to have a given degree. This provides a first look into the behaviour of the graph, and some information on how paths within the network behave.

Probability and probability distributions are discussed further in Chapter 5.

To elaborate, let us denote the probability that a randomly selected node has a degree k by p_k. We say the network follows a Poisson law if

$$p_k = \frac{\mu^k}{k} e^{-\mu},$$

where the constant μ is determined from the data to fit the model. The network follows a power law if

$$p_k = Ck^{-\alpha} e^{-\frac{k}{\kappa}},$$

where the constants C, α, and κ are determined from data to fit the network model. (These constants do not represent any physical quantities.) Knowing this information provides some insight on how the network behaves. Most importantly, if a real life network is modelled and the probability distribution for the nodes is

created, then this information can be used to generate examples of networks that behave similarly to the real network. This allows interventions to be tested in a more robust manner.

In Figure 1 we illustrate how different probability distributions generate different types of networks.

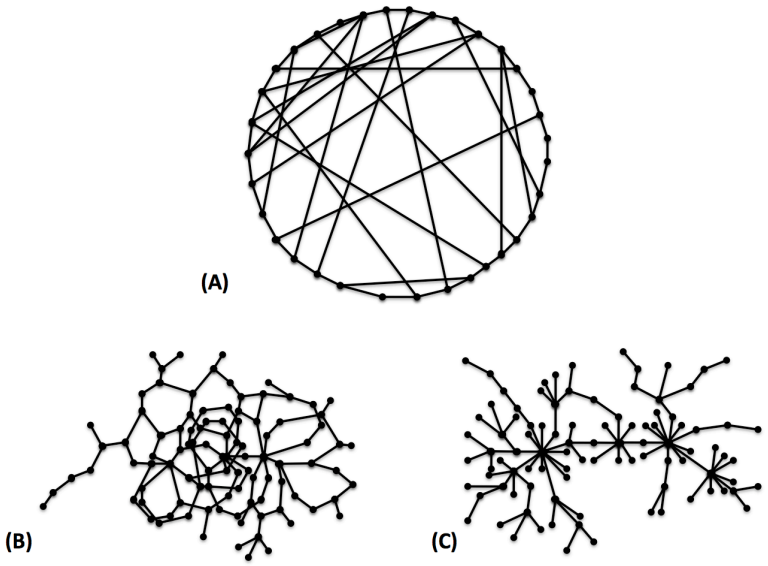

FIGURE 1. **Types of networks:** (A) Small-world network, (B) Power-law network, (B) Poisson-law network.

4. Examples

4.1. Impact of Social Interactions on the Spread of HIV Infection. As mentioned throughout this chapter, one of the key issues in developing a network model is deciding how to generate the network. That is, how to build the vertices (individuals) and the edges (relationships) in order to represent the way in which individuals interact. However, this is only the first step in building most network models. Following this, one usually creates some rules on how interacting individuals influence each other.

For example, building a network that shows all of the relationships in a community is an interesting exercise, but not a model (it does not answer a specific question). Using that network to explore how injection drug users and needle sharing impact the spread of HIV in that community is a model. Dabbaghian, Vasarhelyi, Richardson, Borwein, and Rutherford explored this very question with a simple rectangular network and data based on Vancouver's Downtown Eastside [**55**]. In this example we summarize their work and highlight some of the results therein.

Vancouver's Downtown Eastside is one of the poorest regions in Canada, and contains one of the highest densities of injection drug users. Recently the HIV positive population of the Downtown Eastside has reached what some researchers

refer to as epidemic proportions. One of the major recognized avenues of HIV spread in the Downtown Eastside is needle sharing among injection drug users. In order to understand this issue a little better, Dabbaghian, et. al. sought to create a model of injection drug users. Since traditional models of disease spread based on differential equations, such as the S.I.R. model discussed in Chapter 13, do not typically take into account the impact of social influence, the authors chose to develop a simple network model and overlay a cellular automata simulation.

The network used in the model is the simple rectangular grid. Each vertex represents an individual interacting with its neighbours. Each vertex has exactly four neighbours, lying directly "North", "East", "South", and "West" of the vertex.

Each individual in the model can take on one of five possible states. These are: stayer, susceptible, SIDU, SIDU-HIV, HIV. Individuals labelled as a *stayers* represent those who will never use injection drugs and are therefore immune to contracting HIV (the paper ignores HIV contraction through sexual intercourse). These individuals never change state. Individuals labelled as *susceptible* are not current injection drug users or are injection drug users that do not share needles. Individuals labelled as *SIDU* are needle Sharing Injection Drug Users. Individuals labelled SIDU-HIV are needle Sharing Injection Drug Users that are HIV positive. Finally, individuals labelled HIV are HIV positive individuals who have stopped sharing needles. In Figure 2 we describe how individuals may transition between these five states.

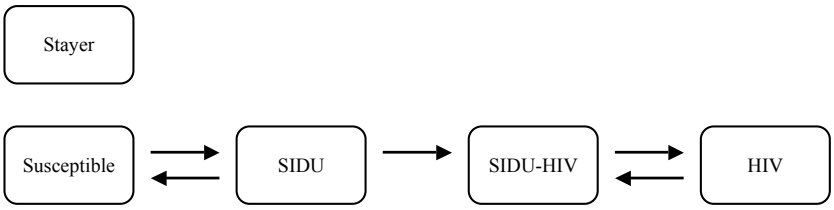

FIGURE 2. **Cellular automata model of HIV spread:** Stayers always remain stayers, but can exert social influence on neighbours that are not stayers. Susceptibles may transition into SIDUs. SIDUs may return to susceptible status, or transition into SIDU-HIV status. SIDU-HIVs may never return to SIDU status, as HIV is incurable, they may transition into HIV status if they stop sharing needles. HIVs may return to SIDU-HIV status if they restart needle sharing.

In the model each individual places a social influence on its four neighbours. The level of this influence is controlled by two parameters labelled α and β. The parameter α represents the ability to discourage sharing needles, provided the individual is not a needle sharer. Conversely, β represents the ability to encourage sharing needles, provided the individual is a needle sharer. Finally, individuals slowly die off, and are replaced with new randomly generated individuals over time.

The paper uses the model to explore several scenarios for interventions to the spread of HIV via injection drug use. First, by setting α and β to 0, they consider the scenario that nobody has any social influence on anybody else. They find in this

case that the HIV prevalence rises quickly for a short time and then crashes as the HIV positive needle sharers die off too quickly to sustain the epidemic. Exploring other values for α and β they are able to find a region in which the epidemic is self-sustaining.

The paper is a good first step is developing a better understanding of how HIV is spread in areas of high injection drug use. According to the authors, future work will examine sexual transmission of HIV, and immigration of already infected individuals.

4.2. Control of Communicable Diseases in Healthcare Facilities. The study of social networks has a long history but it has been largely descriptive. Global properties of the network, or its dynamics, have not been used to make predictions or recommendations until recently. In 2003, work by Meyers, Newman, Martin, and Schrag bridged this gap in the study of the transmission of bacterial pneumonia caused by *Mycoplasma pneuminiae* [**150**]. In this example we examine this work and summarize the results within.

The goal of Meyers et. al. was to develop a model that could be used to determine the size of an epidemic and test different intervention strategies. To do this, they began with a network model consisting of two types of nodes: patients and healthcare workers. The network's edges were directional, indicating the transfer of infection from one person to another. Furthermore, the network was compartmental, reflecting the structure of a healthcare facility. An example of such a model appears in Figure 3.

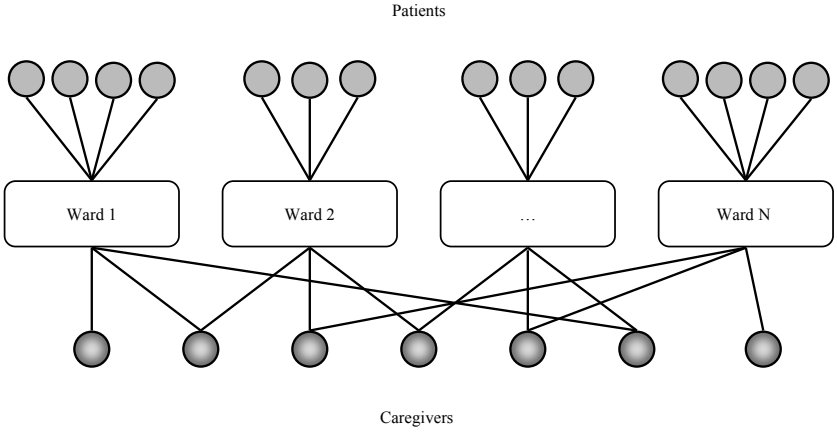

FIGURE 3. **A Network of healthcare facilities:** Reproduced from [**150**].

Next, the model was overlaid with a Markov state model similar to the epidemiological SIR (Susceptible-Infected-Recovered) model. (For further information on Markov models and on the SIR model see Chapter 13.) Accordingly, each node in the model was given one of three health classes: healthy, infective, and previously ill people who are cured and immune. The model then ran along the rule that disease transmission can only occur if there is a direct link between two individuals. By setting various initial conditions and running the model over a series of

time steps, various scenarios for epidemic spread were explored. For example, in its initial stages an epidemic can include a single infected person or several infected persons. The size of the epidemic is defined as the number of nodes in the largest cluster of infected nodes (only one such cluster exists if the epidemic started from a single case).

The model used three independent parameters to define the size of the epidemic: the number of wards where each care-giver works (μ_c), the transmission rates from care-givers to wards (τ_c) and the transmission rates from wards to care-givers (τ_w). While the average number of served wards per worker μ_c is known and can be changed as a control measure, the transmission rates are not observable and should be derived by fitting the model to actual or simulated data. This was done using data from the Centers for Disease Control and Prevention on a mycoplasma outbreak that occurred in a psychiatric institution in 1999. Interestingly, although theoretically a simple Poisson distribution for the probability of a disease transmission should be usable, this did not concur with the data collected, so instead the binomial distribution was used.

Using the model, the authors concluded that the average number of served wards per worker appeared to be the crucial factor in controlling the epidemic. The healthcare workers were found to be the vectors for the spread of the infection. The model demonstrated that limiting the number of wards served by each care-giver and better protection for care-givers is the most effective intervention, even for diseases with long incubation periods, such as pneumonia.

4.3. The Birthrate in Europe from 1950 to 2000. Since the "baby-boom" following World War II, birth rates in Europe have begun a steady decline. Although the reasons for this are unknown, many people have developed hypotheses on the matter. For example, an increased emphasis on education has raised the age of the first time mother, thereby decreasing the total amount of children she might conceive. Other hypotheses focus on the increased cost of child rearing, or the decline of the "stay-at-home mom."

In 2005, Michard and Bouchaud published an interesting network model that demonstrated that the decrease in birthrate behaved in a manner consistent with a model that uses social pressure as one of its driving forces [**151**]. More precisely, Michard and Bouchaud showed that the decline of birthrate demonstrated the behaviour of a *random field Ising model*. In this example we summarize their work and explain some of their results.

In the random field Ising model, agents choose between one of two possible choices, which we shall label as $+1$ and -1 (in our case the choice is whether or not to conceive a child). This choice is influenced by three factors:

(1) personal opinion,
(2) social pressure (the opinions of ones close acquaintances), and
(3) external information.

The model starts by randomly generating a network model in which each node is an individual person and each edge represents a close acquaintance between two people. The model proceeds by randomly setting a personal opinion for each individual (node). Finally, the external public information is a global variable that plays the role of a time-dependent driving force in the decision-making process.

Letting $S_i(t)$ represent the state of node i at time t, we next define the rule

$$S_i(t) = \text{sign}\left(\phi_i + F(t) + \sum_{j \in \mathcal{N}_i} J_{ij} S_j(t-1)\right) \tag{23}$$

where $\mathcal{N}_i$ is the set of friends of agent i, $F(t)$ is the driving force, and ϕ_i is the personal opinion random variable. The function sign takes the value $+1$ if its argument is positive and -1 if it is negative. Finally, we define $S(t) = \sum_i S_i(t)$ as the collective opinion of the model. Our primary interest is now how $S(t)$ changes with respect to time and various driving forces. Equation 23 represents a model

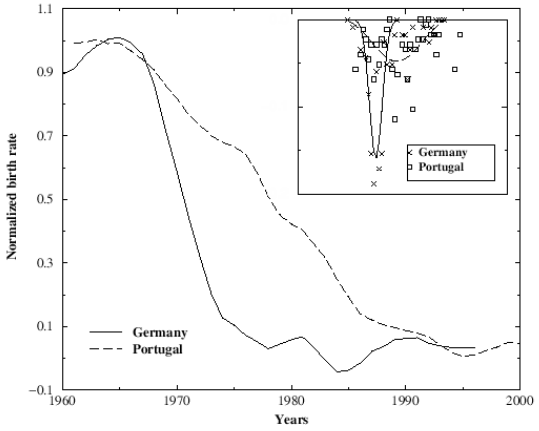

FIGURE 4. **Birth rates of Germany and Portugal:** These curves show the drop-off in birth rate for Germany and Portugal. The drop-off rate for the other countries is intermediate between these two examples. The inset shows the rate of change of the birth rate.
Reproduced from [**151**].

that is driven by the three forces listed above (personal opinion, social pressure, and external information). To test if the real data agrees with this hypothesis we turn to the theory regarding random field Ising models. In particular, in equation (23), the influence of the opinions of friends is given by the coefficients J_{ij}. If these values are near a certain critical value, then collective opinion, $S(t)$, of the society begins by changing slowly if $F(t)$ is increased or decreased. Moreover, if $F(t)$ passes through 0 it will cause $S(t)$ to change rapidly for several time steps, after which the rate of change of $S(t)$ will slow down. In this model, the distribution of the slopes of the curve $S(t)$ forms a peaked curve, similar to a Gaussian distribution. Surprisingly, the height of this peak is related to its width by $h \propto w^{-\frac{2}{3}}$, regardless of the details of F or ϕ [**196**]. Thus, if we wish to test if this is an appropriate model for our problem we should next check that actual data satisfies this property.

Michard and Bouchaud collected data from Eurostat regarding 11 different countries (Belgium, France, Germany, Greece, Italy, Netherlands, Poland, Portugal, Spain, Sweden, Switzerland, and the United Kingdom), and plotted the number of births per woman of child bearing age per year with respect to year. (A sample of

these plots appears in Figure 4.) As can be seen from Figure 4, the birth rate has been falling sharply over time. For each country, the slope of the fecundity curve was calculated at a number of points and the result fitted to a Gaussian distribution. The natural logarithm of the height of each of the peaks was plotted against the natural logarithm of the width in Figure 5. The points cluster remarkably well around a line with slope $-\frac{2}{3}$, demonstrating that the fall in birth rates is consistent with a model of this type. The initial drop in birthrate would have been caused

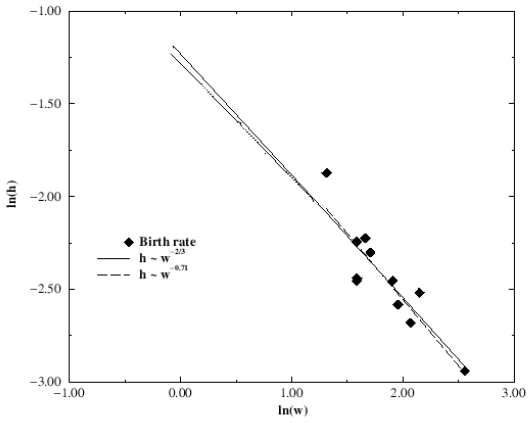

FIGURE 5. **Peak of the Birth Rate Drop-off Rate versus the Width:** The natural logarithm of the peak of the fecundity drop-off rate plotted against the natural logarithm of the width of the fecundity drop-off. A linear regression fit gives a slope of -0.71 ± 0.11. The RFIM predicts a slope of $-\frac{2}{3}$.
Reproduced from [**151**].

by an external factor, such as the availability of birth control pills. However, these results are strong evidence that once the phenomena took root, social pressure became a driving force.

5. Related Reading

Network Models' role in modelling social interactions, give them close associations with Psychosocial Risk Modelling (Chapter 8) and the Health Belief Model (Chapter 10). To a limited degree, graph theory can be thought of as a mathematical tool of use in Markov Models (Chapter 13), System Dynamics (Chapter 14), and Queueing Theory (Chapter 15).

Reference [**17**] investigates queue networks with various classes of customers. Reference [**175**] presents a social organizational strategy framework. Reference [**199**] considers a dynamic social network model in which agents play repeated games in pairings determined by a stochastically evolving social network. Reference [**163**] demonstrates that a large class of standard epidemiological models can be solved exactly for a wide variety of networks. Reference [**143**] examines how complexity theory may offer insight into the behaviour of a population of large-scaled network organizational groups, with a focus on academic medical centers. Reference [**164**] reviews new developments in network systems,

including such concepts as the small-world effect, degree distribution, clustering, network correlations and random graph models. Reference [**214**] suggests the importance of network science and its potential use in ensuring smooth operation of complex networks for the U.S. Army.

Reference [**150**] introduces a network model approach to investigating the spread and control of mycoplasma pneumoniae with a particular focus on the interactions between patients and caregivers in an institution with multiple wards. Reference [**114**] studies several biological systems as networks. Reference [**181**] uses contact network epidemiology to predict the effect of various control policies for mildly and moderately contagious diseases. Reference [**2**] uses a network flow model for long-term bed capacity planning and medium-term care team capacity planning.

References [**196**] and [**197**] examine various aspects of random field Ising models. Reference [**151**] uses a random filed Ising model to look at the collective effects induced by social pressure.

CHAPTER 13

The Future Starts Now

Not the power to remember, but its very opposite, the power to forget, is a necessary condition for our existence. Sholem Asch (1880-1957)

It is singular how soon we lose the impression of what ceases to be constantly before us. A year impairs, a luster obliterates. Lord Byron (1788-1824)

Markov Models

1. Model Overview

To begin this chapter, let us consider a classical model of disease spread. In this model, we examine a collection of individuals that can take on one of three states: susceptible, infected, and recovered. If an individual in the model is susceptible, then he or she has a probability of becoming infected during the next time step. This probability is not based on any demographic factors of the individual, but instead based on the number of infected currently in the model. In particular, the more people that are currently infected, the more likely it is that a susceptible individual comes in contact with an infected individual, and thereby becomes infected. If an individual in the model is infected, then he or she has a probability of becoming recovered during the next time step. This probability is fixed, and based on the standard recovery rate of an infected person (for whatever disease one is studying). Finally, if an individual in the model is recovered, he or she will remain in the recovered state during the next time step. A visual of this model is given in Figure 1.

The simple model described by Figure 1 is often called the S.I.R. model and is a classic example of a Markov model in healthcare. In particular, the model examines a collection of *objects* that are assigned a series of *states* over a period of *time steps*. Moreover, at the end of each time period, the objects move from one state to the next according to transition probabilities (also called transition rates) that depend only on the current state of the system. These are the key aspects of a Markov model. Markov models are models that examine a collection of objects in a system that move through a series of states. Moreover, Markov models work on the assumption that the future state of the object is determined by a random process dependent only on the current state of the system. This assumption is so basic to the methodology of Markov models that it is generally referred to as the *Markov assumption*. Due to the Markov assumption, Markov models are "forgetful" in the sense that a knowledge of the past states of the system is not required to predict

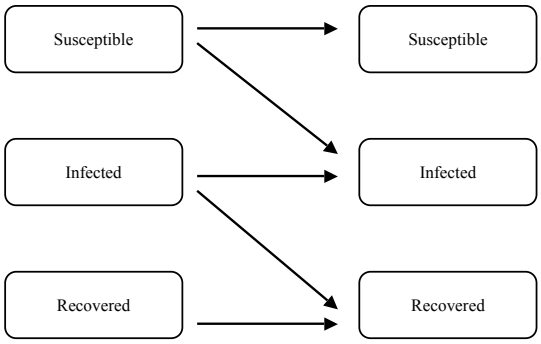

FIGURE 1. **The S.I.R. model of disease spread:** The classical S.I.R. model (Susceptible, Infected, Recovered) of disease spread can easily be viewed as a Markov model.

the future. In spite of this, Markov models can exhibit deep structure through the cumulative effects of repeated stochastic events.

The Markov assumption *is that the future state of an object in the model is determined by a random process dependent only on the current state of the system.*

Before implementing a Markov model it is important to determine the list of all possible states that an object can take. The collection of all possible states that an object can take is called the *state space*. In more complicated models, the list of possible states becomes longer, and the manner in which objects move from one state to another becomes more complex. For example, the objects may be people and the states may be the amount and type of healthcare services each person uses during a given time period. This would result in an extremely large number of possible states, and therefore a rather complicated model to implement.

The simplest type of Markov model is th *finite state Markov chain*. In such models the number of possible states is finite, and transition from one state to the next occurs at predefined points in time. Such models are relatively simple to implement, but can still model some surprisingly complex systems. As such they are often a good first choice for many modelling problems.

Many people believe that the Markov assumption causes Markov models to be extremely limited in application. For example, your probability of gaining weight is partly dependent on your current weight, but also partly dependent on your history of weight gain (see Example 4.2). Even though the Markov assumption forces some level of forgetfulness on the models, it is nonetheless possible to build memory into a Markov model. The way this is done is to create new states that incorporate the memory for the desired trait. For example, we might create a state called "currently normal, but was obese". Markov models that incorporate memory in this manner are sometimes referred to as *higher order Markov models*. If the model incorporates one level of memory it is referred to as a 2^{nd} order Markov model (or a Markov chain of order 2), and models that do not incorporate any memory are sometimes called 1^{st} order Markov models.

Whenever we choose to approach a problem via Markov models, it is of the utmost importance to test whether the Markov assumption is valid.

Whenever we choose to approach a problem via Markov models, it is of the utmost importance to test whether the Markov assumption is valid. Fortunately there is a simple method to test the Markov assumption. Basically, we build a

higher order Markov model and check that its results agree with the original lower order model. If the Markov assumption holds, the two models will produce the same results (see Figure 2, page 134).

There are many generalizations to finite state Markov models, including infinite state models, Markov processes, semi-Markov processes, and Markov Decision Processes. These generalizations are discussed briefly at the end of Section 3, but are beyond the scope of this book.

2. Common Uses

Markov models explore the properties of objects in a system that move through a series of states. The most important aspect of Markov models is the assumption that the system satisfies the Markov assumption: the future state of the object is determined by a random process dependent only on the current state of the system. In healthcare, this assumption is well suited to modelling the movement of patients through disease states. In regards to disease states, Markov models are suitable to answer questions such as:

- *How do immunization rates impact the spread of disease through a population?*
- *How many people will be affected by diabetes in future years?*
- *At what disease state is treatment most suitable to prevent disease spread?*

Aside from modelling disease states, Markov models are also useful for examining if patient history is a factor in behaviour:

- *How does doctor-patient loyalty affect the use of the healthcare system?*
- *To what level does an individual's past BMI status impact their future BMI status?*
- *Does surgeon skill affect patient progress through post-surgery recovery stages?*

3. Mathematical Details

In Markov models we begin with a collection of objects and a list of possible states for each object. For example, the object may be individual people that can take one of two states: healthy or sick. The collection of all possible states that an object can take is called the *state space*. At each time interval, the model assigns every object in the system to exactly one state from the state space. At the end of each time period, the objects move from one state to the next according to *transition probabilities* (or *transition rates*) that depend only on the current state of the system. That the transition probabilities depend only on the current state of the system is the key aspect of Markov models, and generally referred to as the *Markov assumption*.

Due to the Markov assumption, Markov models are "forgetful" in the sense that a knowledge of the past states of the system is not required to predict the future. In spite of this, Markov models can exhibit deep structure through the cumulative effects of repeated stochastic events.

3.1. Finite State Markov Chains. We begin our discussion with the simplest type of Markov models, *Finite State Markov chains*. The words "finite state" mean exactly what one would suppose them to mean; that the list of possible states

for an object is finite. The final word, "chain", refers to the assumption that transition from one state to the next occurs at predefined points in time. For example, in examining the spread of disease we might decide to update each individual's state at the end each day. Alternately, states might be updated on an irregular, but still predefined basis. For example we might be interested in studying an individual's BMI status when they turn 16, 19, 25, 50 and 65. Whether time periods are evenly spread or not makes no difference in the mathematics required to analyze the model.

Let $S = \{s_1, s_2, \ldots, s_i, \ldots, s_N\}$ be the state space of the model (the list of all possible states that an object can take). Let X^0 be a column vector of length N that represents the initial state of the system. That is

$$X^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_i^0 \\ \vdots \\ x_N^0 \end{bmatrix}$$

where x_i^0 is the number of objects in state i at time step 0. In general we shall use

$$X^t = \begin{bmatrix} x_1^t \\ x_2^t \\ \vdots \\ x_i^t \\ \vdots \\ x_N^t \end{bmatrix}$$

where x_i^t is the number of objects in state i at time step t.

Next, let $\Pr^t(i \to j)$ be the probability of an object moving to state s_j at time $t+1$ given that the object was in state s_i at time t. Creating the matrix

$$P^t = \begin{bmatrix} \Pr^t(1 \to 1) & \Pr^t(2 \to 1) & \cdots & \Pr^t(N \to 1) \\ \Pr^t(1 \to 2) & \Pr^t(2 \to 2) & \cdots & \Pr^t(N \to 2) \\ \vdots & \vdots & \ddots & \vdots \\ \Pr^t(1 \to N) & \Pr^t(2 \to N) & \cdots & \Pr^t(N \to N) \end{bmatrix}$$

we find that the state vector for the system at time $t+1$ is the matrix multiplication of P_t and the state vector of the system at time t:

$$X^{t+1} = P^t X^t. \tag{24}$$

The matrix P^t is generally referred to as a *transition matrix*, and may be dependent on the time t or the state of the system at time t.

In order for a Markov chain model to run correctly, transition matrices must satisfy several special properties. First, all elements of the matrix must be non-negative. This stops objects from flowing backwards through the model. Second, each column of the transition matrix must sum to 1. This prevents objects from

appearing or disappearing from the model (models that allow entry into the model, or exit from the model are discussed in Subsection 3.2).[1]

The matrix P^t is generally referred to as a transition matrix.

In particular, the state at time t can be found via the formula

$$X^t = P^{t-1}P^{t-2}\dots P^1P^0X^0. \tag{25}$$

If the transition probabilities do not change over time, that is if $P^t = P^0$ for all t, then the Markov model is called a *time homogeneous*. Time homogeneous Markov models allow for obvious simplifications to formula (25):

$$X^t = [P^0]^t X^0.$$

3.2. Markov Model Involving Entry and Exit. In the previous subsection, we discussed Markov models that did not allow new objects to enter the model, nor allow objects from the model to exit the model. In some cases one wishes to incorporate this idea into the model (for example to represent births or deaths). This can be accomplished in several manners. The easiest manner is to create two new states, one labelled "to-be-born" one labelled "dead". Next set the initial number of objects in the state "to-be-born" to a very high number (say 1000 times the size of the model), and set the initial number of objects in the state "dead" to zero. Finally, one adjusts the transition matrices P^t to allow for objects to be born into the appropriate states, and allow for objects from the appropriate states to transition into the "dead" state.

There are other methods of incorporating birth and death into Markov models, and in general they are quite straightforward. For example, one can rewrite the state-vector equation (equation (24)) to $X^{t+1} = P^tX^t + b^t - d^t$, where b^t is a vector of objects born into each state at time t and d^t is a vector of object exiting each state at time t. The disadvantage of this method, is one must take care that the total number of objects in a given state never becomes negative. In particular, one should never have more objects exit a state, than there are objects in that state. In the first method discussed, this is taken care of automatically when we check that each column of the transition matrix must sum to 1.

An additional advantage of the first method, is that it allows us to keep track of how many objects enter and exit the system over the course of the model. (The number entering the system is the difference of the initial number and final number of objects in the state "to-be-born", the number exiting the system is the final number of objects in the state "dead".)

3.3. Higher Order Markov Models. On the surface, the Markov assumption appears to create models which are extremely limited in application. For example, if we were modelling the spread of disease through a population, then we would be interested in the two states "uninfected" and "infected." However, it is well known that patients who have recovered from a virus are less likely to become infected again from the same disease. Therefore, if the infected state simply feeds back into the uninfected state, the model is unlikely to provide useful information.

Even though the Markov assumption forces some level of forgetfulness on the models, it is nonetheless possible to build memory into a Markov model. The way this is done is to create new states that incorporate the memory for the desired trait. For example, in the case of modelling spread of disease through a population,

[1] A few texts consider transition matrices as the transposition of the above approach; in this case, the sum of each row totals to one.

one could create states labelled "susceptible," "infected," and "recovered." The recovered state now effectively contains the memory that the individual was once infected.

Markov models that incorporate memory in this manner are sometimes referred to as *higher order Markov models*. The order of the Markov model is one more than the level of memory the model attempts to incorporate. For example, if the model incorporates one level of memory it is referred to as a 2^{nd} order Markov model (or a Markov chain of order 2), and models that do not incorporate any memory are sometimes called 1^{st} order Markov models. The order of a Markov model is qualitatively descriptive only. That is, since the higher order models can always be dealt with by adding additional states to the model, higher order Markov models can mathematically be dealt with in the same manner as first order Markov models.

3.4. Testing the Markov Assumption. Higher order Markov models provide us with insight on how to test if the Markov assumption is suitable for a given problem. The basic idea is that if the Markov assumption holds, then building memory into the model via higher order models should have no effect on the transition probabilities. These ideas are clarified in Figure 2.

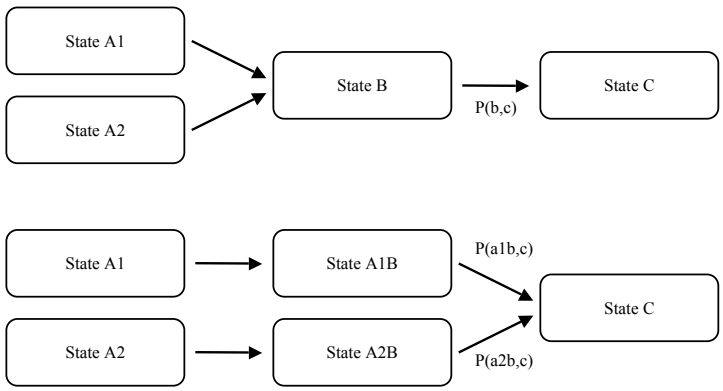

FIGURE 2. **Testing the Markov assumption:** Testing the Markov assumption
One method to test if the Markov assumption holds is to turn a 1^{st} order Markov model into a 2^{nd} order Markov model and check if the transition probabilities are affected. In the 1^{st} order model above (top) we have two states, $A1$ and $A2$, which feed into a state B, which then feeds into state C. To create a second order model (bottom), we expand state B into two states: state $A1B$ and state $A2B$. If the probability of moving from $A1B$ to C is the same as the probability of moving from $A2B$ to C (i.e. $P(b,c) = P(a1b,c) = P(a2b,c)$) then the Markov assumption holds, and a 1^{st} order model suffices. Otherwise, one should test if the 2^{nd} order model satisfies the Markov assumption.

3.5. Infinite State Markov Models. For Markov chains with a finite number of states, the transition probabilities may be represented as a transition matrix (see Subsection 3.1). If the number of states is infinite, then this property does not apply. Instead, the transitions must be described in terms of functions. Recall, for finite chains we used $\Pr^t(i \to j)$ to represent the probability of an object moving to state s_j at time $t+1$ given that the object was in state s_i at time t. If there are an infinite number of states, the indices i and j may no longer be integers, and so building the matrix P^t is no longer possible. Instead one creates a function

$$f(i, j, t) = \Pr^t(i \to j).$$

Various mathematical techniques have been developed to study such functions, most of which focus on the question of whether there is a state s_i that has a high probability of being occupied regardless of starting conditions. These techniques are beyond the scope of this book.

3.6. Markov Processes and Semi-Markov Processes. The Markov models discussed above were Markov chains, meaning that all state transitions occur at fixed predefined time intervals. In the 1920s, a more general class of models, called *Markov processes*, in which transitions occur at arbitrary times was also developed. In these models, time is viewed as a continuous variable, so time steps can occur at any point. One classic example of such a process is the "random walk of a drunkard," in which a point stumbles in a random direction for a random distance. In this case, the concept of time is incorporated into distance, and so the point can be thought to be traveling in a random direction for a random length of time. In literature, the random walk of a drunkard is usually referred to as a *Wiener process* or *Brownian motion*.

Another generalization of Markov chains are *semi-Markov processes*. In a semi-Markov process, the transition probabilities depend not only on the current state of the system, but also on the time that it has spent in that state. The time that the system spends in each state is assumed to vary stochastically according to a probability distribution. Semi-Markov models have wide applicability in queueing theory, reliability modelling, and operations research. Recently, they have been applied to the modelling of chronic diseases, such as HIV.

3.7. Markov Decision Processes. Markov Decision Processes (MDPs) are an extension of Markov Processes and Markov Chains, in which the modeller is allowed to interact with the objects in the system by applying *actions* to the system. Applying an action to the system can be thought of as altering the transition matrix for a selection of time steps. MDPs are used extensively in business to help examine the effect of decision making in situations where outcomes are partly random and partly under the control of the decision-maker. The basics of MDPs are not difficult, once the ideas behind Markov models are understood. However, further discussion on MDPs are beyond the scope of this book.

4. Examples

4.1. A Simple Doctor-Patient Loyalty. To demonstrate the mathematics behind a simple time homogeneous Markov chain, consider a drop-in clinic with three doctors. In this particular drop-in clinic, no appointment is necessary, so patients may not see the same doctor on every visit. However, the patient (when

returning to the clinic) may request a specific doctor. If the doctor is available that day, the patient's wait time increases but considerations are usually made.

We assume that a patient's preference for a doctor is completely determined by the doctor that they visited in their last visit, and the random factor of when that doctor will be available. The probabilities of visiting a given doctor, given the doctor seen during the previous visit, is found in Table 1. Notice that some doctors inspire more patient loyalty than others.

Previous Visit	Next Visit Sees Doctor 1	Next Visit Sees Doctor 2	Next Visit Sees Doctor 3
Saw Doctor 1	0.72	0.09	0.21
Saw Doctor 2	0.18	0.85	0.15
Saw Doctor 3	0.10	0.06	0.64

TABLE 1. Transition probabilities for a hypothetical Doctor-Patient Loyalty model.

Suppose the clinic has 300 patients that return on a regular basis, and we wish to see how these patients impact each doctor's work load. We begin by assuming that each doctor will see 100 of these patients, so

$$X^0 = \begin{bmatrix} 100 \\ 100 \\ 100 \end{bmatrix}.$$

The transition matrix for this Markov chain will be unchanging with time, and equal to

$$P = \begin{bmatrix} 0.72 & 0.09 & 0.21 \\ 0.18 & 0.85 & 0.15 \\ 0.10 & 0.06 & 0.64 \end{bmatrix} \text{ for all } t.$$

Simple matrix-vector multiplication yields

$$X^1 = \begin{bmatrix} 102 \\ 118 \\ 80 \end{bmatrix},\ X^2 = \begin{bmatrix} 100.86 \\ 130.66 \\ 68.48 \end{bmatrix},\ X^3 = \begin{bmatrix} 98.7594 \\ 139.4878 \\ 61.7528 \end{bmatrix}, \ldots,$$

$$X^{20} = \begin{bmatrix} 89.6613 \\ 158.9361 \\ 51.4026 \end{bmatrix}, \ldots,\ X^{50} = \begin{bmatrix} 89.6414 \\ 158.9641 \\ 51.3944 \end{bmatrix}, \ldots,\ X^{100} = \begin{bmatrix} 89.6414 \\ 158.9641 \\ 51.3944 \end{bmatrix}.$$

This might lead us to conjecture that in the long run Doctor 1 will have approximately 90 patients, Doctor 2 will have approximately 159 patients and Doctor 3 will have approximately 51 patients. To confirm this more mathematically, we are interested in the *equilibrium distribution* of X:

$$\lim_{t\to\infty} X^t = \lim_{n\to\infty} P \times P \times P \times \cdots \times PX^0 = \lim_{n\to\infty} [P]^n X^0.$$

To proceed we must *diagonalize* the matrix P. To diagonalize a matrix P is to find an invertible matrix Q and a diagonal matrix D such that $P = Q\,D\,Q^{-1}$. This can be done quickly and easily with modern mathematical computing software. In this case diagonalizing P provides

$$P = Q\,D\,Q^{-1}$$

where

$$Q \approx \begin{bmatrix} -0.583 & -0.867 & -0.481 \\ 0.820 & 0.154 & -0.854 \\ -0.237 & 0.713 & -0.276 \end{bmatrix} \text{ and } D \approx \begin{bmatrix} 0.680 & 0 & 0 \\ 0 & 0.531 & 0 \\ 0 & 0 & 1.00 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \lim_{n\to\infty} P^n X^0 &= \lim_{n\to\infty} QDQ^{-1}QDQ^{-1}\dots QDQ^{-1}X^0 \\ &= \lim_{n\to\infty} QD^nQ^{-1}X^0 \\ &= Q \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} Q^{-1}X^0 \\ &\approx \begin{bmatrix} 0.30 & 0.30 & 0.30 \\ 0.53 & 0.53 & 0.53 \\ 0.17 & 0.17 & 0.17 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 100 \end{bmatrix} = \begin{bmatrix} 90 \\ 159 \\ 51 \end{bmatrix} \end{aligned}$$

This mathematically confirms our predictions for this model. More importantly, if we repeat this analysis with an arbitary

$$X^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix}$$

with $x_1^0 + x_2^0 + x_3^0 = 300$ notice that

$$\begin{aligned} \lim_{n\to\infty} P^n X^0 &= \begin{bmatrix} 0.30 & 0.30 & 0.30 \\ 0.53 & 0.53 & 0.53 \\ 0.17 & 0.17 & 0.17 \end{bmatrix} \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix} = \begin{bmatrix} 0.30(x_1^0 + x_2^0 + x_3^0) \\ 0.53(x_1^0 + x_2^0 + x_3^0) \\ 0.17(x_1^0 + x_2^0 + x_3^0) \end{bmatrix} \\ &= \begin{bmatrix} 0.30(300) \\ 0.53(300) \\ 0.17(300) \end{bmatrix} = \begin{bmatrix} 90 \\ 159 \\ 51 \end{bmatrix} \end{aligned}$$

So regardless of initial conditions, the resulting distribution of patients will be the same.

This model illustrates a key feature of many Markov models, that the model will eventually approach an equilibrium or "steady-state". This means that the distribution of patients among the physicians will eventually approach an equilibrium distribution, which is *independent of the initial distribution*. In this case, 30% of the patients will see Doctor 1, 53% will see Doctor 2, and 17% will see Doctor 3.

4.2. Testing the Markov Assumption for an Male BMI State Model. In this example we consider a simple 3 state Markov model examining changes in BMI status.[2] The portion of the survey data we use consists of following 5316 males over the course of 18 years, and collecting their BMI value ($BMI = mass\ in\ kg/(height\ in\ m)^2$) every two years. In this example we seek to check if the Markov assumption is a valid assumption when studying an individuals BMI status.

[2]The model was designed and tested against the National Longitudinal Survey (NLS) database, which is freely available at `http://www.bls.gov/nls/`.

The basic model consists of the states: "Normal-weight," "Over weight," and "Obese," which we shall abbreviate as: nw, ow, and ob respectively. The states correspond to BMI ranges of

$$\begin{aligned} nw &\Leftrightarrow BMI < 25 \\ ow &\Leftrightarrow 25 \leq BMI < 30 \\ ob &\Leftrightarrow 30 \leq BMI. \end{aligned}$$

(Note, under weight individuals ($BMI < 18.5$) are placed in the nw state, this is due to the lack of under weight individuals in the NLS database.) Since we will tune the model using NLS data, we will use a two year time step between transitions. To allow for easier reading of notation, we shall create a time step index i which will correspond to the age of the individuals at that time step. For example, time step 20 will represent the time at which individuals are aged 20 or 21, time step 22 will represent the time at which individuals are aged 22 or 23, etc. As two years is a relatively long time we will assume that any state can transition into any other state over each time step. Visually this produces the model shown in Figure 3.

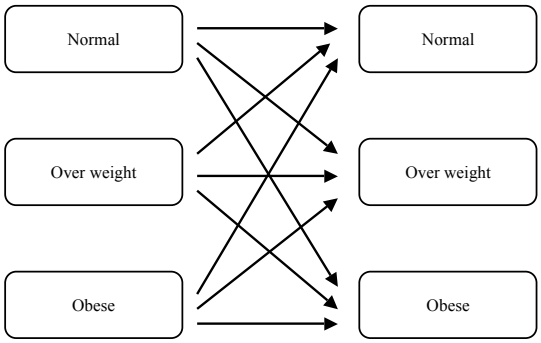

FIGURE 3. **A 3-state Markov model of BMI status:** A visualization of the basic BMI state Markov model.

Let

$$X_i = \begin{bmatrix} nw_i \\ ow_i \\ ob_i \end{bmatrix}$$

be a vector of the number of individuals in each state at time i, where nw_i is the number of normal weight individuals, ow_i is the number of over weight individuals, and ob_i is the number of obese individuals. To determine the number of individuals in each state during time step $i+1$ we multiply X_i by a transition matrix T_i where

$$T_i = \begin{bmatrix} \Pr_i(nw|nw) & \Pr_i(nw|ow) & \Pr_i(nw|ob) \\ \Pr_i(ow|nw) & \Pr_i(ow|ow) & \Pr_i(ow|ob) \\ \Pr_i(ob|nw) & \Pr_i(ob|ow) & \Pr_i(ob|ob) \end{bmatrix}$$

and $\Pr_i(y|x)$ is the probability that at time step i an individual will move to BMI category y given that their current BMi category is x. We let these probabilities be time dependent to represent that fact that as individuals age they will begin

to behave differently. Calculating these probabilities for each time period is easily done by examination of the NLS data set. Doing so, one finds that at the 30th time step

$$T_{30} = \begin{bmatrix} 0.785 & 0.102 & 0.005 \\ 0.211 & 0.793 & 0.145 \\ 0.004 & 0.105 & 0.850 \end{bmatrix}.$$

In order to test the Markov assumption, we next produce a higher order Markov model and check if it behaves the same as the first order model. By this we mean that we change our 3 state model into a 9 state model, where each state stores not only your current obesity status but also your obesity status from the previous time step. In order to facilitate discussion we shall label the 9 states $s_1, s_2, \ldots, s_9$ where

$$\begin{array}{lll} s_1 = \{nw \to nw\} & s_2 = \{ow \to nw\} & s_3 = \{ob \to nw\} \\ s_4 = \{nw \to ow\} & s_5 = \{ow \to ow\} & s_6 = \{ob \to ow\} \\ s_7 = \{nw \to ob\} & s_8 = \{ow \to ob\} & s_9 = \{ob \to ob\} \end{array}$$

and state $\{x \to y\}$ represents that one time step earlier the individual was in state x and currently the individual is in state y. These 9 states can transition according to the logical rule that the next state's x must be the current state's y. If

$$\widetilde{X}_i = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_9 \end{bmatrix}$$

represents the current state of the model the the transition matrix $\widetilde{T}_i$ would take the form

$$\widetilde{T}_i = \begin{bmatrix} \Pr_i(s_1|s_1) & \Pr_i(s_1|s_2) & \Pr_i(s_1|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_2|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_3|s_9) \\ \Pr_i(s_4|s_1) & \Pr_i(s_4|s_2) & \Pr_i(s_4|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_5|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_6|s_9) \\ \Pr_i(s_7|s_1) & \Pr_i(s_7|s_2) & \Pr_i(s_7|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_8|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_9|s_9) \end{bmatrix}.$$

If the Markov assumption holds then each row of $\widetilde{T}_i$ will consist of three (nearly) identical values as there should be no statistically significant difference between the behaviour of individuals in s_1, s_2 and s_3 (for example). However, using the NLS data set one can calculate that $\widetilde{T}_{30}$ would be

$$\widetilde{T}_{30} = \begin{bmatrix} 0.826 & 0.399 & 0.000 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.214 & 0.071 & 0.041 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.000 & 0.017 & 0.000 \\ 0.173 & 0.577 & 0.333 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.763 & 0.819 & 0.514 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.250 & 0.274 & 0.084 \\ 0.001 & 0.024 & 0.667 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.023 & 0.110 & 0.445 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.750 & 0.709 & 0.916 \end{bmatrix}.$$

As one can see, the Markov assumption does not hold (for example $\Pr(s_7|s_1) = 0.001$, but $\Pr(s_7|s_3) = 0.667$). Therefore a basic Markov model is not an appropriate approach to modelling obesity. We could continue by building higher and higher order Markov models until the Markov assumption held, but this is likely an unrewarding task. (On a final note, there is nothing special about examining time step 30. Examining other time steps shows the same disagreement between the data and the Markov assumption.)

4.3. A Mover-Stayer Model for Problematic Drug Use: A Compartmental Markov Model. Standard Markov models assume that the same transition matrices apply uniformly to the entire population. However, it is often the case in epidemiological modelling that the population is divided into different compartments according to their susceptibility or infectiousness. The model then describes how the size of these compartments changes over time by means of equations describing the disease dynamics. A common approach to compartmental modelling is to use a mixed Markov process, which consists of a superposition or mixture of different Markov chains with independent transition matrices. In this example we review a compartmental mixed Markov model used to analyze heroin addiction [**188**].

The basic model views the population as consisting of two types of people. Those who are susceptible to becoming problematic drug users and those who are "prudent" and hence not at risk of becoming drug users. Thus the compartmental Markov model consists of two subgroups, which we will call "movers" and "stayers." The *movers* represent the people who are susceptible to drug abuse, and can therefore move about the various states of drug use. Conversely, *stayers* represent the people who are not at risk of becoming drug users.

In the model of [**188**], movers can become drug users either by coming in contact with other drug users or by contact with drug dealers. The diagram in Figure 4 is a compartmental representation of the model. In this figure, the straight lines represent possible state changes of the movers, and the curves represent the possible interactions between drug users and people susceptible to drug use (movers) in the population.

As shown in the Figure 4, this model uses several compartments to model the phenomenon. If a mover becomes a drug user, then they initially pass through a phase of light use. After a period of light use, they then move on to a phase of heavy but invisible use. When usage becomes problematic, they become visible as heavy drug users and begin their interaction with the healthcare system and social services. Finally, addictive use of drugs leads to possible reform and a possible recidivist phase.

In order understand the spread of drug addiction and build a model that can be used to test the effectiveness of different types of interventions, such as treatment programs or law enforcement, we use the diagram in Figure 4 to write a set of eight coupled difference equations that describe the evolution of the system (see Table 2). In evolving the system, it is assumed that the lengths of stay in each of the compartments are exponentially distributed. To implement the model, a computer program is written to evaluate the difference equations. Since the computer program evaluates the difference equation using a series of predetermined time steps, this is mathematically a Markov model.

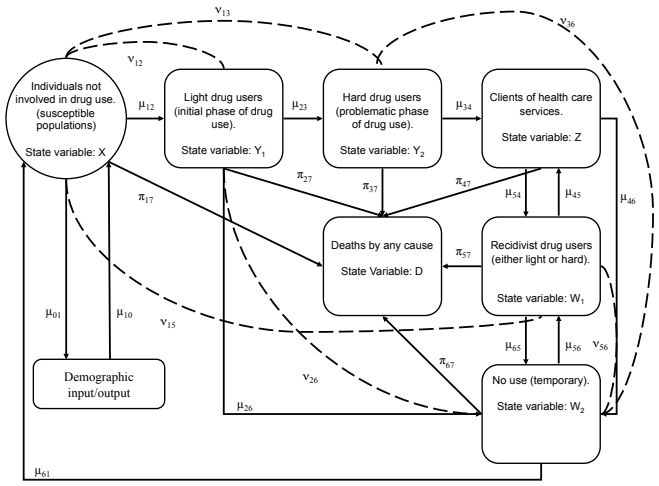

FIGURE 4. **System dynamics diagram of mover-stayer model epidemic drug use:** The parameters μ_{Ij} represent flow of movers from one state to other, the parameters ν_{ij} represent interactions between the different components in the model, and the parameters π_{i7} represent mortality from each of the components. Reproduced from [**188**].

In order to implement the model, it is necessary to obtain values for the various parameters in the model. In this model, μ_{01}, μ_{10} and π_{17} are demographic parameters, which may be obtained from census data. The parameters μ_{23} and μ_{34} are parameters describing the prevalence of problematic drug use and may be obtained from studies of the incidence of drug use. The parameters π_{27} through π_{67} may be obtained from studies of the mortality rate among drug users. The parameters $\mu_{45}, \mu_{46}, \mu_{54}, \mu_{56}, \mu_{65}$ and μ_{61} are the most difficult to obtain; however they may be estimated from therapy data. (In [**188**], values for these parameters are estimated for heroin use in Italy from 1980 to 2000.)

Using this model, the authors explored the effects of both primary and secondary preventive interventions. A primary intervention is one that is applied directly to the susceptible population. It is said to have an effectiveness P, if a proportion P of the movers in the susceptible population become stayers. The effect of secondary preventive interventions can be evaluated by modifying the ν and μ parameters. For example, the consequence of increased law enforcement would primarily be to decrease the parameter μ_{12}. Safe injection sites would primarily have an effect on the parameters ν_{56} and μ_{56}. The impact of healthcare policies on drug use would primarily be on the parameters μ_{45} and μ_{54}.

The model supports the statement that primary interventions are more effective than secondary interventions. However, there is substantial latency in the system after a program of primary intervention is initiated. That is, there would be no sign of any positive response for a significant period of time. In the case of the model applied to heroin drug addiction in Italy, this response latency would likely

$$
\begin{aligned}
X(t+\Delta t) &= X(t) + (-\mu_{01} + \mu_{10} - \pi_{17})X(t) \\
&\quad - (\mu_{12} + \nu_{12}Y_1(t) + \nu_{13}Y_2(t) + \nu_{15}W_1(t))(1 - S(t))X(t) \\
&\quad + \mu_{61}W_2(t) \\
Y_1(t+\Delta t) &= Y_1(t) + (-\mu_{23} - \mu_{26} - \pi_{27})Y_1(t) \\
&\quad + (\mu_{12} + \nu_{12}Y_1(t) + \nu_{13}Y_2(t) + \nu_{15}W_1(t))(1 - S(t))X(t) \\
Y_2(t+\Delta t) &= Y_2(t) + (-\mu_{34} - \pi_{37})Y_2(t) + \mu_{23}Y_1(t) \\
Z(t+\Delta t) &= Z(t) + (-\mu_{54} - \mu_{46} - \pi_{47})Z(t) + \mu_{34}Y_2(t) + \mu_{45}W_1(t) \\
W_1(t+\Delta t) &= W_1(t) + (-\mu_{45} - \mu_{65} - \pi_{57})W_1(t) \\
&\quad + (\mu_{56} + \nu_{26}Y_1(t) + \nu_{36}Y_2(t) + \nu_{56}W_1(t))W_2(t) + \mu_{54}Z(t) \\
W_2(t+\Delta t) &= W_2(t) + (-\mu_{61} - \pi_{67})W_2(t) \\
&\quad - (\mu_{56} + \nu_{26}Y_1(t) + \nu_{36}Y_2(t) + \nu_{56}W_1(t))W_2(t) + \mu_{26}Y_1(t) + \mu_{46}Z(t) \\
D(t+\Delta t) &= D(t) + \pi_{17}X(t) + \pi_{27}Y_1(t) + \pi_{37}Y_2(t) + \pi_{47}Z(t) + \pi_{57}W_1(t) + \pi_{67}W_2(t) \\
S(t+\Delta t) &= S(t)\frac{(1 - \mu_{10} - \pi_{17})X(t)}{X(t+\Delta t)} + S_0\frac{\mu_{01}X(t) + \mu_{61}W_2(t)}{X(t+\Delta t)}
\end{aligned}
$$

$X(t)$	size of the susceptible population at time t
$S(t)$	proportion within the susceptible population who are "stayers" at time t
$S_0(t)$	proportion of the new population entering the susceptible population who are stayers at time t
$Y_1(t)$	population of light drug users at time t
$Y_2(t)$	population of hard drug users at time t
$Z(t)$	population whose drug use has made them known to the healthcare system at time t
$W_1(t)$	recidivist drug users at time t
$W_2(t)$	temporary holding population for users in transition at time t
$D(t)$	number of deaths at time t (cumulative)

TABLE 2. Coupled difference equations represented by Figure 4.

be about 6 years and possibly as long as 8 years. However, when the system does respond to the intervention it does so rapidly and in a highly non-linear fashion. This result is important, as many intervention programs would be abandoned as "failures" if no improvement was seen for 5 years.

However, it should be noted although the model supports the statement that primary interventions are more effective than secondary ones, this model does not evaluate the cost of primary versus secondary interventions. For example, the model predicts a greater effect if the parameter μ_{12} is adjusted slightly than if the parameter ν_{56} is adjusted slightly. However, it does not evaluate how difficult it is to adjust each parameter. In application one must determine how much effort is required to adjust each parameter and weight this against the impact of the adjustment.

5. Related Reading

Markov models are closely related to Systems Thinking and System Dynamics (Chapter 14), as well as Queueing Theory (Chapter 15). Indeed, Markov models provide an alternate approach to developing and implementing many of the ideas in those methodologies. Markov models are often used as a base for other models to build on. In this manner, Markov models are often used in conjunction with Network Models (Chapter 12), Game Theory (Chapter 11), and many statistical models (Part 2).

Reference [**188**] contains details for Example 4.3.

Reference [**90**] looks at the mover-stayer Markov model using various estimators and at the accuracy of these estimators. Reference [**161**] expands previous work on a Markov model for predicting future need for resources by taking varying utilization rates between age groups into account. Reference [**46**] uses a Markov model to track the movement of patients through the disease states of malaria. Reference [**142**] develops a Markov chain for the analysis of a centralized medical record system in a large hospital. Reference [**210**] derives exact HIV incubation distributions under treatment by anti-viral drugs based on first passage probability distributions for some continuous time Markov chains. Reference [**147**] uses a Markov model to describe movements of geriatric patients within a hospital system. Reference [**113**] uses Markov modelling to forecast the number of people with diagnosed and undiagnosed diabetes by age, race, ethnicity and sex. Reference [**27**] uses Markov models to analyze various social science processes, such as bed occupancy rates, brand loyalty, occupational mobility, voter patterns and malaria. Reference [**51**] proposes a method for analyzing surveillance data for communicable pathogens using a "structured" hidden Markov model.

Reference [**203**] uses a Markov decision process to examine the cost-effectiveness of alternative screening strategies for HCV infection in comparison with no screening. Reference [**117**] presents a Markov decision process model to evaluate different screening policies for breast cancer. Reference [**198**] presents a Markov decision process for addressing the unresolved issue of optimal time to initiate HIV therapy.

CHAPTER 14

Viewing the System as a Whole

Imagination is the beginning of creation. You imagine what you desire; you will what you imagine; and at last you create what you will. George Bernard Shaw (1856-1950)

No man is an island entire of itself; every man is a piece of the continent, a part of the main. John Donne (1572-1632)

System Dynamics and Systems Thinking

1. Model Overview

In conventional management thinking, a problem can be broken down into individual parts, and each of these is optimized separately. In healthcare, this approach is failing as strong interactions exist between various parts of the healthcare system. To deal with this problem, many researchers are beginning to apply *systems thinking* to the field of healthcare, and turning to *system dynamics* to help quantify their models.

Systems thinking is more a style of thought than an actual modelling technique. Overall, systems thinking can be viewed as a form of qualitative modelling that focuses on viewing the system as a whole instead of a collection of individual parts. This method is strongly based on the belief that the components of a system will act differently when united than when separated. It argues that by viewing the system as a whole, fundamental insights can be gained, and that persistent difficulties can be resolved by studying the system as a single entity.

Although the name may appear to be new, many policy makers will already be familiar with systems thinking. Generally systems thinking models are the type of models that show up in boardrooms during talks on "how X impacts Y." Often, but not always, they are best visualized as a collection of boxes connected by arrows and influence signs. Each box represents a part of the system, and each arrow and influence sign combination represents how that box impacts another box in the system. A positive sign means that an increase in the box from which the arrow leaves causes an increase in the box into which the arrow arrives. A negative sign means that an increase in the box from which the arrow leaves causes a decrease in the box into which the arrow arrives.

As an example, consider Figure 1, which shows how new medicines can both save lives and increase the chance of medicinal error. In this figure, we have four boxes: "New Medicines," "Lives Saved," "Medicinal Errors," and "Staff Training." Since new medicines can save lives, there is an arrow with a positive sign going from "New Medicines" to "Lives Saved." That is, an increase in the number of new medicines will cause an increase in the number of lives saved. Similarly, the model

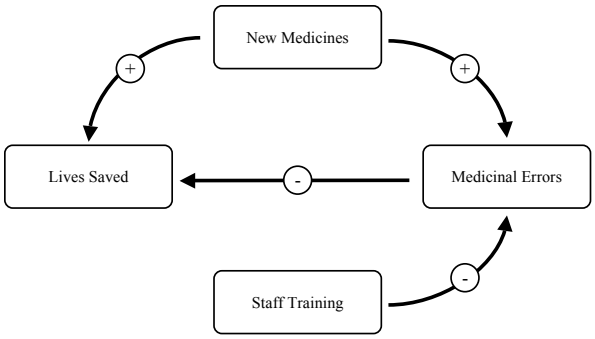

FIGURE 1. **Systems thinking model relating new medicines to medicinal errors:** New medicines can provide better treatment for diseases, thereby saving lives. However, new medicines also increase confusion amongst hospital staff, which increases the chance of medicinal errors. This decreases the number of lives saved (note the negative on the arrow connecting box "Medicinal Errors" and "Lives Saved"). Finally the model suggests that staff training could be used to counter this problem.

suggests an increase in the number of new medicines leads to an increase in the possibility of making a medicinal error, which in turn leads to a decrease (note the negative sign) in the number of lives saved. To counter this effect, one must use staff training to decrease the possibility of medicinal errors. The model suggests that in order for new medicines to have their maximum impact, their introduction must be accompanied by staff training.

In order to quantify systems thinking, one can turn to system dynamics models. First developed in the early 1960s, system dynamics models are based on the economics concepts of *stocks* and *flows*. To understand stocks and flows it is useful to think of the analogy of water flowing through a series of reservoirs and pipes. As the valves on the pipes open and close, the water flows from reservoir to reservoir in a different pattern. To connect this with our previous systems thinking models, each box in the systems thinking model is a reservoir and the arrows connecting the boxes become the pipes. Figuring out the exact equations to describe the flow through a given pipe given a state of the system is how the model becomes quantified.

System dynamics models (at least from the perspective of this book) are defined by their use of stocks and flows to describe feedback loops and complex systems.

The strength of system dynamics modelling lies largely in the fact that it is both qualitative and quantitative in nature. The qualitative nature of system dynamics comes from the fact that it builds upon a systems thinking model approach. Thus the original model can be created without using data, but solely focusing on the qualitative nature of the system. This is best obtained based on the experience and insights of professionals and managers. This knowledge base, although qualitative, is the foundation of the actual decision process in the system, and as such is considered more comprehensive. (A positive side effect is that beginning modelling

via systems thinking usually makes the model more understandable and believable later.)

The model shifts to being quantitative when equations that describe each arrow in the model are added. The equations should be developed and tested using actual data, which gives the model mathematical accuracy. The equations usually reduce to a complex system of non-linear differential equations. It is possible (but unlikely) that these equations can then be solved analytically. More likely, a numerical differential equation solver or discrete time computer simulation is employed to provide testing to various system scenarios.

System dynamics models pose a few drawbacks to modelling in healthcare. Most notably, system dynamics treat individuals in the system like water. It can remember where a patient is coming from, and where a patient is going to, but it cannot distinguish between two patients in the same reservoir, nor can it remember a patient's entire past history. In some cases, such as modelling wait times for surgery, these details are significant, in most other cases they are not required and are usually lost in the mass of the system. In cases where the time a patient has waited in a given spot (i.e. the ability to distinguish patients in a reservoir) is important, queueing theory models are probably best to employ, see Chapter 15. In cases where the entire history of a patient is important, discrete event simulation is probably the best option, see Chapter 9.

2. Common Uses

Systems thinking is a holistic approach to modelling, based on the belief that the components of a system will act differently when isolated from the system. It argues that, by viewing the system as a whole, fundamental insights can be gained. Using these ideas, systems thinking approaches questions of how various parts of the system interact. For example,

- *What factors are driving the changes in hospital operating budgets?*
- *How is the shift in population age demographics going to impact the healthcare system?*
- *How will a pandemic affect the healthcare system?*

can all be discussed via systems thinking.

System dynamics is the natural choice for quantifying the ideas developed in a systems thinking model. In healthcare, system dynamics has been successfully employed to solve problems such as,

- *How can one implement intervention strategies for better control of disease?*
- *Would hiring more cleaning staff alleviate the hospital bed crisis?*
- *How do privately run healthcare facilities impact the need, demand, and use of publicly run facilities?*

Unfortunately, this versatility often makes it difficult to approach system dynamics via analytical means, so numerical analysis techniques usually become necessary.

3. Mathematical Details

3.1. Systems Thinking. *Systems thinking* is method of thought as opposed to an actual modelling technique. It is a form of qualitative modelling that focuses

on viewing the system as a whole, instead of a collection of individual parts. The goal is to describe how the various parts of a system interact qualitatively.

Often systems thinking models are best visualized as a collection of boxes connected by arrows and influence signs. Each box represents a part of the system, and each arrow and influence sign combination represents how that box impacts another box in the system. A positive sign means that an increase in the box from which the arrow leaves causes an increase in the box into which the arrow arrives. A negative sign means that an increase in the box from which the arrow leaves causes a decrease in the box into which the arrow arrives.

Systems thinking models can be created and viewed in many other ways than the influence diagrams described above. As one of the major goals of any systems thinking model is to clarify interactions between various parts of a system, any visual or descriptive model that accomplishes this suffices. If the model does not accomplish this, one should "rethink" the system.

An example of a systems thinking model that employs the influence visualization method can be found in Figure 1. An example of a systems thinking model that does not employ the influence visualization is detailed in Example 4.1.

3.2. System Dynamics. After a systems thinking approach is employed, and an influence diagram is created, a modeller often wishes to quantify their model in order to be able to make predictions regarding policy changes and future work loads. To do this, we often turn to system dynamics. *System dynamics* models are defined by their use of stocks and flows to describe feedback loops and complex systems, therefore in order to understand system dynamics we must begin with the definitions of stocks and flows.

The term *stock* is derived from the business concept, which refers to the value of an asset at a balance date. In a more general sense a stock is better described as an entity that is accumulated over time by inflows and/or depleted over time by outflows. In healthcare, the entity in question is often patients, so a stock might measure the number of patients in a hospital emergency department, the number of patients infected with a specific disease, or the number of patients who require home care nursing. Returning to systems thinking, stocks measure the contents of the boxes in an influence diagram.

The term *flow* is also derived from concepts in economics, and refers to the total value of changes to a stock during a given period. Thus in healthcare, flows could represent the number of patients entering and exiting a hospital emergency department, the number of patients becoming infected or recovering from a specific disease, or the number of patients who degrade to the point where they need home care nursing minus the number of patients who improve (or degrade) to the point where they no longer need home care nursing. Returning to systems thinking, in a system dynamics model of an influence diagram flows measure the amount of influence an arrow has on a given box.

Of course stocks and flows can measure objects other than patients. In fact, one of the strengths of systems modelling is that stocks and flows can measure anything that is quantifiable. The number of beds in a hospital, the number of washrooms in use at a given time, and the number of staffed ambulances sitting idle at a given time could all be modelled using stocks and flows. In short, if you can measure it, you can model it.

With stocks and flows defined, and an influence diagram created, the next step in creating a system dynamics model is to determine the equations that govern each stock. These equations generally rely on the time t and the state of the system $S(t)$ at that time. For example, consider the influence diagram in Figure 1. To flesh this out into a full system dynamics model we make some small changes to the model to develop the diagram shown in Figure 2.

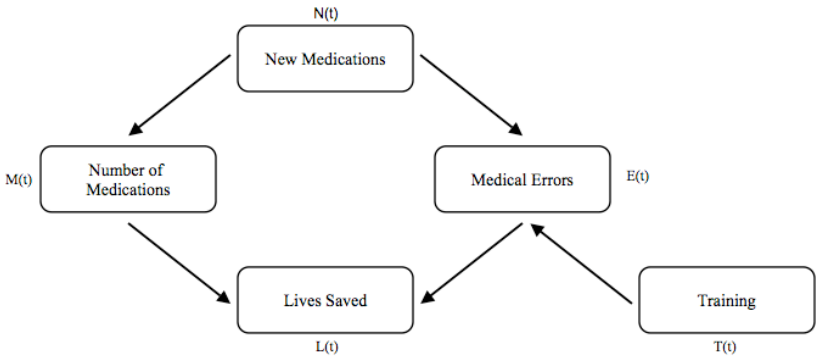

FIGURE 2. **System dynamics model relating new medicines to medicinal errors:** A reconstruction of the systems thinking model in Figure 1 as a system dynamics model.

In Figure 2 the function $N(t)$ represents the number of new medicines introduced during the month t. This function could be created by the user to test certain scenarios, or determined using historical data to predict future trends.

The function $T(t)$ is the number of hours of staff training provided in month t. Like $N(t)$, this number is chosen by the user to test various training strategies, or set based on historical training strategies.

The function $M(t)$ represents the number of medicines employed at the hospital during month t. Since the change in the number of medicines employed per month is the number of new medicines introduced, we have

$$\frac{d}{dt}M(t) = N(t). \tag{26}$$

Recall, $\frac{d}{dt}f$ is the derivative of the function f with respect to time, that is, the change in the function f over a given time.

The function $E(t)$ represents the number of medicinal errors in a given month. This number is increased as new medicines arrive and decreased as staff training takes effect. For the sake of example, we assume both of these effects are linear. That is, doubling the time spent training per month doubles the effect of training on the number of errors. (This assumption is somewhat unrealistic, but it makes the math achievable without use of a computer. Plus, as long as the number of drugs introduced and the amount of training done does not fluctuate too much, a linear approximation is reasonably accurate.) This leads to the differential equation

$$\frac{d}{dt}E(t) = \alpha N(t) - \beta T(t), \tag{27}$$

where α and β are positive constants.

Finally, the function $L(t)$ is the total number of lives saved from the start of the model until month t. Each month, L increases with the number of medicinal

options available, and decreases by the number of medicinal errors. For the sake of example, assume each medicinal option has the potential to save γ lives. That is

$$\frac{d}{dt}L(t) = \gamma M(t) - E(t). \tag{28}$$

Equations (26), (27) and (28) combine to form a second order system of ordinary differential equations, dependent on the input functions $N(t)$ and $T(t)$. To see this, differentiate equation (28) to obtain

$$\frac{d^2}{dt^2}L(t) = \gamma \frac{d}{dt}M(t) - \frac{d}{dt}E(t),$$

then use equations (26) and (27) to replace $\frac{d}{dt}M(t)$ and $\frac{d}{dt}E(t)$ to obtain

$$\frac{d^2}{dt^2}L(t) = \gamma N(t) - (\alpha N(t) - \beta T(t)).$$

Integrating this twice we obtain the solution

$$\begin{aligned} L(t) &= \int_0^t \int_0^t ((\gamma - \alpha)N(t) + \beta T(t))\, d\tau_1 d\tau_2 \\ &= (\gamma - \alpha) \int_0^t \int_0^t N(t) d\tau_1 d\tau_2 + \int_0^t \int_0^t \beta T(t)) d\tau_1 d\tau_2. \end{aligned}$$

If $N(t)$ and $T(t)$ are simple functions (such as constants) these integrals are easily evaluated. If they are more complicated functions, one may have to resort to numerical solvers to complete the solution.

The above example demonstrates much of the mathematics required to develop and solve system dynamics models. Of course the above model did not include any feedback loops, hence the differential equations were simple to solve. If the model contains feedback loops, then the differential equations become more challenging, and often can no longer be solved by analytic methods. In these cases it is usually very useful to lean on the growing selection of system dynamics software available. Some of this software is reviewed in Appendix A.

4. Examples

4.1. Going Solid: The Danger of Lacking Wiggle Room. In 2005 Cook and Rasmussen published an exercise in systems thinking that examines how hospital operating budget and staff workload might impact the safety margins for practitioners of the healthcare system [**50**]. In this example we summarize their thinking, and provide some further interpretation.

Many hospitals have begun operating as what Cook and Rasmussen refer to as a "solid system." This system includes the practice of admitting patients into surgery based on the event that a bed will become available by the time surgery is complete, not based on the event that a bed is currently available. On the surface, this sounds like an excellent system; surgeries can begin sooner, and therefore more surgeries can be performed. However, consider the following real event:

> *Patient A was admitted into surgery based on the fact that patient B would leave the recovery room before surgery was completed. Unfortunately, patient B did not leave the recovery room, because the bed he was supposed to move to was still occupied by patient C. Patient C was scheduled to move from the intensive care unit to the regular ward, but was stopped by patient D who was still occupying the desired bed. Patient D was actually released from*

the hospital, but was still occupying the bed as his transportation had not arrived. The transportation was delayed due to an accident on the freeway.

As one can see from the example, this particular hospital had adopted the "solid system" approach to management. Moreover, they had adopted the approach at all levels of the hospital, making them (in theory) highly efficient. Unfortunately, without the wiggle room created by a less efficient system, patient A was left lying on the operating room table, occupying highly expensive space, people, and equipment. The question we would like to answer is, why are hospitals becoming "over efficient"?

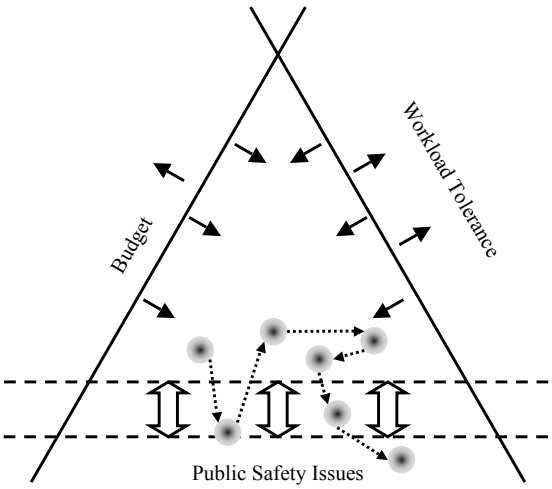

FIGURE 3. **Factors impacting hospital operating procedures:** Budget, workload tolerance, and public acceptance each play a role in hospital operation procedures. Budget and workload tolerance are strong factors, constantly pushing operating procedures (the grey dots) towards unsafe practices. Public acceptance levels waver over time, and generally only arise when accidents become frequent enough to arouse public interest. In the long run this may result in the public becoming accustomed to unsafe operation procedures.
Based on the work of [**50**].

Consider three of the major factors that affect how a hospital operates: *budget*, *workload tolerance*, and *public acceptance*. The hospital's operating budget is clearly a factor in how a hospital operates. The next factor, the staff's workload tolerance, is based on the fact that if staff are overworked (or perceive themselves to be overworked) then they tend to neglect performing tedious duties in order to focus on what they consider more important. Finally, public opinion is of concern, as if the hospital has too many accidents then they will be penalized due to public outcry. For example, leaving a patient on an operating room table long after the surgery is complete is generally considered bad for the hospital's image.

Now consider these three factors as sides of a triangle, and the actual operating state of the hospital as a point inside of the triangle (see Figure 3). Each factor is exhibiting a force on the operating point, making it quiver inside of the triangle. The magnitude of this quivering is determined by variations in the three forces impacting the operating point. Since the forces produced by the budget and workload tolerance are fairly steady, the operating point is constantly pushed towards the public acceptance boundary. Conversely, the force from the public acceptance is not steady, as it only arises when accidents become frequent enough to arouse public interest. As a result, the operating point may exist very close to the public acceptance boundary, and may even occasionally cross the boundary with no repercussions. By flirting with the margin of public acceptance, the public becomes hardened to accepting a greater number of accidents and the public acceptance boundary is loosened without inciting public outcry.

This simple model demonstrates how the constant pushes of budget and workload tolerance are overwhelming the standard operating proceedings, forcing hospitals into states that appear more efficient. It also provides some insight on how to combat this problem, in the idea that the public acceptance force must be made less erratic. How to alter this force could be explored further by systems thinking, or by developing psychosocial models on how the public views hospital practices (see Chapter 10 for more information on psychosocial modelling).

4.2. A System Dynamics model of yo-yo dieting. Recently researchers have begun to refer to obesity as the "global epidemic" of the 21st century. Obesity rates all over the world are on the increase. From a medical view-point obesity and weight gain are associated with increased risk for a number of diseases (such as diabetes, sleep apnoea, heart disease, and gallbladder disease). From a social view-point, being over-weight or obese is still considered socially unacceptable. For both of these reasons, many individuals turn to dieting in an attempt to lose weight. Indeed, according to a survey by Goldbeter, about 25% of men and nearly 50% of women reported trying to lose weight in 1985 [**88**]. However, dieting is often accompanied by repeated bouts of weight loss and regain, a phenomenon known as weight cycling or "yo-yo dieting." In 2006, Goldbeter developed and studied a simple system dynamics model that examines the dynamics of human weight cycling [**88**]. In this example we provide a short overview of the model examined in that paper.

The paper begins with the development of a qualitative model to examine human weight cycling, which studies three variables and how they interact. In particular the model incorporates an individual's *weight*, *dietary intake*, and *personal resolution to lose weight*. We shall label these as follows:

- W = Weight (Goldbeter uses P),
- I = Dietary Intake (Goldbeter uses Q), and
- R = Personal Resolution to lose weight (Goldbeter uses R).

The qualitative model has these three factors interacting as shown in Figure 4.

In order to quantify the model, Goldbeter begins by normalizing the variables I and R between 0 and 1. He then suggests that the following system of differential

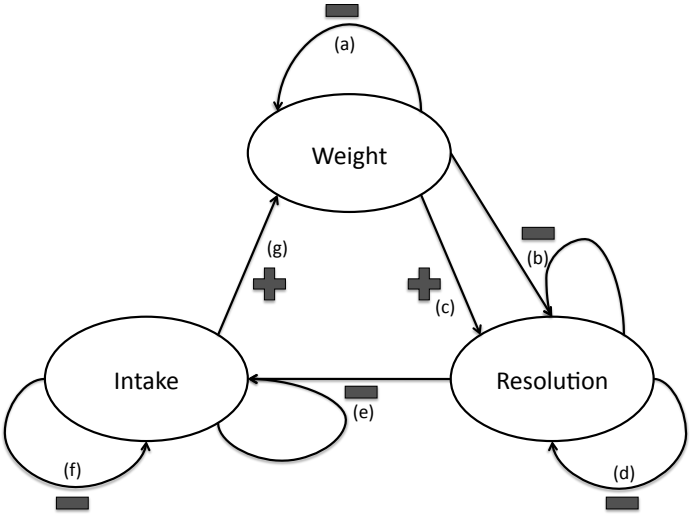

FIGURE 4. **Systems thinking model examining three factors in human weight cycling:** arrow (a) states that an increase in weight causes a decreases the tendency to gain weight; arrow (b) states that an increase in weight and resolution causes a decrease in the tendency to gain resolution to lose weight; arrow (c) states that an increase in weight causes an increase in resolution to lose weight;arrow (d) states that an increase in resolution causes a decrease in the tendency to gain resolution to lose weight; arrow (e) states that an increase in resolution and intake causes a decrease in the tendency to increase intake; arrow (f) states that an increase in intake causes a decrease in the tendency to increase intake; arrow (g) states that an increase in intake causes an increase in the tendency to gain weight
Based on the work of [**88**].

equations might be used to analytically capture the interactions in the model.

$$\begin{aligned}\frac{dW}{dt} &= a_1 I - a_2 \frac{W}{K_1 + W} \\ \frac{dI}{dt} &= b_1 \frac{1 - I}{K_2 + (1 - I)} - b_2 R \frac{I}{K_3 + I} \\ \frac{dR}{dt} &= c_1 W \frac{1 - R}{K_4 + (1 - R)} - c_2 \frac{R}{K_5 + R}\end{aligned}$$

where $a_1, a_2, b_1, b_2, c_1, c_2, K_1, K_2, K_3, K_4, K_5$ are unknown constants. (Recall $\frac{df}{dt}$ is the rate of change of the function f with respect to time.) The first of these equations captures that the rate of change of an individual's weight ($\frac{dW}{dt}$) is positively

correlated with intake and negatively correlated with weight. The second equation captures that the rate of change of an individual's intake ($\frac{dI}{dt}$) is negatively correlated with intake and negatively correlated with the product of intake and resolution. The final equation captures that the rate of change of an individual's resolution ($\frac{dR}{dt}$) is positively correlated with weight, negatively correlated with resolution, and negatively correlated with the product of weight and resolution.

Although complicated in appearance, the above system of equations is not difficult to solve with the aid of a computer (once values for $a_1, a_2, b_1, b_2, c_1, c_2, K_1, K_2, K_3, K_4, K_5$ are selected). Indeed, using a standard numeric ODE solver we can sketch the solution to the system of equations over time. In Figure 5 we provide plots of $W(t)$, $I(t)$, and $R(t)$, for the case of $a_1 = 0.1, a_2 = 0.1, b_1 = 1, b_2 = 1.5, c_1 = 6, c_2 = 0.75, K_1 = 0.2, K_2 = 0.001, K_3 = 0.001, K_4 = 0.001$, and $K_5 = 0.001$, with initial conditions $W(0) = 0.05$, $I(0) = 0.9$, and $R(0) = 0.02$. (Note, these values were chosen arbitratily.Also, although $W(t)$ represents weight, it should not be thought of as kilograms or pounds, but instead as an arbitrary weight measurement).

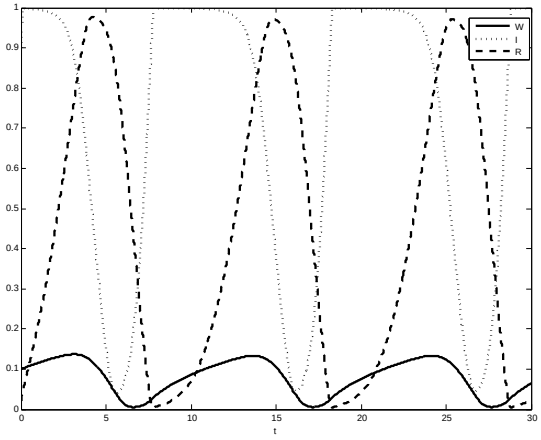

FIGURE 5. **Human weight cycling plots suggested by Goldbeter's model:** These plots show how $W(t)$, $I(t)$, and $R(t)$ change over time when $a_1 = 0.1, a_2 = 0.1, b_1 = 1, b_2 = 1.5, c_1 = 6, c_2 = 0.75, K_1 = 0.2, K_2 = 0.001, K_3 = 0.001, K_4 = 0.001$, and $K_5 = 0.001$, with initial conditions $W(0) = 0.05$, $I(0) = 0.9$, and $R(0) = 0.02$. (Note that although $W(t)$ represents weight, it should not be thought in terms of a specific scale, such as kilograms or pounds.)
Based on the model of [**88**].

Examining Figure 5 we see how even a simple model using the tools of system dynamics can capture some interesting behavior regarding weight cycling. In particular, notice that the weight (W) slowly grows to until a threshold is met and then a sudden decrease occurs. This slow growth in weight is accompanied by a growth in the resolution to lose weight (R). When the resolution to lose weight hits

a certain threshold, the individual's dietary intake (I) suddenly decreases, driving weight loss.

Goldbeter's model does contain some flaws, but these may be corrected in time. Most notably, Goldbeter's model shows a weight cycling that is not accompanied by a slow increase in weight as the individual ages. This contradicts the fact that an individual's weight tends to increase over their life-span. Nonetheless Goldbeter's model provides some insight on how system dynamics can be used to study the question of human weight cycling.

4.3. A Special Issue in System Dynamics Review. The System Dynamics Society is a nonprofit organization devoted to the development and use of systems thinking and system dynamics around the world. Its membership spans over fifty countries and includes researchers, consultants, and practitioners in the corporate and public sectors. For those who study or apply systems thinking and system dynamics it is considered one of the most prestigious sources of knowledge on the subject. The society publishes several newsletters and journals. Foremost amongst them is the "System Dynamics Review" which publishes peer-reviewed articles on systems thinking and system dynamics and their applications to societal, technical, managerial, and environmental problems. In 1999 the System Dynamics Review published a special issue devoted to the use of system dynamics in healthcare. This issue provides an excellent sampling of the variety of places where system dynamics can be applied in modelling in healthcare. In this example we provide a brief review of the papers that appear in the 1999 issue of System Dynamics Review (volume 15 issue 3).

The first three articles in the special issue examine the question of patient wait time. In the first paper, González-Bustoa and García create a system dynamics simulation to study waitlists in Spain [**89**]. Their model is significant as it shows how interactions between two types of patients, elective surgery and emergency surgery patients, affect each other's wait time through the joint use of resources. Their model also incorporates several outside factors, such as surgeon income expectations (in Spain a fee-for-service system is used), to create an elasticity of demand. In the second paper, Ackere and Smith examine waitlists in the UK National Health Service [**217**]. Unlike González-Bustoa and Garciá's work, Ackere and Smith focus only on elective surgery patients. However, their model also examines the interactions between the demand for surgery, resources for surgery, and average wait time for surgery, as well as incorporates elasticity of demand. In the third paper, Wolstenholme examines patient flow for UK National Health Service across all levels of service (from general practitioner to residential care) [**227**]. Wolstenholme uses his model to explore how various system interventions (such as a 20% increase in hospital bed capacity) will impact patient wait time for each type of service. As all three of these papers point out, these system dynamics approaches to modelling wait time allow for the modelling system feedback that is not possible using the traditional queueing theory method.

In the fourth paper in the special issue, Dangerfield and Roberts use system dynamics (and optimization) to study the question of incubation periods for AIDS [**58**]. Their model is based on earlier work by the same authors [**56**], and the paper examines how optimization methods can be used to determine good data distributions for the model.

The last three articles in the special issue are written by individuals who are working in the healthcare industry, instead of academics doing healthcare research. As such the papers are much more applied in nature. In the fifth article, Royston, Dost, Townshend and Turner outline how system dynamics has been used by the Department of Health in England [**190**]. The paper covers a variety of examples, such as developing policies for disease screening and emergency care, and in each case explains how system dynamics provided the necessary modelling tools to approach the questions of interest. In the sixth paper, Hirsch and Immediato discuss what they feel are the three major challenges faced by the U.S. healthcare industry, and how system dynamics can be used to address these challenges [**107**]. The three challenges they note are: the transition from a fee-for-service to a capitated payment system; the shift from autonomous providers to integrated delivery of care; and an expanded definition of healthcare to include health improvement and prevention rather than a narrow focus on treatment of illness. In the final paper, Cavana, Davies, Robson, and Wilson discuss how system dynamics modelling played a role in the New Zealand Ministry of Health determining the factors that interact to drive quality of care in the health and disability sector [**43**]. In this case, system dynamics provided a communication tool for economists, policy makers, and clinicians to develop a shared mental model of the complex system that drives the quality of care.

5. Related Reading

Systems thinking produces qualitative models that are the basis of many other modelling techniques. Psychosocial Models (Chapter 10), such as the Health Belief Model, apply the ideas behind systems thinking to the setting of human behaviour. Models developed through systems thinking are often implemented as system dynamics models. System dynamics models bear strong connections to Markov Models (Chapter 13) and queueing theory models (Chapter 15). Descriptive statistics (Chapter 5) and regression analysis (Chapter 6) are often required to tune system dynamics models.

Reference [**180**] explores new approaches to issues in clinical practice and organizational leadership using complex adaptive systems. Reference [**75**] examines how managers' actions shape the dynamic characteristics of organizations. References [**76**], [**77**], and [**78**] provide some personal recollections on the development of system dynamics by J.W. Forrester.

Reference [**57**] is an introduction to the special issue of the System Dynamics Review discussed in Example 4.3. References [**89**], [**217**], [**227**], [**58**], [**190**], [**107**] and [**43**] also appear in that issue. Reference [**56**] provides the background to the model results discussed in reference [**58**].

Reference [**59**] reviews some system dynamics models used to address European healthcare issues and possible future roles of these models. Reference [**108**] reviews strengths and difficulties in the application of system dynamics to health care. Reference [**153**] looks at the background of system dynamics modelling and its uses in modelling public health. Reference [**112**] reviews system dynamics modelling in healthcare, with a focus on chronic disease prevention. Reference [**205**] looks at how systems thinking and modelling can enhance the ability to generate and learn from evidence and use this in promoting effective changes in public health policy.

Reference [**136**] describes the components of an emergency and urgent care system within one health authority and suggests ways in which patient flow and system capacity could be improved. Reference [**140**] outlines a systems thinking approach called qualitative politicized influence diagrams that may be useful in managing the general practice system

behaviour. Reference [**169**] argues that it is possible to focus on the structural complexity of system dynamics models to design a partition strategy that maximizes the test points between the model and the real world. Reference [**226**] focuses on the use of system archetypes in system dynamics modelling. Reference [**228**] describes a series of system dynamics challenges and explores the issue of mismatch in organizations between process and boundary structure. Reference [**109**] introduces a conceptual framework applying a system dynamics approach to rural disaster preparedness and planning.

Reference [**16**] develops a system dynamics model that suggests that there is a need for a separation of responsibility into medical care of individual patients and prevention/population health, as they require different organizational structures. Reference [**34**] uses system dynamics to model emergency and on-demand healthcare in Nottingham, England. Reference [**111**] presents results from a system dynamics simulation used to model a hypothetical community in poor health and suffering from various intertwined afflictions. Reference [**133**] formulates a system dynamics model to explore the factors that contribute to long wait times for urgent admissions to acute care in hospitals. Reference [**134**] discusses a system dynamics model of an accident and emergency department. Reference [**186**] reports on the use of simulation modelling for redesigning specimen collection centers at a medical diagnostic laboratory. Reference [**50**] examines a dynamic model of risk and safety and its applications in healthcare.

Reference [**155**] compares discrete-event simulation and system dynamics as methods for simulating the dynamics of systems. Reference [**209**] looks at discrete-event simulation and system dynamics in the context of supply chain.

CHAPTER 15

Dealing with Lines and Capacity

People nowadays like to be together. Not in the old-fashioned way of, say, mingling on the piazza of an Italian Renaissance city, but, instead, huddled together in traffic jams, bus queues, on escalators and so on. It's a new kind of togetherness which may seem totally alien, but it's the togetherness of modern technology. James Graham Ballard (1930-)

Affairs are easier of entrance than of exit; and it is but common prudence to see our way out before we venture in. Aesop (circa 600 BC)

Queueing and Traffic Models

1. Model Overview

The question of providing timely access to healthcare is of great interest to many policy-makers. From a humanitarian point of view, it is the people's duty is to promote human welfare. From the political point of view, the ability to provide a minimum level of healthcare to whomever is in need is a source of pride for many countries. However, from the business perspective, there is generally insufficient budgets to provide everyone with all the healthcare they desire. To help determine operating budgets, and locate "bottle-necks" in the system, modellers often turn to queueing theory and traffic models.

Queueing theory and *traffic models* are mathematical models that describe the dynamics of objects arriving, waiting in a queue, and then being served by a server.

Mathematically, the two forms of modelling are essentially the same, the key difference lying in what outcome we are trying to measure. In queueing theory models, we focus on understanding wait times within the system, while in traffic models we focus on locating points of congestion within the system. Since wait times are generally determined by the point of most congestion, and congestion points are defined as points in the system where an object spends an inordinate amount of time, it is clear that the two problems are intimately intertwined.

In developing queueing theory or traffic models, we begin with a system and a collection of objects that seek to enter and exit the system. In healthcare, one of the best examples of queueing theory is modelling wait lists for surgery and one of the best examples of applying traffic theory is in modelling bed counts in hospitals. In the queueing example, wait list modelling, the system one wishes to enter (and eventually exit) is the operating room and the objects that wish to enter the system are the patients waiting for surgery. In the traffic example, bed count models, the system we wish to study is the number of available beds in a hospital, and again

the objects wishing to enter the system are patients (typically post-surgery). The similarities between these two examples is obvious. In fact the only real difference is that in waitlist modelling we are interested in how long an individual patient has to wait in order to enter the system, while in bed count modelling we are interested in the number of beds in use at any given moment. That is, in the first we are interested in the properties of the objects, while in the second we are interested in the properties of the system. Since the interactions between the objects and the system are what determines these properties, it is clear that the study of queueing theory and traffic theory are simply two sides of the same coin.

Queueing theory and traffic theory are two sides of the same coin.

When developing queueing theory models or traffic models, we should think of the *objects* as an abstraction of people waiting in a line. In addition to this line we have a *server*, which is an abstraction of a teller that serves the line. In the case of healthcare, the server might represent a patient leaving the hospital, thus opening a bed for another patient to move into.

In order to model how objects move through a system, we must determine the arrival pattern, the service pattern, the number of service channels, the system capacity, and the queue discipline. More complicated queueing models may also incorporate multiple services stages and impatience.

The *arrival process* is the distribution of arrival times of objects entering the queue, and the *service process* is the distribution of service times of those objects in the queue. Most commonly, these rates are stochastic and the Poisson or exponential distribution is used (see Chapter 5 for details on probability distributions). These are the natural rates for queueing theory as they represent the probability of a series of consecutive independent events.

The *number of service channels* is the number of possible ways an object can exit the queue. If there is more than one server, the queue will probably interact in various manners. For example, arrival patterns may dictate that a new arrival will automatically enter the shortest queue.

The *system capacity* refers to the maximum number of objects allowed in the system. In some models there may be multiple *service stages*, that is, when an object exits one queue it is automatically placed into a following queue. In these cases, each service stage will have its own service capacity. If we are studying concepts of expected waitlist length, it is important to define this accurately. Conversely, if we are developing models to analyze how the capacity of the system changes, then we may simply define the maximum system capacity as infinite, so maximum expected capacities can be determined.

The *queue discipline* refers to the manner in which customers are selected for service, and is often considered the most important decision in developing a queueing theory or traffic theory model. There are four common queue disciplines: First In First Out (FIFO), Last In First Out (LIFO), Service In Random Order (SIRO), and Priority schema. *First in first out* commonly arises in line-ups where patients will feel they are being treated unfairly if it is not used. *Last in first out* arises in dealing with stacked objects where the top object must be removed before lower objects are accessible. This most commonly occurs when dealing with warehousing issues, including blood banks and organ donor clinics. *Service in random order* is applied when the actual queue discipline is not under the control of the modeller. For example, if one is modelling hospital bed counts, then the server represents

when a patient leaves the hospital, making a bed available for use. Finally, *priority schema* are used in emergency room and surgery queueing.

In more advanced queueing theory models, we often wish to exhibit impatience in the objects. *Impatience* represents objects that leave the system prematurely or refuse to enter the system due to issues with the queue length.

After developing a queueing theory or traffic theory model, we usually desires information on average queue lengths, maximum queue lengths, average systems capacity, maximum system capacity, etc. To do this we seek understanding of the equilibrium state of the queue. *Equilibrium* is the idea that, if run long enough, the model may approach a point where the length of the queue and capacity of the system are independent of the time variable. This is discussed further in Section 3.2. It is worth emphasizing here that *not all queues will approach an equilibrium state.*

2. Common Uses

Queueing theory is applicable in any situation where we are trying to model objects moving through a system. Its two major uses are in studying wait times for the objects to travel through the system, and to examine fluctuations in the capacity of the system. Queueing theory regarding wait time is of great help in studying questions regarding waitlists for surgeries, or admittance rates into various hospital departments. For example:

- *How does the number of surgeons impact how long a patient waits before receiving surgery?*
- *How does hospital emergency room admittance change throughout the day?*
- *How does the hospital determine how many elective procedures to do, while still maintaining the necessary access for emergencies?*

With regard to capacity of the system, queueing theory in healthcare is often used to study the amount of bed space available in a hospital and to locate potential bottle necks in the system. Queueing theory may address questions such as:

- *How does hospital bed usage vary throughout the day, week, month, or year?*
- *How do different hospital department occupancy rates interact?*
- *Is back-log in one hospital department causing extra stress on another department?*

3. Mathematical Details

Both queueing and traffic models address problems where objects or people move through a system. Both models begin with a system and a collection of objects that seek to enter and exit the system. The theory behind these two modelling techniques is essentially the same, and the major difference lies in what outcome we want to measure. In general, traffic models are focused on congestion in dynamical systems, whereas queueing models focus on understanding wait times within the system. Since the interactions between the objects and the system are what determine these properties, it is clear that the study of queueing theory and traffic theory are simply two sides of the same coin. Hence, although the remainder of our discussion shall refer to queueing models, it should be recognized that the theory below is equally applicable to traffic models.

3.1. Building a Queueing model. In order to study how the system and objects interact, we next develop the idea of a server. If the objects are an abstraction of customers waiting in a line, then a *server* is a concept that abstracts the idea of the teller that serves that line. In the case of waitlist modelling, the server might represent an open operating room slot that can be used to serve one patient from the waitlist. In the case of bed count modelling the server might represent a random event that sends a patient home (thus making a hospital bed available for use).

To build a queueing model one must define:
A: the arrival pattern*,*
B: the service pattern*,*
X: the number of service channels*,*
Y: the system capacity*, and*
Z: the queue discipline*.*
Once these items are selected, the constructed queues are often called an A / B / X / Y / Z *queue.*

Now that we have the idea of objects, systems, and servers firmly established, we can state that the basic ingredients of queueing (and traffic) models are the *arrival pattern*, the *service pattern*, the *number of service channels*, the *system capacity*, and the *queue discipline*. More complicated queueing models may also incorporate *multiple services stages* and *impatience*. We will discuss each of these in turn.

Arrival and service processes. The arrival process is the distribution of arrival times of objects entering the queue. The arrival process may be either deterministic or stochastic. If it is stochastic, a probability distribution must also be selected to describe the arrival rate (most commonly this is a Poisson distribution, but occasionally others are used).

The service process is the distribution of service times for the objects in the queue. As with the arrival process, this may be either deterministic or stochastic, and a variety of probability distributions can be considered.

Number of service channels and multiple service stages. The number of service channels is the number of possible ways an object can enter into service. In a simple case, this might be viewed as a number of parallel queues, each of which have independent servers to allow for exiting the queue (see Figure 1).

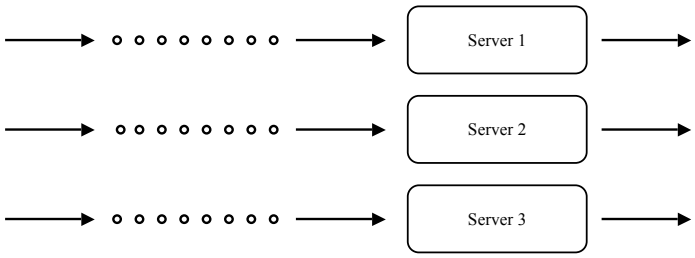

FIGURE 1. **A multiple service channel queue:** A schematic representation of a queueing theory model with multiple service channels and parallel queues.

If these independent queues are not interacting then each can be studyied individually. More realistically, if there is more than one server, there may be complex interactions within the system. For example, arrival patterns may dictate that a new arrival will automatically enter the shortest queue, or if impatience is used (see below) then objects in the queue may move to shorter queues as they become available. This latter concept is called jockeying.

In some models it may also be appropriate to have multiple service stages. That is, when an object exits one queue they are automatically placed into a following queue. This is captured in the Figure 2

FIGURE 2. **A multiple service stage queue:** A schematic representation of a queueing theory model with multiple service stages.

In multistage queues one may wish to incorporate the idea of *blocking*. The idea in blocking is that certain stages of the queues have maximum occupancy levels. Thus even if an object in an earlier queue has been served, it may be blocked from entering a later queue.

The most complicated queueing models are multistage queues that take the form of a complex network with feedback loops. In these queues, an object exiting one level of the queue may result in a number of different possible outcomes. It may exit the system, enter a queue, or enter a number of different queues. The process of deciding which queue to enter next may be based on occupancy, time, or may be strictly random. (An example of such a queueing model is given in Example 4.2.)

System capacity. The system capacity refers to the maximum number of objects allowed in the system. In most applications in healthcare, the system capacity is finite. However, if we are developing models to analyze how the capacity of the system changes then we may simply define the maximum system capacity as infinite, so maximum expected capacities can be determined.

Queue discipline. Perhaps the most important aspect to developing a queueing model is the idea of queue discipline. Queue discipline refers to the manner in which customers are selected for service. The four most common disciplines are:

(1) First In First Out (FIFO),
(2) Last In First Out (LIFO),
(3) Service In Random Order (SIRO), and
(4) Priority schema.

At first glance the first in first out rule may seem like the only fair and logical rule, prompting one to ask why the other disciplines would ever be considered. In the case of healthcare, it should be immediately clear that sometimes priority schema will take precedence over the classical FIFO rule. In fact, in most queueing theory textbooks, emergency room and surgery queueing are used as the classic examples of when priority schema should be developed.

Service in random order queues often arise when the actual queue discipline is not under the control of the modeller. For example, if one is modelling hospital bed counts, then the server represents when a patient leaves the hospital, making a bed available for use. In general this is not under the control of the doctor, but has a large random aspect involved. To clarify, although a doctor may know several hours (or even days) before a patient leaves what the departure time will be, the doctor cannot know the departure time of a random patient as they enter the hospital (i.e. before diagnosis occurs). Moreover, the doctor can certainly not say to a healthy

patient, "I'm sorry you can't leave yet. You see, John over there isn't healthy yet, and he got here before you."

The applications of the last in first out rule to healthcare are more obscure. The LIFO rule, often called the stack rule, most commonly arises in computer programming and warehouse management, where it is easier to take off the top of the pile than the bottom. Surprisingly, LIFO is often used in blood banks, despite the fact that storing blood for too long can cause spoilage. (This is likely because it is quite rare for blood banks to remain stocked for long.)

Impatience. In more advanced queueing models, it may be important for the objects to exhibit impatience. This is of particular interest when the objects being considered are people (as is often the case in healthcare). Some examples of impatience include:

- Balking: The customer may decide not to enter the queue upon arrival, perhaps because it is too long.
- Jockeying: If there are multiple queues in parallel the customers may switch between them.
- Reneging: The customer may decide to leave the queue after waiting a certain time in it.
- Drop-offs: Customers may be dropped from the queue for reasons outside of their control.

Mathematically, the concepts of reneging and drop-offs can be treated as one, but it is often easier to understand the model if these are treated separately.

3.2. Analyzing a Queueing Model. Having built a queueing model, we now wish to extract from it information such as average queue lengths, maximum queue lengths, average systems capacity, maximum system capacity, etc. If the model is based on deterministic arrival and service rates, then this is typically a fairly simple procedure that can be done very effectively via computer simulation. If the arrival or service rates of the model are stochastic, as is the case with most applications in healthcare, the process becomes more complicated.

In either case, one of the most important concepts in queueing theory is that of *equilibrium*. Equilibrium is the idea that, if run long enough, the model may approach a point where the length of the queue and capacity of the system are independent of the time variable. It is worth making it very clear that not all queues will approach an equilibrium state (see Figure 3).

Note that not all queues will approach an equilibrium state.

In the case of deterministic models, reaching an equilibrium state often means that the length of the queue and the capacity of the system become constant from one time period to the next. However, it may also mean that the length of the queue and capacity of the system alternate between two states, or cycle through a known pattern of states.

In the case of stochastic queueing models, the concept of equilibrium becomes much more complicated. Instead of approaching a constant or cyclic state, the length of the queue and capacity of the system may (if one is lucky) approach a time-independent probability distribution. This means that at any given time period, the length of the queue and capacity of the system will be unknown, but follow a known probability distribution. (For information on probability distributions see Chapter 5.) Since the term equilibrium is no longer sufficient, this state is usually referred to as a *statistical equilibrium* (see Figure 3).

To say a queue is approaching statistical equilibrium does not mean the queue length and system capacity are approaching a constant value, rather it means that in the long-time limit, the queue lengths

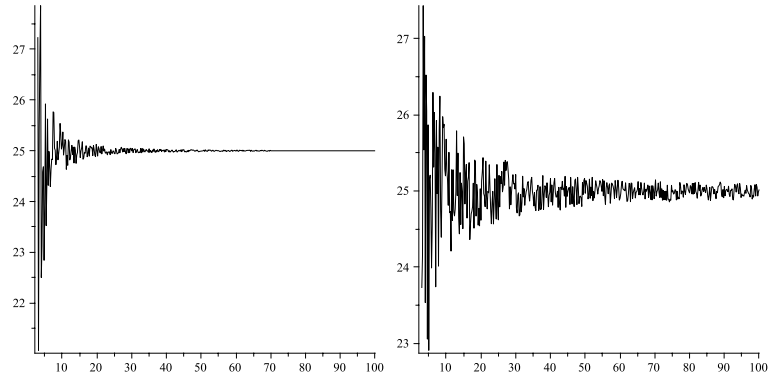

FIGURE 3. **Queueing models reaching various equilibrium states:** The above graphs might represent model output for queue length over time. The first graph (left) quickly reaches an equilibrium state of 25. The second graph (right) does not reach a fixed equilibrium state, but does approach a statistical equilibrium with a mean of about 25 and a standard deviation of about 0.1.

The mathematics required to analyze if a queue has an equilibrium state can be very simple, or very complicated, depending on the model developed. For simple queues it is often possible to develop a closed form analytic solution that describes the equilibrium state of the queue (two examples of this are given in Example 4.1). In more complicated cases, the queue may be solved via simulation methods (see Chapter 9). The trouble with "solving" queueing models via simulation methods is that it is often difficult to determine exactly when a state of equilibrium has been achieved, especially if the queue is complicated and stochastic in nature. This is often worked around by running the simulation for a "warm-up" period before trusting the results of the simulation.

4. Examples

4.1. Washing dishes in the hospital cafeteria. In this example we develop several artificial queueing models that simulate dish washing in a small town hospital cafeteria[1].

The objects in our queue will be dirty dishes (plates), and the system will be the storage racks used to collect the dishes. Dishes arrive into the queue after a hospital employee uses them to eat a meal. Dishes exit the queue after the hospital's dish washer cleans them and places them into the clean dish racks. We will assume that the cafeteria workers follow a LIFO queueing discipline. This means that dirty dishes are added to the top of a "pile" of dirty dishes, and the cafeteria workers clean dishes on the top of the pile first. In our model we will assume that there is only one server channel. This can be viewed as either lumping all the dish washers into one unit or as a single employee whose job is to clean dishes.

If we replace the dirty dishes in this example with occupied hospital beds, and the clean dishes with available hospital beds, then we can easily use this framework to develop a queueing theory model for hospital bed count.

Our first and most basic model will be a deterministic queue. After some data collection we determine that the hospital cafeteria sells 12,000 meals each day, and

[1]By artificial, we do not mean that the model is unrealistic, only that the model is calibrated with artificial data.

therefore produces 12,000 dirty dishes per day. The dish washers are capable to cleaning 500 dishes per hour. We therefore set our arrival rate to $A = 12,000/day$ and our service rate as $B = 500/hour$. Our queue may now be modelled as follows. Let t represent time in hours, and $N(t)$ represent the number of dirty dishes in the dirty dish racks at time t. Then

$$\begin{aligned} N(t) &= \max\{0, A * \lfloor \tfrac{t}{24} \rfloor - B * \lfloor t \rfloor + N_0\} \\ &= \max\{0, 12000 * \lfloor \tfrac{t}{24} \rfloor - 500 * \lfloor t \rfloor + N_0\}, \end{aligned}$$

where N_0 represents the number of dirty dishes at time $t = 0$, and the brackets $\lfloor \cdot \rfloor$ represent the flooring function (i.e. round down to the nearest integer). Notice we take the maximum of the computed number and zero, this forces the dirty dish count to always be nonnegative.

Examining this model it is easy to compute the equilibrium of the system. Since each day we are adding 12,000 dishes to the queue, and removing (up to) $500 * 24 = 12,000$ dishes from the queue, it is clear that the queue will balance itself out every day. Thus at the beginning of each day, the queue will jump to $N_0 + 12,000 - 500$ every twenty four hours, and then decease by 500 each hour until it reaching N_0. Thus the equilibrium for this queue is a pattern that cycles every 24 hours.

Of course this model is naive in several manners. It is unlikely that all the dirty dishes arrive at exactly midnight every night. Second, it is unlikely that the dish washers process exactly 500 dishes, every hour, on the hour. Basically, we have completely ignored the stochastic nature of the problem. To correct this, consider the following more advanced queueing model.

We assume that the arrival rate and service rate are stochastic. In particular we will set the arrival pattern as a Poisson distribution with a mean of $12,000$ $dishes/day$ and the service pattern as a Poisson distribution with a mean of 500 $dishes/hour$. (Recall, the Poisson distribution is the natural choice for modelling arrival rates. See Chapter 5, Subsection 3.3, for more information on the Poisson distribution.) Our new queue can be visualized as shown in Figure 4.

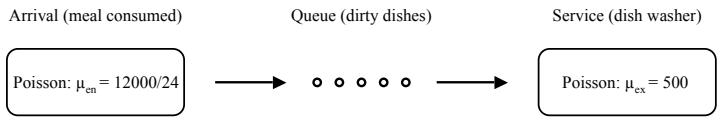

FIGURE 4. **Dish washing queue:** A schematic model of the dish washing queue developed in example 4.1. (The abbreviations *en* and *ex* refer to *enter* and *exit* respectively.)

We now seek the statistical equilibrium for this queue. We begin by defining $\Pr_n(t)$ as the probability that the queue contains n elements (dirty dishes) at time t. The *change* in probability from t to $t + 1$ can now be computed as follows:

$$\Pr_n(t+1) - \Pr_n(t) = \mu_{en} \Pr_{n-1}(t) + \mu_{ex} \Pr_{n+1}(t) - \mu_{en} \Pr_n(t) - \mu_{ex} \Pr_n(t). \tag{29}$$

Equation (29) can be interpreted as follows:

$$\begin{array}{rcl} \mu_{en}\Pr_{n-1}(t) & : & \text{plus the chance a queue of length } n-1 \text{ increases to length } n, \\ \mu_{ex}\Pr_{n+1}(t) & : & \text{plus the chance a queue of length } n+1 \text{ decreases to length } n, \\ -\mu_{en}\Pr_{n}(t) & : & \text{minus the chance a queue of length } n \text{ increases to length } n+1, \\ -\mu_{ex}\Pr_{n}(t) & : & \text{minus the chance a queue of length } n \text{ decreases to length } n-1, \end{array}$$

sums to

$$\Pr_n(t+1) - \Pr_n(t) : \text{ the } \textit{change} \text{ in probability from time } t \text{ to time } t+1.$$

The first two terms of the right hand side of equation ((29)) are added, since they increase the probability of a queue of length n, while the final two terms are subtracted since they decrease the probability of a queue of length n. For the special case of $n = 0$ equation (29) is replaced by

$$\Pr_0(t+1) - \Pr_0(t) = \mu_{ex}\Pr_1(t) - \mu_{en}\Pr_0(t). \tag{30}$$

Next we assume that for some large value of t the queue has reached statistical equilibrium, implying $\Pr_n(t) = \Pr_n(t+1) = \Pr_n$. Thus, equations (29) and (30) reduce to

$$\begin{array}{rcl} \Pr_1 & = & \frac{\mu_{en}}{\mu_{ex}}\Pr_0 \\ \Pr_{n+1} & = & \frac{\mu_{en}+\mu_{ex}}{\mu_{ex}}\Pr_n - \frac{\mu_{en}}{\mu_{ex}}\Pr_{n-1} \end{array}$$

This system is solved by the iterative formula

$$\Pr_n = \left(\frac{\mu_{en}}{\mu_{ex}}\right)^n \Pr_0 . \tag{31}$$

To develop equation (31) correctly, we should actually begin by developing a series of differential equations, and then setting the derivatives to 0. Equations (29) and (30) provide a discretized interpretation of this mathematical technique.

Since the queue must have some length $(0, 1, \ldots)$, we note that the sum of all the probabilities $\Pr_n$ must be 1:

$$1 = \sum_{n=0}^{\infty}\Pr_n = \sum_{n=0}^{\infty}\left(\frac{\mu_{en}}{\mu_{ex}}\right)^n \Pr_0 = \Pr_0 \sum_{n=0}^{\infty}\left(\frac{\mu_{en}}{\mu_{ex}}\right)^n .$$

This sum converges if and only if $\frac{\mu_{en}}{\mu_{ex}} < 1$, therefore a statistical equilibrium can be achieved if and only if $\frac{\mu_{en}}{\mu_{ex}} < 1$. If $\frac{\mu_{en}}{\mu_{ex}}$ is less than 1, then the sum converges to

$$\sum_{n=0}^{\infty}\left(\frac{\mu_{en}}{\mu_{ex}}\right)^n = \frac{1}{1-\frac{\mu_{en}}{\mu_{ex}}},$$

which tells us $\Pr_0 = 1 - \frac{\mu_{en}}{\mu_{ex}}$, and provides (with equation (31)) the statistical equilibrium for the model.

From here the modeller may be satisfied, or may wish to develop this model further. Some examples might include adding a maximum length to the queue (representing the total number of dishes the cafeteria owns), using multiple time-dependent arrival and departure rates that represent different times of the day (week, month, or year), creating multiple server queues where each dish washer works at a different rate, or creating multiple staged queues that represent moving the dirty dishes from the table to the dish rack and then cleaning them. Once the modeller is satisfied with the quality of the model, the model can be used to explore certain "interventions." For example, the impact of a cafeteria dish washer strike could be examined by running the queue to equilibrium then dropping the departure rate to 0, while the effect of hiring more dish washers could be examined by increasing the departure rate.

4.2. Interrelating Hospital Capacity across Departments. Both from a humanitarian and a business perspective, it is of interest to reduce inefficiencies in the healthcare system. One approach is to examine the problem of bed allocation in hospitals. In this example we summarize work by Cochran and Bharti on designing and implementing a queueing theory model for hospital bed allocation [**45**].

Cochran and Bharti's goal was to produce an accurate queueing theory network model to describe patient flow through a hospital. Their modelling process began by interviewing staff, and charting the patient flow in a 400 bed hospital in the United States of America. Using this information they developed a complicated network that described possible patient flows in the hospital. A simplified version of this network can be found in Figure 5. The model was tuned using statistics

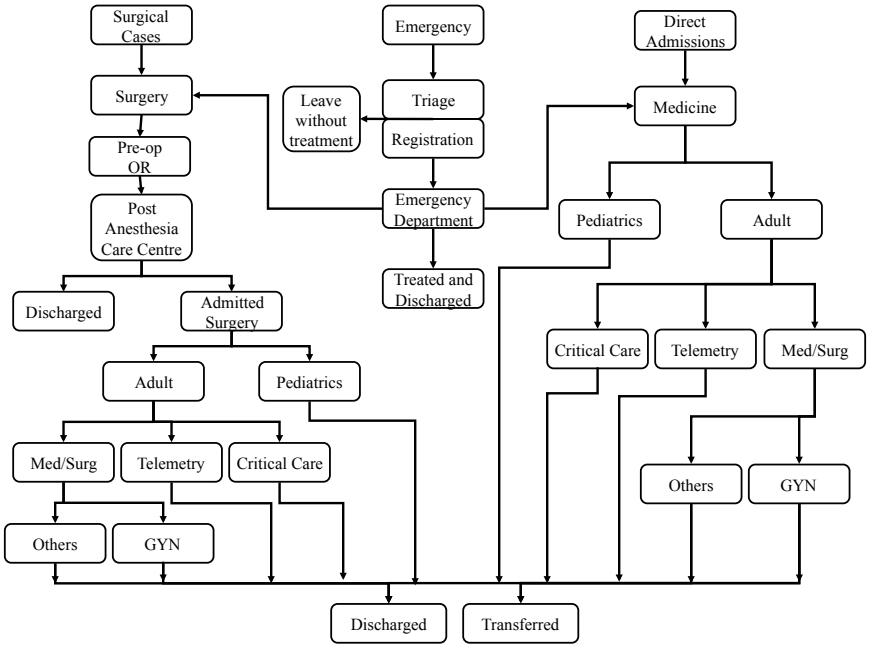

FIGURE 5. **Hospital patient flow queue:** A simplified version of Cochran and Bharti's queueing theory model describing patient flow through a hospital.
Based on the work of [**45**].

obtained from economic data stored by the hospital. In particular, they required admittance rates, probability of being an elective admittance versus an emergency room admittance, average length of stay data for each unit in their model, and (when applicable) the relative probabilities of where a patient will move after being serviced at a given hospital unit.

The next step in Cochran and Bharti's analysis was to simplify the model to the point where the model could be solved analytically. In particular this meant making the following assumptions:

- no difference was made between exiting the hospital due to recovery and exiting the hospital due to death,

- each empty bed is available to all patients arriving in the unit (i.e. all beds are always appropriately staffed and equipped for any patient),
- there is only one type of patient and no priority structure,
- the length of stay for each unit is exponentially distributed,
- the length of stay for each unit is independent of the state of the system, and
- the relative probabilities of where a patient will move after being serviced at a given unit are independent of the state of the system.

Although on some level each of these assumptions is wrong, together they mean that the resulting model was what is referred to as a *Jackson Network Queue*. The importance of this is that, regardless of how complicated the model is, any Jackson Network Queue can be analytically solved to find an exact closed form stochastic equilibrium. To check that the above assumptions were not too restricting, the closed form solution for the queue was compared to actual historical hospital occupancy rates, and a very close match was found.

Whenever an analytical closed form solution can be created, it should be used to check that the simulation model is working properly.

The next step in Cochran and Bharti's analysis was to develop a discrete event simulation program for the model. This was first done using the restricting assumptions above, and the simulation was compared to the closed form solution. (This is an excellent troubleshooting step, and should be performed whenever possible.) Next, the restricting assumptions were relaxed one at a time until only the final two remained (the relative probabilities and length of stay rates are independent of the state of the system). In addition, the discrete event simulation model included blocking of patients caused by the finite bed capacities of each unit, two classes of patients (emergency and regular patients) with emergency patients given priority for beds, and probability distributions that varied according to the time of day and the day of the week. It was again checked, and confirmed, that the final simulation produced results very similar to historical hospital data.

Having developed the simulation, Cochran and Bharti proceeded to develop optimization strategies for improving hospital efficiency. First they noted that there were large discrepancies in the bed loads between different departments. Using both the analytic and simulation models the optimal bed allocation to balance loads was calculated, and suggested reallocations were provided to the hospital. Using the simulation model, moments of unbalance in the temporal loads were determined, and strategies to rebalance the loads were developed. For example, they showed how blocking could be decreased if elective procedures were conducted during off-peak times.

4.3. Hip and Knee Replacement Surgical Waitlist in British Columbia. Over the past 40 years, knee and hip replacement surgery has advanced to the point where it is the standard approach (in Canada) for treating chronic joint pain. The improvement in techniques, along with the aging population of Canada, has led to an increased demand for these procedures and, consequently, increases waitlist length. To understand the attributes that impact waitlist length, and to explore potential interventions, a research team at the IRMACS center developed a working queueing theory model for hip and knee replacement surgical waitlists [**184**],[**218**]. In this example we outline the model, and provide some of the analytical evaluation of the model.

The basic model consists of individuals entering the queue on a continuous basis. Individuals can exit the queue either through surgery or by dropping out.

The surgery server is assumed to be the classic FIFO server (first in first out), while the drop-out server was SIRO (service in random order). This means that to have surgery, an individual must wait until they are at the front of the queue, but an individual may drop off the queue at any time. The arrival rate and surgery rate are both assumed to be continuous with rates r and s respectively (in this manner the queue can be modelled via differential equations). The drop-off rate is also continuous, but with a rate that varies in proportion to the current length of the queue: kN (k is the drop-out proportion, and N is the current length of the queue).

In order to work with the queue, a discrete event simulation was developed and analytical methods were employed. Using the discrete event simulation, the researchers were able to "tune" the model until the parameters r, s, and k produced results similar to those found in the available data. The available data consisted of two data sources: the Discharge Abstract Database (DAD) and the Surgical Wait List (SWL) registry. The DAD is a validated data set that includes hospital, surgeon, and procedure, but no information on wait times. Conversely, the SWL registry includes entry and exit dates for surgical waitlists, but is unvalidated data. Using common patient identifiers (surgeon plus operation date for example), these two lists were combined and the linked cases were used for data flow analysis in preparation for the simulation.

The queue length N can be mathematically approximated[2] by the differential equation

$$\frac{dN}{dt} = r - s - kN.$$

This equation states that the change in the size of the queue is proportional to the people joining the queue, minus the number of people exiting via surgery and the number of drop-outs. Clearly, if the rate of joining (r) is greater than the rate of surgery (s), then the queue will grow. However, the rate of drop-off (kN) grows with N, and will eventually approach $r - s$. Thus, eventually the rate of growth of the queue becomes negligible. (This was supported in the output of simulations.)

This system can be solved analytically, obtaining formulas for the queue size, waits and total dropouts, and giving a valuable comparison. In particular, the differential equation for the queue size has the solution

$$N(t) = \frac{r-s}{k} - \left(\frac{r-s}{k} - N_0\right) e^{-kt}, \tag{32}$$

where N_0 is the number of individuals in the queue at time 0.

As we are further interested in the wait time for a given individual, for a *fixed* individual P we define the wait time as $W = t_{out} - t_{in}$, where t_{in} is the time the patient enters the waitlist and t_{out} is the time the patient enters surgery. We also define the size of the waitlist in front of P at time t as $Q(t)$. The function $Q(t)$ satisfies the differential equation

$$\frac{dQ}{dt} = -s - kQ,$$

when t is restricted to the time interval $t_{in} \leq t \leq t_{out}$. (The change in the size of the queue in front of P is proportional to the number of people exiting via surgery

[2]This is called the fluid approximation for large queues

and the number of drop-outs.) This equation is solved by

$$Q(t) = -\frac{s}{k} + \left(\frac{s}{k} + Q_{t_{out}}\right) e^{-k(t-t_{out})},$$

where $Q_{t_{out}}$ is the number of people in front of P at time t_{out}. As $Q_{t_{out}}$ must be 0 this reduces to

$$Q(t) = -\frac{s}{k} + \frac{s}{k} e^{-k(t-t_{out})}.$$

Finally, we link $N(t)$ and $Q(t)$ by noting that $N(t_{in}) = Q(t_{in})$. Thus

$$\begin{aligned} N(t_{in}) &= -\tfrac{s}{k} + \tfrac{s}{k} e^{-k(t_{in}-t_{out})} \\ N(t_{in}) &= -\tfrac{s}{k} + \tfrac{s}{k} e^{kW} \\ e^{kW} &= \tfrac{k}{s}\left(N(t_{in}) + \tfrac{s}{k}\right) \\ kW &= \ln\left(\tfrac{k}{s} N(t_{in}) + 1\right) \\ W &= \tfrac{\ln\left(\frac{k}{s} N(t_{in})+1\right)}{k}. \end{aligned}$$

Thus we have a closed form solution to the expected total wait time for P, provided we know the length of the waitlist at the time P enters the queue.

5. Related Reading

Queueing theory models have close connections to Markov Models (Chapter 13) and system dynamics models (Chapter 14). Queueing theory models also rely heavily on probability distributions (Chapter 5).

Examples of queueing theory and traffic theory applied to healthcare are abundant. A few examples include references [**17**], [**48**], [**28**], [**89**], [**133**], [**71**], [**91**], [**110**], [**159**], [**97**] and [**92**]. Reference [**17**] investigates queue networks with various classes of customers. Reference [**48**] derives the queue size distribution for infinite server queues with Poisson arrivals and exponential service times. Reference [**28**] uses a discrete event simulation model of an emergency room to determine issues contributing to wait times in the delivery of primary care in emergency rooms. Reference [**89**] constructs a simulation model to improve understanding of the behaviour of waitlists over time and the effect of policy on wait times. Reference [**133**] formulates a system dynamics queueing theory model to explore the factors that contribute to long wait times for urgent admissions to acute care in hospitals. Reference [**71**] describes a simulation model designed to provide decision support for patient scheduling for elective surgery in the public hospital system. Reference [**91**] uses a queueing model to investigate the effect of changing several conditions on bed allocation and bed occupancy rates in hospitals. Reference [**110**] discusses the circumstances in which queueing theory predicts it is optimal to have waiting time for public health treatment. Reference [**159**] explores the application of queueing theory and principles of industrial engineering, adapted to clinical settings, in substantially reducing delay in current health care systems. Reference [**97**] provides an overview of patient flow and the application of queueing theory to model this flow. Reference [**92**] describes a novel method of incorporating impatience into queueing theory for healthcare.

Literature of waitlists in healthcare includes references [**141**], [**54**], [**158**], [**98**], [**206**], [**99**], [**139**], [**192**], [**39**], [**215**], [**166**], [**138**] and [**191**]. Reference [**141**] examines a case in economics where queues of suppliers and customers emerge in a nonstochastic setting when price is below or above the market-clearing level. Reference [**54**] discusses the main arguments advanced to explain the waiting lists in Britain's National Health Service and presents a framework that focuses on "demand" and "supplier" induced demand considerations. Reference [**158**] presents findings from a simple simulation model about waiting lists for hospital inpatient treatment. Reference [**98**] reviews the implications and results of implementing maximum wait time guarantees in Sweden. Reference [**206**] discusses how supply has been addressed in Victoria by changing the financial incentives

in hospitals and the consequent reductions in waiting lists. Reference [**24**] compares the differences in prices and wait times of 7 health services between US and Canadian hospitals. Reference [**99**] focuses on physicians as implementers of health policy reforms. Reference [**139**] investigates misconceptions about waiting lists and outlines a number of initiatives that could contribute to more durable solutions to the waiting list problem in Canada. Reference [**192**] looks at the decline in Canadians' approval of their healthcare system, the reasons behind these declines and why these perceptions may hamper progress in the healthcare area. Reference [**39**] describes the relevance of the functional relationship between point-count measures assessing severity of patients' conditions and the extent of benefit expected from waitlist services for the development and validation of priority criteria. Reference [**215**] investigates the level of capacity required to operate a booked admissions policy for elective inpatient services where patients are given a date for hospital admission months in advance rather than being out on a waiting list and being informed of their admission date at short notice. Reference [**166**] provides an overview of the Western Canada Waitlist Project. Reference [**138**] explores some of the logical and ethical implications of the work done for the Western Canada Waitlist project. Reference [**191**] explores the issues and methods related to developing acceptable waiting times for selected medical services in Canada.

Reference [**4**] contains the details omitted from Example 4.3.

Reference [**45**] uses a queueing theory network model to minimize bed blocking in hospitals, and is discussed in Example 4.2.

CHAPTER 16

Finding the "Best" Intervention

The man who is a pessimist before 48 knows too much; if he is an optimist after it, he knows too little. Mark Twain (1835-1910)

I am an optimist. It does not seem too much use being anything else. Winston Churchill (1874-1965)

Optimization

1. Model Overview

At the beginning of this book (Chapter 1, page 4) we defined a model to be a *simplified representation* of a real world situation used to help answer a *specific question*. This chapter is not about creating models. That is, the tools discussed in this chapter do not create simplified representations of real world situations. Instead the tools discussed in this chapter are designed to be used on models, sometimes in order to determine a theoretically optimal intervention, other times simply to tune the model to best fit real world data. In either case, what we seek to do is determine what set of input parameters for the model create the "best" model output.

In mathematics (and in this book), the term optimization *refers to the study of how to find the minimum or maximum of a function over an allowed set.*

In order to answer this question it is clear that we must begin by defining what is meant by the word "best". Mathematically, this means we need to create a *quantitative objective function*. This is a function that takes the model output and turns it into a single number. We then seek to determine what selection of input parameters for the model cause the quantitative object function to produce the best number. This might represent the smallest error, or the greatest benefit.

For example, suppose we have created a model that examines how the number of surgeries impacts the wait time for knee and hip surgery (as in Example 4.3), and we want to "optimize" the number of surgeries performed each month in order to keep patient wait times at an "acceptable level". At this point, we must carefully address the question of what is meant by the statements "optimal number of surgeries" and "acceptable level." The second statement, "acceptable level", could be addressed by stating that x percent of patients should receive surgery within y months. The "optimal number of surgeries" might then be defined as the number of surgeries that result in the minimal total cost to achieve this level. With strong definitions in place, we can turn to the tools optimization to answer the question.

As in the above example, in many circumstances the objective function takes the form of minimizing the total cost of the intervention. However, in some cases more complex objective functions may be required. For example, we might seek to maximize the impact of a medicine, or minimize the number of individuals that will be infected by a given disease.

Another very common use of optimization in modelling is in the idea of parameter tuning. A crucial step in the implementation of any mathematical model is to make the model's output align with observed real world data. The most common manner of doing this is to tune a set of model parameters in a manner that minimizes the "error" between the model output and observed real world data. In simple models the tuning of parameters can often be accomplished by well established methods (such as least-squares or maximum likelihood, see Chapter 6), while more complicated models may rely more heavily on advanced optimization methods.

In either case, the basic method is the same. We begin by defining a qualitative objective function that measures the error between the model output and observed data. For example, the classic mean square error,

$$error = \sum (model\ output - observed\ data)^2$$

is used in the least squares method for linear regression (Chapter 6).

After creating an objective function, the modeller must next optimize the function. We call this step solving the *optimization problem*. This can be done in many different manners. If we are lucky, the problem is easy enough to solve exactly, possibly by hand. However, it is much more likely that the problem is either too large or too complicated for this to be the case. In such circumstances, we can often turn to computers and optimization algorithms for assistance. Optimization algorithms are collections of formal steps that are guaranteed to provide better and better estimates to the solution of an optimization problem. Since, in most real world circumstances it is sufficient to find the solution within one or two percent, optimization algorithms provide a large toolbox of methods for solving optimization problems.

Fortunately, optimization and the development of optimization algorithms is a fairly well established field of mathematics, so many optimization algorithms have already be designed and implemented as computer codes. Unfortunately, the diversity of the field, and the diversity in the types of problems that require optimization, means that most codes are designed with a specific problem style in mind. In the Mathematical Details section of this chapter, we try to provide some insight on how to examine an optimization problem with the goal of selecting a good optimization algorithm to solve the problem.

2. Common Uses

Strictly speaking, optimization is not a modelling technique but a method of examining models to determine what intervention will have the "best" effect. Most commonly, optimization problems consider techniques to minimize cost under some given constraint. For example

- *What combinations of drugs minimize the cost of pharmaceuticals while producing the desired effect?*
- *Where should we build a new hospital to minimize cost given that everybody should be able to reach it in less than 30 minutes?*
- *How should we schedule nurses to minimize cost given that certain staffing and hospital service constraints must be met?*

are all examples of optimization problems. Each of these problems can be posed in the *dual form*, which examines how to maximize the impact given a fixed budget:

- *What combination of drugs provides the maximal impact given we can only afford a fixed budget?*
- *Given a fixed budget, where is the best place to build a new hospital in order to maximize public accessibility?*
- *How to we maximize our patient service in terms of nursing coverage given a fixed budget?*

Many other problems can be posed as optimization problems. In most cases, the first step to solving such a problem is to develop a model of the problem upon which optimization techniques can be applied. Because of this, it is important to know what types of problems are easily solved by today's optimization tools, and what types of problems are intractable using today's optimization tools.

3. Mathematical Details

Optimization is a method of examining models to determine what intervention will have the "best" effect. In order to answer this question it is clear that we must begin by defining what is meant by the word "best". To answer this, we must define the term best in regards to a *quantitative objective function*. A quantitative objective function is a function with the property that for every input the function must return a real number or the value $+\infty$, and must take on a non-infinite value in at least one location. Mathematically, functions with these properties are called *proper*[1].

Let $f(x)$ be a proper function from $\mathbb{R}^n$ to $\mathbb{R}$ and S be a non-empty set in $\mathbb{R}^n$. Optimization is the field of study interested in solving the problem

$$\min\{f(x) : x \in S\},$$

or (in English) minimize the *objective function* $f(x)$ such that x lies in the *constraint set* S. By minimize we mean find a point $\bar{x}$ such that $f(\bar{x}) \leq f(x)$ for all other x in S. Before we discuss the role of the constraint set, let us note that should we be interested in maximizing a function,

$$\max\{f(x) : x \in S\},$$

then we can always create a minimization problem by applying the following theorem:

> **Theorem:** Let $f(x)$ be a proper function from $\mathbb{R}^n$ to $\mathbb{R}$ and S be a non-empty set in $\mathbb{R}^n$. Then, any point which maximizes $f(x)$ over S also minimizes $-f(x)$ over S, and vice versa. Consequently,
>
> $$\max\{f(x) : x \in S\} = -\min\{-f(x) : x \in S\}.$$

[1]Here, and henceforth, we consider optimization in terms of minimizing the quantitative objective function. If we are interested in maximizing the function, then most of these definitions are "turned upside-down". For example, if we are studying maximization then a proper function is one such that for every input the function must return a real number or the value $-\infty$, and must take on a non-infinite value in at least one location.

The addition of the constraint set can make similar optimization problems behave very differently. To see this consider the following three problems:

$$\min\{x^2 - x : x \in \mathbb{R}\}, \tag{33}$$

$$\min\{x^2 - x : x = -2, -1, 0, 1, \text{or } 2\}, \text{and} \tag{34}$$

$$\min\{x^2 - x : x = \frac{m}{3^n} \text{ for some } m, n = 1, 2, ...\}. \tag{35}$$

(In problems (33), (34), and (35), we have the constraint sets $S = \mathbb{R}$, $S = \{-2, -1, 0, 1, 2\}$, and $S = \{x : x = \frac{m}{3^n} \text{ for some } m, n = 1, 2, \ldots\}$ respectively.) In each of the above problems the objective function is the same, however the problems are very different. It is not difficult to show that the problem (33) is solved at $x = 0.5$ and gives an objective value of -0.25. Problem (34) is easier, but comes with a twist. Since there are only 5 options for x it is a simple matter to check each and determine the minimum value is 0. However, this value occurs at both $x = 0$ and $x = 1$, so the problem has multiple solution points. Finally, in problem (35) we have the strangest situation. By selecting m and n carefully we can construct points with objective function values arbitrarily close to -0.25, but we can never achieve this value as it only occurs when $x = 0.5$ (which can never be created as a fraction with the denominator being odd). Thus this optimization problem is technically unsolvable.

To get around the issues arising in problem (35), mathematicians usually search for the *infimum* (alternately *supremum*) of a problem instead of the minimum (alternately maximum). The infimum of f over a set S, $\inf\{f(x) : x \in S\}$, is the highest lower bound for the problem. That is, a minimum value that does not necessarily have to be obtained.

In order to solve an optimization problem it is important to classify what type of problem it is. To begin, we differentiate between two important classes of optimization problems: continuous and discrete.

In many optimization problems the constraint takes the form of intervals in $\mathbb{R}$ or simple shapes in $\mathbb{R}^n$. The important point of this is that the constraint set does not consist of a list of isolated points, but instead consists of a type of continuum of points. Such problems are referred to as *continuous optimization problems*. Problem (33) is a continuous optimization problem. Conversely, in some optimization problems the constraint set takes the form of a list of possible solutions. This list may be finite, or infinite in length, but in either case elements are distinct in nature. Such problems are referred to as *discrete optimization problems*. Problems (34) and (35) are discrete optimization problems.

We now give descriptions of some of the more commonly arising types of optimization problems. In the simplest cases we actually describe how to solve the problem, but in most cases we only discuss how to recognize the type of problem, and then reference appropriate algorithms or computer software for the given problem.

3.1. Analytically Solvable Problems.

3.1.1. *Continuous Problems:* In the simplest case, where $f(x)$ is differentiable and S is $\mathbb{R}$ (the set of all real numbers) or a closed interval in $\mathbb{R}$, we can often resort to first year calculus. To solve these problems, first recall that the derivative of $f(x)$ at the point x represents the slope of the function at x. If the function is at a minimum (or maximum) then the slope at that point must be zero. Therefore,

to solve the problem we can simply differentiate $f(x)$ to get $\frac{d}{dx}f(x)$, find all the points where $\frac{d}{dx}f(x) = 0$, and compare the function values at these points. If S is a closed interval in $\mathbb{R}$ then we must also remember to check the endpoints of the interval as possible locations for the minimum.

To illustrate, let us apply this to $\min\{x^2 - x : 1 \leq x \leq 7\}$. First $f(x) = x^2 - x$, so $\frac{d}{dx}f(x) = 2x - 1$. Since $2x - 1 = 0$ only when $x = 0.5$, we must check 0.5 as a possible location of the minimum. Since x must be in the closed interval $1 \leq x \leq 7$ we must also check the endpoints 1 and 7. Checking these we find

$$f(0.5) = -0.25, \ f(1) = 0, \text{ and } f(7) = 42.$$

Although the minimum value in this list is -0.25, the point $x = 0.5$ is not feasible for the problem (as x must be greater or equal to 1). Therefore the minimum objective value is 0 and occurs at $x = 1$.

The mathematics described above can also be performed in multiple dimensions. Basically, points are replaced with vectors and derivatives are replaced with gradients. However, in higher dimensions we must be more careful with the edges of the constraint set, as they will not be just two points.

3.1.2. *Discrete Problems:* The other case where analytical methods may sometimes be applied is when the constraint set is a finite list of elements. If the list is small enough then we can simply perform an exhaustive search to determine the optimal answer. With the aid of a computer the exhaustive search can generally be automated, so even large finite lists can be approached in this manner. In practice, however, discrete optimization problems have so many elements in the finite list that, even with the aid of a computer, an exhaustive search would take years to complete.

3.2. Numerical Methods for Continuous Optimization Problems. Now suppose that although the optimization problem is continuous in nature, it is complicated enough that solving the problem analytically is not an option. This can result from working with a high number of dimensions, the objective function involving integrals, the constraint set taking a complicated form, or the objective function being non-differentiable (among many other possible reasons). In this case we must turn to numerical solving methods. Which method to select depends on the form of the problem.

3.2.1. *Linear Problems.* One of the simplest forms for a continuous optimization problem is that of a *linear program*[2]. A linear program is an optimization problem of the form

$$\min\{c^\top x : Ax = b\} \tag{36}$$

or

$$\max\{b^\top y : A^\top y \leq c\} \tag{37}$$

where b and c are column vectors and A is a matrix ($^\top$ denotes the transposition operation). For example,

$$\min\left\{5x_1 + x_3 : \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\right\},$$

[2]The word *program* instead of *problem* is historical dating back to world war II. The phrase *Linear Problem* is perfectly acceptable, but not often used by mathematicians.

is a linear program with

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},\ A = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 2 \end{bmatrix},\ b = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix},\ \text{and } c = \begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix}.$$

Mathematically, problem (36) is called the *primal problem* and problem (37) is called the *dual problem*.

The interesting thing about linear programs is that if A, b, and c are fixed and the problem (36) has a solution, then its solution has the same objective function values as problem (37). In fact, if x and y can be found where problem (36) and problem (37) take on the same value, then this is necessarily the solution.

This amazing fact is called *duality theory* and has resulted in some very powerful algorithms for solving linear programs. Using today's computers, linear programs with millions of variables can be solved in just hours.

3.2.2. *Quadratic and Semi-definite Problems.* The theory and abilities of linear programming can be extended to two broader classes of problems referred to as *quadratic programs* and *semi-definite programs*[3]. In quadratic programs the linear problem (36) is generalized by the addition of a quadratic term in the objective function. In semi-definite programming, the linear problem (36) is generalized in a manner that replaces the optimization of the vector x with optimization over a matrix X contained in the "semi-definite cone". Understanding the full meaning of this sentence is generally a graduate level course in mathematics, and well beyond the scope of this book. The important part is that many of the powerful optimization algorithms from linear programming have been generalized to these settings. In particular, semi-definite programming allows for easy solving of problems that involve quadratic objective functions or quadratic constraints (along with many other problems).

3.2.3. *Differentiable Convex Problems.* Let us return now to the more general structure of an optimization problem

$$\min\{f(x) : x \in S\}.$$

We shall define the set S to be *convex* if any two points contained in S can be joined by a straight line that never leaves S. That is, if $x_1 \in S$ and $x_2 \in S$ then $\lambda x_1 + (1-\lambda)x_2 \in S$ for all $0 \le \lambda \le 1$. We shall call the objective function *convex* if by drawing the function and shading in the area above the function we creates a convex set. That is, f is convex if $\{(x, \alpha) : \alpha \ge f(x)\}$ is a convex set. (The set $\{(x, \alpha) : \alpha \ge f(x)\}$ is called the *epi-graph* of f.)

Suppose the function f and the set S are both convex. Further suppose that f is differentiable and the gradient of f is readily available. Under these conditions we can solve the optimization problem $\min\{f(x) : x \in S\}$ using a wide variety of well-studied methods. Some generalized examples follow:

> **Steepest Descent:** At any point x, the gradient of f, $\nabla f(x)$, represents the direction that is directly "up-hill" from x. As such, to find the minimum we can repeatedly take steps in the direction $-\nabla f(x)$. This is called *steepest descent*.

[3]Like in linear programming, the word *program* is tradition. The phrase *semi-definite problem* is perfectly acceptable, but not often used by mathematicians.

Newton's Method: If the point x is a minimum for f then necessarily $\nabla f(x) = 0$. If we can find a second derivative of f (or even just approximate a second derivative of f), then we can apply Newton's root finding method to the function $g = \nabla f$.

Bundle Methods: By finding the function value f and the gradient value ∇f at a point x, we know roughly what the function looks like very close to x. By using these approximations for a large collection of points, we can create a piecewise linear approximation to the function f, which can be easily optimized. By refining the approximation function we can quickly locate the minimum of the function.

All of these methods are well studied, and proofs of convergence exist in many forms. Most of these methods are simple to program and many non-commercial codes exist. However, most of these codes are not user friendly, so figuring out how to use them can be a difficult task.

3.2.4. *Differentiable Non-convex problems.* If the function f or the set S is non-convex (that is, does not satisfy the above definitions), then optimization is faced with a difficult challenge. In particular, convex problems have the very satisfying property that if $\nabla f(x) = 0$ then the point x is a minimum for f. Non-convex problems do not satisfy this property, so it is very easy to find a point x that appears to be a minimum but is not. This problem is illustrated in Figure 1

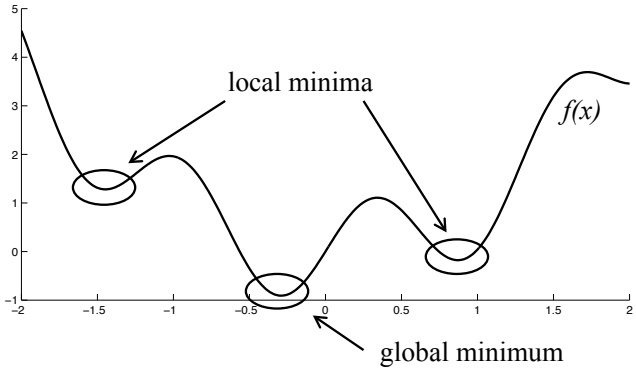

FIGURE 1. **The difficulty with non-convex optimization:** In a nonconvex problem we can often locate points which appear to minimize the function but in reality do not. In the graph above, a computer implemented optimization method could easily mistake one of the two local minima to be the global minimum.

There are many suggested methods for dealing with this problem, but realistically it is nearly impossible to guarantee optimality for non-convex problems. In practice, one of the best methods to deal with non-convex problems is to pretend they are convex. This is done by randomly generating a large collection of starting points and then running convex optimization methods on the problem starting at each point. Each starting point will result in convergence to one possible minimum, and taking the best of these results is a good guess at the optimal solution.

3.2.5. *Non-differentiable problems.* Another difficultly that may arise in applying an optimization algorithm to a problem is that the function f may not be differentiable, or the gradient of f is may not be readily available. In this case we turn to non-differentiable optimization methods.

In some cases, the gradient exists (and is calculable) at enough points, that even though the function is non-differentiable, differentiable methods can be applied. If this is not the case, then we usually resort to a form of *pattern search* (also called *direct search*). Pattern search is basically a formalized method for trial and error. We begin by selecting a number of points, and determining the objective function value for each. This information is then used to determine where to generate a new selection of points.

One of the most famous pattern search methods is the *Nelder-Mead* algorithm. Although many more recent methods provide improvements to this method, Nelder-Mead continues to be a popular choice due to its ease of implementation.

3.3. Numerical Methods for Discrete Optimization Problems. When decision variables can assume only discrete values from a specified set, the problem is called a *discrete optimization problem.* When that specified variable set is a set of integers, we call it an *integer program.* When the specified set consists of combinatorial structures (sets, subsets, permutations, partitions, Hamiltonian paths, or subgraphs), the problem is called a *combinatorial optimization problem.* Most combinatorial optimization problems may be formulated as integer programs, however this often results in the integer program formulation having an exponential number of constraints, so it is usually avoided.

On the surface, discrete optimization problems may sound easier than continuous optimization problems, as there are less possible answers, but in practice they are much more difficult. The difficulty lies in the fact that we can no longer lean on the mathematical power of functional analysis to help solve the problem. The result is that most optimization methods for discrete problems are based more on *heuristical* approaches than proven algorithms. However, there are a few *exact solution methods* for discrete optimization problems. We discuss these next, and then turn our attention to some of the heuristical approaches.

3.3.1. *Exact Solutions.* One of the most powerful theorems of calculus is the fact that the minimum of a differentiable function occurs at a point where the derivative is 0. This theorem is inapplicable in discrete optimization, as the discontinuity of the constraint set does not allow for the taking of derivatives. As a result, there are very few optimization algorithms for discrete optimization that are actually proven to converge.

The most basic method that always works is to test every possible element in the constraint set. If the constraint set is finite and small then this can be done quite easily with the assistance of a computer. However, in the majority of discrete optimization problems the constraint sets have an exponential number of elements. This means that solving the problem with n variables requires checking a multiple of 2^n solutions. Consider for example, a problem with 2^n elements in the constraint set, and suppose we can check one point every microsecond (1/1000 of a second). If we try the problem with ten variables we require just 1024 microseconds, approximately 1 second. If we try the problem with 20 variables, this number raises to 17 minutes. For 30, 40, and 50 decision variables the time jumps to 12 days, 34 years, and 3500 years respectively. By the time you reach 75 decision variables

checking all feasible solutions would require longer than the scientifically accepted age of the earth.

To make matters worse, some problems have $n!$ elements in the constraint set. Under the same circumstances (each solution requires one microsecond to check), instances with 5, 10, and 20 decision variables would require 1/10 of a second, 1 hour, and 70 million years, respectively.

Luckily, there are methods and techniques that avoid explicit exploration of all feasible solutions. The *branch and bound* method explores only a portion of the set of feasible solutions yet still guarantees the correct answer (when run to completion).

3.3.2. *Heuristics.* A heuristic is a reproducible method for improving one's knowledge on a problem. In optimization this means an algorithm that, when run on an optimization problem, will find a solution no worse than the best known solution. Good heuristical methods are those that usually produce sufficiently good results when applied in commonly occurring conditions.

Heuristics can be divided into construction heuristics and improvement heuristics. A *construction heuristic* builds a feasible solution to a problem in small steps, usually by "growing" a series of partial solutions to the problem. *Improvement heuristics* improve a feasible solution in a series of iterative steps.

Many of the improvement methods are local search methods. These are methods that iteratively search solutions near the best known feasible solution seeking some improvement. Stopping criteria for these algorithms are usually when there is no improving solution in the neighbourhood of the current solution. The disadvantage of these methods is that generally we have no way of knowing if we have found a global optimum, or simply a local optimum. Some of these modern heuristical methods include simulated annealing, evolutionary algorithms (also known as genetic algorithms), ant colony algorithms, and tabu search.

> **Simulated Annealing:** The ideas behind simulated annealing are derived from metallurgy, where annealing refers to the process of controlled heating and cooling of a metal in order to improve its properties, such as strength and hardness. In simulated annealing we begin with a collection of potential solutions and a "temperature gauge" T. If the temperature gauge is high, then the algorithm produces new candidate solutions that have a low relationship to current best candidate solution. If the temperature gauge is low, then the algorithm produces new candidate solutions that have a strong relationship to the current best candidate solution. The idea is that by controlled increase and decrease of the temperature gauge, we can focus in on good candidate solutions, while avoiding the tendency to become stuck at a local minima of the problem. Simulated annealing is popular for large discrete optimization problems.
>
> **Evolutionary Algorithms:** Also called *Genetic Algorithms*, evolutionary algorithms use ideas inspired from the theory of biological evolution. In particular, evolutionary algorithms consider candidate solutions to coincide with individuals in a population, and the objective function to be a measure of survivability for each individual. The algorithm generates many random candidate solutions (individuals) and then evaluates which solutions have the best survivability. These solutions are then allowed to breed (by taking slightly randomized averages between two candidate

solutions) and mutate (by adding random perturbations to a single candidate solution). Repeating this procedure many times, we hope to evolve towards an optimal solution. The weakness in such methods is the large amount of computing power required to perform evolutionary algorithms.

Ant Colony Algorithms: The ant colony algorithm is a heuristic method specifically designed for the study of route finding. That is, locating the shortest route that satisfies certain conditions (such as, determining the shortest route to get an ambulance to an emergency situation given we must stay on the road). The inspiration comes from the methods ant colonies use to locate the shortest path to a food source. Initially ants wander randomly until they locate food, then they return to the colony while laying down a pheromone trail for other ants to follow. Ants are attracted to the pheromone covered path, but pheromone slowly evaporates over time. Since, the more time it takes for an ant to travel down a given path the longer the pheromone has to evaporate, shorter paths are given preference. This preference reinforces the pheromone levels of the shortest path, and, with luck, convergence to the shortest route occurs. Ant colony algorithms mimic this behaviour with simulated ants and pheromones. Such algorithms have been relatively successful in solving difficult route finding problems, such as the traveling salesman problem. However, like all heuristic methods they have no assurance of locating the optimal solution.

Tabu Search: Tabu search can be thought of as an add-on to any pattern search method (see Subsection 3.2.5). To begin, a pattern search method is applied in order to find a local optimum. In order to avoid becoming stuck at a local optimum (instead of finding the global optimum), the tabu search occasionally defines a local area as "tabu," thus forcing iterations to move away from this area, even if this causes movement towards a worse solution. Over time the tabu of a given area is removed, allowing the pattern search to return there if no better solution has been located. Tabu search has been shown relatively effective at helping us avoid becoming stuck at the first local optimum we encounter. Thus it often improves the final result. However, for nonconvex problems it is still quite possible to end in a local optimum instead of a global one.

Empirical studies have shown that many heuristics may be very successful. More importantly, for many discrete optimization algorithms heuristical methods are the only practical option.

3.4. Dynamic Optimization Problems. At times, the optimization problems arising in healthcare may have dynamic components (i.e. the data changes with time) and should be modelled by a dynamic optimization problem. A *dynamic optimization problem* is one where the problem changes as new data becomes available. At each time step, the optimizer must output a solution to the problem that provides a good level of optimization and allows for flexibility when new data arrives. A prime healthcare example is emergency vehicle dispatching (ambulances must be dispatched in a manner that retrieves patients in a somewhat optimal time, but reserve the flexibility for new calls to change dispatch priorities).

This is a relatively new field of optimization so little can be said about the best methods to approach such problems. Many researchers solve dynamic optimization

problems using what are called *online algorithms.* These algorithms are typically a blend of exact and heuristical methods. Another option includes the idea of solving a problem over a limited rolling time horizon.

4. Examples

4.1. Nurse Scheduling as a Linear Program. The nurse (physician, surgeon, etc.) scheduling problem deals with finding the minimum number of nurses required in a department so that patient needs are met. In this example we demonstrate how nurse scheduling can be approached as a linear program.

Assume that the resources needed in the department are constant over successive intervals of 4 hours each, and that particular needs involve the following: 4 nurses are needed in the department between 8 am and noon, 8 between noon and 4 pm, 10 between 4 pm and 8 pm, 7 between 8 pm and midnight, 12 between midnight and 4 am, and 4 between 4 am and 8 am.

Consider first a situation in which there is a three-shift schedule (8 am–4 pm, 4 pm–midnight, and midnight–8 am). After introducing decision variables x_1, x_2, and x_3 that represent the number of nurses in each of the three shifts, the optimization problem becomes:

$$\begin{array}{lrcrcrcr} \min & x_1 & + & x_2 & + & x_3 & & \\ \text{subject to} & x_1 & & & & & \geq & 8 \\ & & & x_2 & & & \geq & 10 \\ & & & & & x_3 & \geq & 12 \end{array}$$

Examining this problem it is easy to see that the optimal solution is obtained when $x_1 = 8, x_2 = 10$, and $x_3 = 12$. Thus the minimum number of nurses required is 30.

It is interesting to observe that if we model the same nurse scheduling problem by a slightly different mathematical programming formulation, we may be able to achieve a better solution. Specifically, let us allow for a six-shift schedule where the starting shift times can be 8 am, 12 pm, 4 pm, 8 pm, 12 am, or 4 am, and all shifts are 8 hours long. The new optimization problem is:

$$\begin{array}{lrcrcrcrcrcrcll} \min & x_1 & + & x_2 & + & x_3 & + & x_4 & + & x_5 & + & x_6 & & & \\ \text{subject to} & x_1 & + & & & & & & & & & x_6 & \geq & 4 & (8\text{am} - 4\text{pm shift}) \\ & x_1 & + & x_2 & & & & & & & & & \geq & 8 & (12\text{pm} - 8\text{pm shift}) \\ & & & x_2 & + & x_3 & & & & & & & \geq & 10 & (4\text{pm} - 12\text{am shift}) \\ & & & & & x_3 & + & x_4 & & & & & \geq & 7 & (8\text{pm} - 4\text{am shift}) \\ & & & & & & & x_4 & + & x_5 & & & \geq & 12 & (12\text{am} - 8\text{am shift}) \\ & & & & & & & & & x_5 & + & x_6 & \geq & 4 & (4\text{am} - 12\text{pm shift}) \end{array}$$

This problem is slightly more difficult to solve, but it can be easily checked that $x_1 = 0, x_2 = 10, x_3 = 0, x_4 = 12, x_5 = 0$, and $x_6 = 4$ is a feasible solution to the problem (in fact this is the optimal solution). Notice this only requires 26 nurses, so rewriting the problem has saved 4 nurses.

In both of these examples the optimal solution was fortunately formed of integers (we never had to employ half a nurse). This is a result of the artificial nature of the example. Realistic problem parameters (number of nurses required during the day) would likely result in a non-integer optimal solution. Various remedies

for this exist, one of which is to consider the problem as a discrete optimization problem.

4.2. Dispatching Ambulance Vehicles. The efficiency of emergency medical services in reducing mortality is strongly related to the time needed by a paramedic team to arrive at the scene. This time largely depends on the location of the nearest ambulatory service. With the development of new telecommunication and computer technologies, comes the opportunity to collect real-time data that can be used to improve ambulance response times. As such recent research has explored methods of optimizing the ambulance allocation and the ambulance redeployment problems. Both problems examine methods to maximize the expected number of calls covered within a given time frame, subject to constraints such as number of vehicles and potential locations of vehicle bases.

The ambulance allocation problem consists of determining which ambulance should be sent to answer a given call. In a mathematical sense, we would weight the time it would take a given ambulance to reach a call against the probability that the ambulance would be needed elsewhere, and determine which ambulance is the best to send in terms of minimizing total expected response time for all calls (not just response time to the given call). However, in a practical sense, society would not accept such a solution, so the ambulance allocation problem is generally solved by the simple rule: "send the nearest ambulance".

Given that we are not able to tamper with the ambulance allocation problem, we must turn to the ambulance redeployment problem to improve response time to calls. The ambulance redeployment problem consists of relocating available ambulances to potential location sites when calls are received. Basically, when one ambulance moves to answer a call, the remaining ambulances shift locations in order to ensure maximal coverage of the service area.

In 1983 Daskin proposed a method to write the ambulance redeployment problem as a linear program that would be resolved every time either a new ambulance became available (i.e. finished responding to a call) or an ambulance became unavailable (i.e. left to answer a call) [**60**]. Although this approach provided a theoretically sound solution to the problem, it required the problem to be solved in real-time. Furthermore, to follow the framework of this approach, we was required to redeploy ambulances each time a call arrived. Ambulance crews were dissatisfied with this approach.

In order to improve on this approach, we move to dynamic optimization methods. In order to avoid ambulances being in continuous motion, we must seek *robust* solutions. These are, solutions that may not be perfectly optimal at any given moment, but are reasonably optimal in a large variety of situations. This remains a difficult problem, but some approaches have been proposed. For example, the inclusion of some (or all) of the following constraints help improve the robustness of the redeployment problem:

- only a limited number of ambulances can be moved when a redeployment occurs;
- vehicles moved in successive redeployments cannot always be the same;
- repeated round trips between two location sites must be avoided;
- long trips between the initial and final location sites must be avoided;
- an assignment to a call should be avoided near the end of a working shift; and

- at the end of a shift, the ambulance has to be moved closer to the central service point where the vehicles are based.

In [**84**], this problem is approached via a parallel tabu search heuristic. The main component of this algorithm is the pre-computation of redeployment scenarios that allows immediate decision-making when calls are received. Simulations based on real data have confirmed the efficiency this approach.

Another approach is provided by [**106**], who adds the constraint

- ambulances may only be redeployed when completing a call,

to the problem. This constraint is satisfying to ambulance crews and provides robustness to the problem at the cost of efficiency. The redeployment problem was solved using simulation and tested in several cities world wide with reasonable success.

4.3. Shared Transportation Problem as a Discrete Optimization Problem. In this example, we consider the problem of how to deliver transportation services to disabled and elderly people. This is an example of the *shared transportation problem*, sometimes called the *dial-a-ride problem*.

In this problem we assume that each healthcare program has been put in place to help disabled and elderly people meet their transportation needs. The program is accessed by a patient phoning in a request to be picked up and transported to a health client at a certain time. To save costs, transportation is shared, and consists of a fleet of buses serving transport requests. The overall optimal schedule can be maintained even if some trips have to follow longer routes, as long as customers are picked up and dropped off on time.

The problem can be modelled by the *pickup and delivery problem with time windows* (PDPTW). The PDPTW is a vehicle routing problem that deals with finding an optimal set of routes for a fleet of vehicles in order to serve a set of *transportation requests*. A transportation request is defined by a pair of locations: a person or a package has to be picked up at the pickup location and delivered at the delivery location. Each location is associated with a specific time interval allocated for the visit to that location. This interval is known as the *time window* of the location. Each vehicle has a capacity constraint. The solution to the problem consists of a set of routes and schedules. A route is a sequence of locations to be served by one vehicle. A schedule for a route is the sequence of times when each location on the route will be serviced.

More formally, the problem may be described as follows. Let the number of vehicles be m and the number of customers be n. We enumerate the vehicles $V = \{1, 2, \ldots, m\}$. Each vehicle v has a start location $s(v)$ and an end location $e(v)$, usually called depots. Since there are n customers we have n pick-up locations and n drop off locations. We label these $P^+ = \{p^+(i) : i \in 1, 2, ...n\}$ and $P^- = \{p^-(i) : i \in 1, 2, \ldots, n\}$, where $p^+(i)$ is the pick-up location for customer i and $p^-(i)$ is the drop-off location for customer i. Let N be the set of all locations: $N = P^+ \cup P^- \cup \{s(v) : v \in V\} \cup \{e(v) : v \in V\}$. We enumerate N in the order

$$\begin{aligned} N = \quad & \{p^+(1), p^+(2), \ldots, p^+(n), p^-(1), p^-(2), \ldots, p^-(n), \\ & s(1), s(2), \ldots, s(m), e(1), e(2), \ldots, e(m)\}, \end{aligned}$$

so a customer picked-up at location $i \in N$ is dropped-off at location $i + n \in N$.

The required time window for location $i \in N$ is $[a_i, b_i]$, meaning the customer must be picked-up/dropped off between time a_i and b_i. For each two distinct stop

locations $i, j \in N$, we let $t_{i,j}$ and $c_{i,j}$ represent direct travel time and travel cost from location i to location j.

Let the maximum load of vehicle v be Q^v.

Three types of variables are used in this mathematical formulation: binary flow variables $X^v_{i,j}$, time variables T_i, and load variables L_i. The binary flow variable $X^v_{i,j}$ has value 1 if vehicle v travels from node i to node j. The time variable T_i is the time when node i is serviced, and the load variable L_i is equal to the load in the vehicle after servicing node i.

The optimization problem for this PDPTW is the integer program given below.

$$\text{minimize} \quad \sum_{v \in V} \sum_{i,j \in N} c_{i,j} X^v_{i,j} \tag{38}$$

$$\text{subject to} \quad \sum_{v \in V} \sum_{j \in N} X^v_{i,j} = 1, \qquad i \in P^+ \tag{39}$$

$$\sum_{j \in N} X^v_{i,j} - \sum_{j \in N} X^v_{j,i} = 0, \qquad i \in P,\ v \in V \tag{40}$$

$$\sum_{j \in P^+} X^v_{s(v),j} = 1, \qquad v \in V \tag{41}$$

$$\sum_{i \in P^-} X^v_{i,e(v)} = 1, \qquad v \in V \tag{42}$$

$$\sum_{j \in N} X^v_{i,j} - \sum_{j \in N} X^v_{j,n+i} = 0, \qquad i \in P^+,\ v \in V \tag{43}$$

$$T^v_i + t^v_{i,n+i} \le T^v_{n+i}, \qquad i \in P^+,\ v \in V \tag{44}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ T^v_i + t^v_{i,j} \le T^v_j, \qquad i, j \in P,\ v \in V \tag{45}$$

$$X^v_{s(v),j} = 1 \ \Rightarrow\ T^v_{s(v)} + t^v_{s(v),j} \le T^v_j, \qquad j \in P^+,\ v \in V \tag{46}$$

$$X^v_{i,e(v)} = 1 \ \Rightarrow\ T^v_i + t^v_{i,e(v)} \le T^v_{e(v)}, \qquad i \in P^-,\ v \in V \tag{47}$$

$$a_i \le T^v_i \le b_i, \qquad i \in P,\ v \in V \tag{48}$$

$$a_{s(v)} \le T^v_{s(v)} \le b_{s(v)}, \qquad v \in V \tag{49}$$

$$a_{e(v)} \le T^v_{e(v)} \le b_{e(v)}, \qquad v \in V \tag{50}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ L^v_i + l_j = L^v_j, \qquad i \in P,\ j \in P^+,\ v \in V \tag{51}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ L^v_i - l_{j-n} = L^v_j, \qquad i \in P,\ j \in P^-,\ v \in V \tag{52}$$

$$X^v_{s(v),j} = 1 \ \Rightarrow\ L^v_{s(v)} + l_j = L^v_j, \qquad j \in P^+,\ v \in V \tag{53}$$

$$L^v_{s(v)} = 0, \qquad v \in V \tag{54}$$

$$0 \le L^v_i \le Q^v, \qquad i \in P^+,\ v \in V \tag{55}$$

$$X^v_{i,j} \in \{0, 1\}, \qquad i, j \in N,\ v \in V \tag{56}$$

(57)

Constraint (39) states that each pickup location is left by exactly one vehicle. Constraint (40) means that the number of vehicles coming to location j is equal to the number of vehicles leaving location j. Constraints (41) and (42) ensure that each route starts with a pickup location and ends with a delivery location, not counting depots. Constraint (43), called the pairing constraint, deals with the fact that each

pickup location and its corresponding delivery location have to be served by the same vehicle. Less formally, a patient will be driven by one vehicle. Constraint (44), called precedence constraint, ensures that each pickup site is located before its corresponding delivery location — in other words, patients have to be picked up before they can be dropped off. Constraints (45)–(47) represent compatibility between routes and schedules and constraints (48)–(50) are time window constraints ensuring that each location is served within its own time window. Constraints (51)–(53) represent compatibility between routes and vehicle capacity and constraints. Constraints (54)–(55) are capacity constraints ensuring that no vehicle is filled above capacity, or has an occupancy that is negative.

The above problem is extremely complicated, and clearly cannot be solved by hand. Research by Savelsbergh and Sol has shown that branch and bound methods can been successfully employed to solve this problem [**194**].

5. Related Reading

Reference [**165**] is a recent textbook on applied optimization algorithms. Reference [**195**] overviews the basics of linear programming, including the classic simplex method. Reference [**187**] explores some of the more advanced methods for solving linear programs. References [**221**] and [**220**] discuss the advancement of linear programming into semi-definite programming.

Reference [**33**] explains a large collection of methods for convex optimization, it is available online at `http://www.stanford.edu/~boyd/cvxbook/`. Reference [**31**] covers some more recent advancements in numerical techniques for convex optimization.

Reference [**162**] presents the classic Nelder-Mead algorithm. Reference [**229**] provides a survey of the advancements in pattern search algorithms up to 1995. Reference [**49**] is a survey paper outlining the general framework of pattern search methods. Reference [**137**] is an example of recent research and development in pattern search methods. Reference [**219**] develops the tools needed to implement a pattern search method using parallel computing. Reference [**10**] details a pattern search method for constrained optimization problems.

Reference [**127**] discusses optimization by simulated annealing. Reference [**11**] examines evolutionary algorithms. Reference [**67**] is a survey of the mathematics behind ant colony algorithms. Reference [**86**] explores the meta-heuristical approach of tabu search.

Reference [**94**] discusses the current state of online and real-time optimization with the aim of reaching a broad audience. Reference [**183**] examines the idea of solving dynamic optimization problems over a limited rolling time horizon.

Reference [**194**] discusses characteristics that separate pickup and delivery problems from standard vehicle routing problems as well as problem types and solution methods, and is discussed in Example 4.3.

Reference [**60**] is one of the first models for examining optimal ambulance deployment, and discussed in Example 4.2. Reference [**84**] looks at the problem of redeployment of a fleet of ambulances using a dynamic model that includes a parallel tabu search heuristic to compute redeployment scenarios, and is discussed in Example 4.2. Reference [**106**] also examines the problem of ambulance redeployment, and is discussed in Example 4.2. Reference [**36**] traces the evolution of ambulance location and relocation models from 1973 to 2003. Reference [**87**] reviews the development and current state of operations research for deployment and planning analysis pertaining to Emergency Medical Services and Fire Departments. Reference [**7**] describes some decision support tools for dynamic ambulance relocation and automatic ambulance dispatching.

APPENDIX A

Computer Programming Packages Useful in Modelling

There are numerous pieces of computer software that can be of assistance in modelling. In this appendix we provide a list of some of the more popular programs, along with a brief description of each. The appendix is divided into three sections, the first focusing on statistical software packages, the second on general mathematical packages, and the third on computer software designed to run specific models.

1. Statistical Software

Microsoft Excel: Microsoft Excel is one of the most popular spreadsheet programs for personal computers today. It contains many basic statistical commands including linear and logistic regression. However, statistics is limited by the maximum spreadsheet size and the program's fairly slow operating speed. It is best suited for producing basic summary statistics from small sample sizes. It is available from Microsoft™for both Microsoft Windows and Mac OS X, `http://office.microsoft.com/`.

S, R, and S-PLUS:. S is a powerful environment for the statistical and graphical analysis of data. It provides the tools to implement all basic statistical methods along with many advanced methods such as generalized linear regression and time series analysis. S-PLUS is a commercial implementation of S developed and supported by Insightful Corporation™, `http://www.insightful.com/`. R is a free open source version of S used in many academic settings. It is available at `http://www.r-project.org/`. S-PLUS is available for Microsoft Windows and Unix, while R is available for Microsoft Windows, Mac OS X, and Linux.

SAS:. SAS, originally the "Statistical Analysis System," is one of the most popular operations management tools in industry. Although not specifically designed as a statistical software package, most versions include a large variety of advanced statistical packages. It is available from S.A.S.™for Microsoft Windows, Mac OS X, and Unix operating systems, as well as a variety of mainframe installations, `http://www.sas.com`.

2. Mathematical Software

Berkeley Madonna: Berkeley Madonna is a multi-purpose numerical differential equation solver. It is able to solve systems of differential equations rapidly and graphically display the system evolution at run time. However it has limited uses outside of differential equations. It is available for both Microsoft Windows and Mac OS X. A free trial download is available at `http://www.berkeleymadonna.com/`.

Maple: Originally developed at the University of Waterloo, Maple is now one of the preeminent software packages in mathematics. It is a good general purpose computer algebra system capable of both symbolic and numerical analysis of functions, systems of equations, and systems of differential equations. It also comes with a limited optimization library and support for many other branches of mathematics. One of its few drawbacks is that many of its libraries are not loaded as compiled code so it can run quite slowly. This is rapidly being corrected in newer versions of the software. Maple is available from MapleSoft^{TM}for Microsoft Windows, Linux, and Mac OS X : `http://www.maplesoft.com/`.

Mathematica: Mathematica is a full featured suite for the mathematical sciences. It has extensive capabilities in both computer algebra calculations and numerical analysis. It is available from Wolfram Research^{TM}for Microsoft Windows, Mac OS X, Linux, Solaris, and most other unix operating systems `http://www.wolfram.com/`.

MATLAB:. Due to its superior numerical algorithms, MATLAB is one of the most widely used software packages for numerical analysis in applied mathematics, science, and engineering. It has a large number of toolkits supporting a wide range of mathematical specialization. Its popularity in the mathematical sciences has also resulted in numerous freely available toolkits designed and implemented by academics. It is developed and supported by MathWorks^{TM}and is available for Linux, Solaris, Mac OS X, and Microsoft Windows, `http://www.mathworks.com/`.

3. Simulation and Modelling Codes

Modelling Languages

ANML:. (Another Modelling Language) is a general purpose modelling language for describing various systems such as communication networks. The language, which is object-oriented, consists of three general constructs: models, schemas and databases. Models are descriptions of specific system scenarios, schemas specify the rules for creating models, and databases serve as a repository of components for easy reuse in different models. ANML is based on the Domain Modelling Language, which was developed as part of the Scalable Simulation Framework and the Extensible Markup Language. ANML was developed at the University of Calgary, the open source is available at `http://warp.cpsc.ucalgary.ca/Software/ANML/anml.php`.

dML:. (deX Modelling Language) is an object-oriented language based upon C++, which is part of the deX modelling package. dML is designed to facilitate the development of parallel simulations, either on multi-processor systems or on clusters.

GPSS:. This was the first simulation programming language. It was developed at IBM and appeared in 1961. Strictly speaking, GPSS is not a complete programming language, but rather a set of FORTRAN routines. GPSS programs follow a block-diagram representation, which represent the process flow in the simulation. It is particularly well-suited to modelling queueing problems and remains popular to this day. A commercial GPSS compiler is currently available from Wolverine Software. In addition, a MATLAB toolkit is available for processing GPSS simulations in MATLAB.

PARSEC:. Developed by the Parallel Computing Laboratory at UCLA, it is an acronym for "PARallel Simulation Environment for Complex systems". It is

a C-based discrete-event simulation language that adopts the process interaction approach to discrete-event simulation. PARSEC provides support for executing a discrete-event simulation model using several different asynchronous parallel simulation protocols on a variety of parallel architectures. As such, it is well-suited for deploying discrete-event simulations on computer clusters. PARSEC is freely available for academic use.

SHIFT:. This is a programming language for describing dynamic networks of hybrid automata, which may exhibit both discrete and continuous behaviour. The components interact via a network, which may itself evolve in time. SHIFT is well-suited for applications such as automated highway systems, air traffic control systems, robotic shopfloors, and similar systems whose operation cannot be captured easily by conventional models. SHIFT was developed by the California PATH (California Partners for Advanced Transit and Highways) Project at the University of California, Berkeley and both a compiler and a run-time system are freely available.

SIMSCRIPT I/II/III:. Simscript is an "english-like" high level simulation language designed for discrete-event and hybrid discrete/continuous modelling. The first version, SIMSCRIPT I, was developed by the RAND Corporation for the US Air Force and released in 1962. This initial version was implemented as a FORTRAN preprocessor, producing FORTRAN code that was then subsequently compiled with a FORTRAN compiler. SIMSCRIPT II, which was also developed by the RAND Corporation, was released in 1968. The CACI Products company currently sells and supports a commercial version called SIMSCRIPT II.5. They have also recently released SIMSCRIPT III, which extends SIMSCRIPT II to provide full support for object-oriented programming.

Simula: First released in 1965, Simula is a full-featured object-oriented programming language designed for discrete-event simulations. It introduced object-oriented concepts and has had great influence on all modern class-based, object-oriented programming languages. Simula remains in use today and Cim is a currently available compiler for it. Cim is implemented as a Simula to C converter, followed by compilation of the C code using a C compiler. It is open source and licensed under the GPL. It should run under most unix-like operating systems.

SLX:. SLX is a new simulation language from Wolverine Software. It utilizes a layered approach, starting with the SLX kernel at the bottom, a traditional simulation language in the middle, and more application-specific dialects at the top. Thus, simulation programmers may construct models using the upper layers of language and are not required to delve into the lower-level details. However, models requiring features not available at the upper level may still be modelled by using lower-level programming to build new higher-level constructs. In addition to its multilayered structure, another focus of SLX is to provide support for DES models with parallel processes.

Libraries and Application Program Interfaces

baseSim: baseSim is a simulation library for Borland Delphi. It supports a variety of discrete-event simulation models, including Monte Carlo simulation models. It is sold and supported by iBright Ltd.

C++Sim: This is an object-oriented C++ simulation library developed at the University of Newcastle upon Tyne. It provides SIMULA-like simulation routines,

random number generators, queueing algorithms, and thread package interfaces. This library is freely available for teaching and research use.

CSIM:. CSIM is a commercial library of simulation routines for C or C++. It provides routines for discrete-event simulation models of complex systems. CSIM is a product of Mesquite Software.

DESMO-J:. DESMO-J ("Discrete-Event Simulation and Modelling in Java") is an object-oriented framework targeted at programmers developing simulation models in Java. It provides support for building models with a graphical user interface. Developed at the University of Hamburg, DESMO-J is part of the larger Eclipse project for developing open source modelling and simulation tools. DESMO-J is licensed under the Apache License.

DSOL:. DSOL is a Java-based suite for continuous and discrete-event simulation. The focus of DSOL is to view simulation as a set of loosely-coupled, web-enabled services. As such, DSOL focuses on the development of simulation models with web-based interfaces. DSOL was developed at the Delft University of Technology and is open source.

RedShift: RedShift is a simulation library written in Ruby and C. Its syntax is based on that of SHIFT and Lambda-SHIFT. RedShift is open source and freely available.

Simulación 4.0: Simulación 4.0 is a visual basic library designed to provide support for simple simulations within Microsoft Excel. Simulación 4.0 is freely available.

SimPy: SimPy is an object-oriented, process-based, discrete-event simulation library for Python. It provides the modeller with components of a simulation model including processes (for active components such as customers, messages, and vehicles) and resources (for passive components such as servers, checkout counters, and tunnels). It also provides monitor variables to aid in gathering statistics. SimPy is an acronym for "Simulation in Python". It is open source and released under the Lesser GNU Public License.

SSF:. SSF (Scalable Simulation Framework) is an open standard for a discrete-event simulation application program interface (API). SSF is designed to support parallel simulations that support very large collections of simulation entities running on a computational cluster. Implementations of SSF in both C++ and Java are freely available. Furthermore, the SSF specification is defined in an abstract manner, allowing it to function as a model for high-level modelling languages or graphical modelling environments. Associated with SSF is the Domain Modelling Language (DML), which is an open standard for defining model configurations.

Simulation Engines

JiST:. JiST ("Java in Simulation Time") is a high-performance discrete-event simulation engine that runs over a standard Java virtual machine. The JiST system architecture consists of four distinct components: a compiler, a byte-code rewriter, a simulation kernel and a virtual machine. JiST simulations are written in standard Java and compiled to byte-code using a regular Java language compiler. These compiled classes are then modified by JiST using a byte-code-level rewriter to run over a simulation kernel. This architecture provides exceptional performance and also allows for efficient parallelization execution on large clusters. JiST was developed at Cornell University and is freely available for academic use.

SimKit: The SimKit engine is based on an object-oriented, logical-process view of discrete-event simulation. In a SimKit simulation, each physical process is characterized by a logical process. The logical processes communicate by exchanging messages called events. SimKit implementations in both C++ and Java are available. It is open source and was developed at the University of Calgary.

WARPED:. WARPED is a highly parallel simulation engine, implemented in C++. WARPED makes extensive use of the object-oriented structure of C++ and it defines a number of specialized classes for simulation. A variety of libraries are also available. Included with WARPED is KUE, a library for building queueing models. WARPED was developed at the University of Cincinnati and has been released into the public domain.

Simulation Packages

AnyLogic: AnyLogic is a Java-based simulation package from XJ Systems. It has support for stochastic modelling, interactive 2-D and 3-D animation, as well as optimization. A number of different modelling paradigms are available with AnyLogic, including process flow diagrams, system dynamics, agent-based modelling, and state charts. AnyLogic runs under Microsoft Windows 2000 or Windows XP.

Arena: Arena is based on the SIMAN simulation language. However, its user interface is completely graphical. Models are built from graphical objects called modules. Modules are then organized into structures called templates. A number of standard templates are included with Arena. Arena is developed and supported by Rockwell Automation, Inc. It is available for Microsoft Windows. In addition to Arena, Rockwell Automation also offers OptQuest, an optimization suite for Arena.

AutoMod: Automod is a graphical modelling package that provides true to scale 3-D virtual reality animation, making simulation models easy to understand and explain. It uses CAD-like features to define the physical layout of manufacturing, material handling, and distribution systems. Although primarily designed for operations analysis of manufacturing systems, it may also be applied to a variety of other types of simulation modelling. AutoMod is available from Brooks Software and it runs under Microsoft Windows.

eM-Plant: This package focuses on the modelling and simulating of production systems and processes. It provides the capability to optimize material flow, resource utilization, and logistics. It takes an object-oriented approach to model design and has extensive features for visualization and animation of models. eM-Plant is developed and supported by UGS, and runs under Microsoft Windows.

Enterprise Dynamics: This is an object-oriented dynamic analysis and control package, available from Incontrol Enterprise Dynamics. Model design is based on blocks and templates. It supports both 2-D flowchart animation and 3-D models. Enterprise Dynamics runs under Microsoft Windows.

Extend: This simulation suite from Imagine That Inc. supports both discrete-event simulation and numerical solutions of systems of differential equations. In 1988, Extend was the first simulation package to introduce a graphical interface utilizing a block-diagram approach to model-building. There is also support for animation of the process flow diagram. A C-like programming environment is available for defining new blocks. Extend runs under the Microsoft Windows and Mac OS X.

iThink and Stella: iThink and Stella from ISEE Systems use a graphical systems dynamics approach to model design. The models are designed graphically using flow diagrams to show process flow and feedback loops. iThink and Stella are similar packages, with iThink focused more on business applications and Stella focused towards education and research. Both packages run under Microsoft Windows and Mac OS X.

JSIM. This is a Java-based simulation and animation environment that focuses on web-based simulation. It was developed at the University of Georgia. Simulation models may be built using either the event package (event-scheduling paradigm) or the process package (process-interaction paradigm). In addition, a graphical design interface allows process models to be be rapidly built graphically. JSIM is open source and licensed under a BSD-style license. It requires Java 5.0.

SimuLink: Simulink is a graphical environment for developing models of dynamical systems produced by MathWorks. Essentially it provides a graphical interface to MATLAB. SimEvents is an extension to SimuLink, which is specialized for discrete-event simulation. SimuLink is available for Linux, Solaris, Mac OS X, and Microsoft Windows.

MicroSaint: MicroSaint is a graphical discrete-event simulation package from Micro Analysis and Design, Inc. It provides two graphical views of the simulation: a network flow diagram and a 2-D animation of the model. Support for optimization is also provided by the OptQuest module, which is included with the package. The package runs under Microsoft Windows.

OMNeT++: This is a discrete-event simulation package with strong GUI support. It utilizes a modular open-architecture, making it flexible and easy to extend. Its initial application area was the simulation of communication networks, however, it is now widely used to simulate queueing networks, IT systems, and business processes. A variety of examples are included with the software distribution. OMNeT++ runs under most versions of unix-like operating systems, including Linux, Mac OS X, and FreeBSD, as well as under Microsoft Windows. It is open source and free for academic use. Commercial use requires a license from SimulCraft, Inc.

Ptolemy Project: Based at the University of California at Berkeley, this project is developing an open source software system for modelling, simulation, and design of concurrent, real-time systems. Their main focus is on models that have concurrently executing components, which interact with each other. Ptolemy has a graphical interface for constructing simulation models using block diagrams. It supports data-flow, discrete-event simulation, process networks, and finite-state machine models of computation. Ptolemy is written in C++ and compiles using gcc on most unix-like operating systems. A follow-up Java-based system called Ptolemy II is under development.

SansGUI:. This is a modelling and simulation environment for developing and deploying scientific and engineering simulators without the need for writing any graphical user interface code. SansGUI supports the development of graphical interfaces for models that are implemented in Microsoft Visual C/C++, Compaq Visual Fortran 6.1, or any programming language or environment that can generate Win32 DLLs callable by Microsoft Visual C/C++. This includes simulations implemented in MATLAB. SansGUI is available from ProtoDesign, Inc, and runs under Microsoft Windows.

SDX:. This is a Fortran-based problem solving environment for both continuous and discrete dynamical systems. Typical applications include aerospace, applied mathematics, biological systems, and control systems. It is available from Eclipse Software and runs under Microsoft Windows.

ShowFlow. This is a simulation software package that places a strong emphasis on graphically representing the simulation from a systems dynamics perspective. It is developed by Webb Systems Ltd. and runs under Microsoft Windows.

Simile: This is a simulation package for complex dynamical systems in the earth, environmental and life sciences. It is developed and supported by Simulistics Ltd. Models are developed graphically using a system dynamics approach. Simile then converts the graphical representations into C++ code that is compiled and executed. Simile runs under Microsoft Windows, Mac OS X, Linux, and FreeBSD.

SIMPROCESS:. This is an integrated process-based simulation software package based on SIMSCRIPT II.5. It provides a graphical interface for constructing a discrete-event simulation model using processes, activities, entities, resources, and connectors. SIMPROCESS is available from CACI Products. It runs under Microsoft Windows, Linux, and Solaris.

SIMUL8: The focus of SIMUL8 is to provide a simulation package that is easy to use for non-experts. It relies heavily on a graphical interface, with models being constructed by assembling graphical building blocks. Templates are used to allow a variety of common simulation scenarios to be developed easily. The simulation model and data are saved in XML, for easy access by other programs and web interfaces.

Traffic: This is a suite of programs for the analysis of queue models. It is distributed and supported by Erlang Software. A source code license may be purchased. The Traffic programs run under Microsoft Windows, Linux, and Solaris. With a source code license, it should compile under most unix-like operating systems.

WITNESS:. This package uses a graphical interface with modules and templates for model design. The Lamner Group offers two versions of WITNESS: one focused on modelling in the service sector and the other focused on the manufacturing sector. The models may be displayed in 2-D animation. A variety of extension modules are available, supporting extra capabilities such as 3-D animation, CAD design, and optimization. WITNESS runs under Microsoft Windows.

Packages for Agent-Based Simulation

Brahms: This is a modelling language, composer, compiler, virtual machine, and simulation viewer for agent-based simulations. This system is distributed by Agent Solutions and licensed to NASA. It is free for academic and research use. It is currently not available for commercial use. Brahms runs under Microsoft Windows, Linux, Solaris, and Mac OS X.

Ps-i: This is a Tcl/Tk based environment for agent-based simulations. It has built-in routine optimization for improving simulation performance, plus the ability to change model parameters while the simulation is running. Ps-i is open source and licensed under the GNU Public License.

SeSAm: Developed at the Universität Würzburg, the Shell for Simulated Agent Systems (SeSAM) provides a generic environment for agent-based simulation. It provides easy visual agent modelling, flexible environment and situation definitions,

an integrated graphical simulation analysis, and the ability to run distributed simulations on a cluster. Furthermore, SeSAM is a full programming language. The package is able to deal with complex multi-agent systems simulations for complex models with flexible agent behaviour and interactions. SeSAM is open source and licensed under the Less GNU Public License. It requires Java 5.0 or better.

SimWalk: This is an agent-based simulation package primarily targeted at modelling pedestrian flows in complex environments. However, it can also be used to model marketing scenarios and other types of social interactions.

Bibliography

[1] J.J. Abramson and Z.H. Abramson. *Survey Methods in Community Medicine: Epidemiological Research, Programme Evaluation, Clincal Trials*. Churchill Livingstone, London, UK, 1999.

[2] E. Akcali. A network flow model for health care resource planning. In *INFORMS Optimization Society Conference: Optimization and Health Care*, San Antonio, Texas, February 3-5 2006.

[3] G Albrecht, S Freeman, and Higginbotham N. Complexity and human health: The case for a transdisciplinary paradigm. *Culture Medicine and Psychiatry*, 22(1):55 – 92, 1998.

[4] Azadeh Alimadad, Peter Borwein, Vahid Dabbaghian-Abdoly, Ron Ferguson, Ellen Fowler, Yuri Gusev, Michael Hayes, Warren Hare, Michel Joffre, Snezana Mitrovic-Minic, A. R. Rutherford, Krisztina Vasarhelyi, and Les Vertesi. How many more cases are needed? A report on hip and knee surgery waits in british columbia. Technical report, CSMG Technical Report, IRMACS, Simon Fraser University, 2006. Prepared for the British Columbia Ministry of Health.

[5] Ronald M. Andersen. Behavioral model of families' use of health services, 1968. Research Series No. 25. Chicago, IL: Center for Health Administration Studies, University of Chicago.

[6] Ronald M. Andersen. Revisiting the behavioral model and access to medical care: Does it matter? *Journal of Health and Social Behavior*, 36(1):1–10, 1995.

[7] T. Andersson, , and P. Várbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, pages 1–7, 2006.

[8] Margarete Arndt and Barbara Bigelow. Commentary: The potential of chaos theory and complexity theory for health services management. *Health Care Management Review*, 25(1):35 – 38, 2000.

[9] Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963.

[10] C. Audet and J. E. Dennis, Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17(1):188–217 (electronic), 2006.

[11] Thomas Bäck. *Evolutionary algorithms in theory and practice*. The Clarendon Press Oxford University Press, New York, 1996. Evolution strategies, evolutionary programming, genetic algorithms.

[12] Rose D Baker. Sensitivity analysis for healthcare models fitted to data by statistical methods. *Health Care Management Science*, 5(4):275 – 281, 2002.

[13] Osman Balci. Verification, validation and accreditation of simulation models. In S. Andradttir, K. J. Healy, D. H. Withers, and B. L. Nelson, editors, *Proceedings of the 1997 Winter Simulation Conference*, pages 135 – 141, 1997.

[14] Osman Balci. Quality assessment, verification, and validation of modeling and simulation applications. In R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 122 – 129, 2004.

[15] A. Bandura. Health promotion by social cognitive means. *Health Education and Behavior*, 31:143–164, 2004.

[16] Yaneer Bar-Yam. Improving the effectiveness of health care and public health: A multiscale complex systems analysis. *American Journal of Public Health*, 96:459–466, 2006.

[17] Forest Baskett, K. Mani Chandy, Richard R. Muntz, and Fernando G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.*, 22:248–260, 1975.

[18] Gary S. Becker. Investment in human capital: A theoretical analysis. *The Journal of Political Economy*, 70:9–49, 1962.

[19] Gary S. Becker. *The Economic Approach to Human Behaviour*. University of Chicago Press, Chicago, 1976.

[20] Gary S. Becker. The economic way of looking at life, 1992. Nobel Lecture, December 9, 1992.

[21] Gary S. Becker. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. University of Chicago Press, Chicago, 1993.

[22] Gary S. Becker and Kevin M. Murphy. A theory of rational addiction. *The Journal of Political Economy*, 96:675–700, 1988.

[23] Gary Stanley Becker. Human capital and personal distribution of income: an analytical approach. In *Woytinsky lecture no. 1*. Ann Arbor: Institute of Public Administration, 1967.

[24] C M Bell, M Crystal, A S Detsky, and D A Redelmeier. Shopping around for hospital services: a comparison of the united states and canada. *JAMA: The Journal Of The American Medical Association*, 279(13):1015 – 1017, 1998.

[25] S. Bergheim. *Live long and prosper! Health and longevity as growth drivers*. Deutsche Bank Research, Frankfurt, Germany, March 2006.

[26] P. P. Biemer and L. E. Lyberg. *Introduction to Survey Quality*. John Wiley and Sons, Inc.,, New Jersey, USA, 2003.

[27] L. Billard. Markov models and social analysis. *International Encyclopedia of the Social and Behavioural Sciences*, pages 9242–9250, 2004.

[28] John T. Blake and Michael W. Carter. An analysis of emergency room wait time issues via computer simulation. *INFOR*, 34(4):263 – 273, 1996.

[29] Kristian Bolin, Lena Jacobson, and Björn Lindgren. The family as the health producer — when spouses are Nash-bargainers. *Journal of Health Economics*, 20:349–362, 2001.

[30] Kristian Bolin, Lena Jacobson, and Björn Lindgren. The family as the health producer — when spouses act strategically. *Journal of Health Economics*, 21:475–495, 2002.

[31] J. Frédéric Bonnans, J. Charles Gilbert, Claude Lemaréchal, and Claudia A. Sagastizábal. *Numerical optimization*. Universitext. Springer-Verlag, Berlin, second edition, 2006. Theoretical and practical aspects.

[32] K. Boulding. *Economic Analysis*. Harper and Row, New York, 1966.

[33] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. http://www.stanford.edu/ boyd/cvxbook/.

[34] S. C. Brailsford, V. A. Lattimer, P. Tarnaras, and J. C. Turnbull. Emergency and on-demand health care: modelling a large complex system. *Journal of the Operational Research Society*, 55(1):34 – 42, 2004.

[35] H. Bronnum-Hansen. How good is the Prevent model for estimating the health benefits of prevention? *J. Epidemiol. Community Health*, 53:300–305, 1999.

[36] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[37] B. Burström and P. Fredlund. Self rated health: Is it as good a predictor of subsequent mortality in lower as well as in higher social classes? *J. Epidemiol. Community Health*, 55:836–840, 2001.

[38] Barbara M Byrne. *Structural Equation Modeling with AMOS*. CRC Press, second edition, 2009.

[39] Hadorn David C. Setting priorities on waiting lists: point-count systems as linear models. *Journal of Health Services Research and Policy*, 8:48–54, 2003.

[40] A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*. Cambridge University Press, Cambridge, UK, 1998.

[41] A. C. Cameron, P. K. Trivedi, F. Milne, and J. Piggott. A microeconometric model of the demand for health care and health insurance in Australia. *The Review of Economic Studies*, 55:85–106, 1988.

[42] John S. Carson. Introduction to modeling and simulation. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 9–16, 2004.

[43] Robert Y. Cavana, Philip K. Davies, Rachel M. Robson, and Kenneth J. Wilson. Drivers of quality in health services: different worldviews of clinicians and policy managers revealed. *System Dynamics Review (Wiley)*, 15(3):331 – 340, 1999.

[44] Harry Clarke and Svetlana Danilkina. Talking rationally about rational addiction. preprint, Dept. of Economics, La Trobe University, 2006.

[45] Jeffery K Cochran and Aseem Bharti. A multi-stage stochastic methodology for whole hospital bed planning under peak loading. *International Journal of Industrial and Systems Engineering*, 1:8–36, 2006.

[46] J. E. Cohen and B. Singer. Malaria in Nigeria: Contrained continuous-time Markov models for discrete-time longitudinal data on human mixed-species infections. In *Some Mathematical Questions in Biology*, pages 69–133. Providence: American Mathematical Society, 1979.

[47] A D Colle and M Grossman. Determinants of pediatric care utilization. *The Journal Of Human Resources*, 13 Suppl:115 – 158, 1978.

[48] Collings, T. and Stoneman, C. The $m/m/\infty$ queue with varying arrival and departure rates. *Operations Research*, 24(4):760–773, 1976.

[49] A. R. Conn, K. Scheinberg, and Ph. L. Toint. Recent progress in unconstrained nonlinear optimization without derivatives. *Math. Programming*, 79(1-3, Ser. B):397–414, 1997. Lectures on mathematical programming (ismp97) (Lausanne, 1997).

[50] R Cook and J Rasmussen. "going solid": a model of system dynamics and consequences for patient safety. *Quality & Safety In Health Care*, 14(2):130 – 134, 2005.

[51] Ben Cooper and Marc Lipsitch. The analysis of hospital infection data using hidden markov models. *Biostatistics (Oxford, England)*, 5(2):223 – 237, 2004.

[52] Kenneth Craik. *The Nature of Explanation*. Cambridge University Press, Cambridge, 1943.

[53] Thomas F. Crossley and Steven Kennedy. The reliability of self-assessed health status. *Journal of Health Economics*, 21:643–658, 2002.

[54] John G. Cullis and Philip R. Jones. National health service waiting lists: A discussion of competing explanations and a policy proposal. *Journal of Health Economics*, 4(2):119 – 135, 1985.

[55] V. Dabbaghian, K. Vasarhelyi, N. Richardson, P. Borwein, and A. R. Rutherford. The impact of social interactions on the spread of HIV infection among injection drug users: A Cellular Automaton model. *CSMG Technical Report*, pages 1–18, 2008.

[56] B. Dangerfield and C. Roberts. Fitting a model of the spread of aids to data from five european countries. In *O.R. Work in HIV/AIDS*, pages 7–13, Birmingham, 1994.

[57] B. Dangerfield and C. Roberts. Foreword to the special issue on health and health care dynamics. *System Dynamics Review (Wiley)*, 15(3):197 – 199, 1999.

[58] B. Dangerfield and C. Roberts. Optimisation as a statistical estimation tool: an example in estimating the aids treatment-free incubation period distribution. *System Dynamics Review (Wiley)*, 15(3):273 – 291, 1999.

[59] B. C. Dangerfield. System dynamics applications to european health care issues. *The Journal of the Operational Research Society*, 50(4):345–353, 1999.

[60] M. S. Daskin. A maximum expected coverage location model: Formulation, properties and heuristic solution. *Transportation Science*, 17:48–70, 1983.

[61] M. S. Daskin and L. K. Dean. Location of health care facilities. In F. Sainfort M. L. Brandeau and W. P. Pierskalla, editors, *Operations research and health care: A handbook of methods and applications*, International series in operations research and management science, pages 43–76. Kluwer, Boston, 2004.

[62] Paul Davidsson. Agent based social simulation: A computer science view. *Journal of Artificial Societies and Social Simulation*, 5(1), 2002.

[63] G. A. Diamond, A. Rozanski, and M. Steuer. Playing doctor: application of game theory to medical decision-making. *Journal of Chronic Diseases*, 39:669–677, 1986.

[64] P Diehr, D Yanez, A Ash, M Hornbrook, and D Y Lin. Methods for analyzing health care utilization and costs. *Annual Review Of Public Health*, 20:125 – 144, 1999.

[65] Ana V. Diez-Roux. Multilevel analysis in public health research. *Annu. Rev. Public Health*, 21:171–192, 2000.

[66] Eddy K. A. Van Doorslaer. *Health, Knowledge and the Demand for Medical Care: an econometric analysis*. Van Gorcum, Assen/Maastricht, 1987.

[67] Marco Dorigo and Christian Blum. Ant colony optimization theory: a survey. *Theoret. Comput. Sci.*, 344(2-3):243–278, 2005.

[68] Steven B. Dowd. Applied game theory for the hospital manager: Three case studies. *The Health Care Manager*, 23:156–161, 2004.

[69] F. Y. Edgeworth. *Mathematical Psychics*. Kegan Paul, London, 1881.

[70] Matthew J. Eichner. The demand for medical care: What people pay does matter. *American Economic Review*, 88(2):117 – 121, 1998.

[71] J. E. Everrett. A decision support simulation model for the management of an elective surgery waiting system. *Health Care Management Science*, 5(2):89 – 95, 2002.
[72] Michael G. Farnworth. A game theoretic model of the relationship between prices and waiting times. *Journal of Health Economics*, 22:47–60, 2003.
[73] Centre for Disease Control and Prevention. HIV/AIDS Surveillance Report, 2007. Technical report, Department of Health and Human Services, Centres for Disease Control and Prevention, Atlanta, USA, 2009.
[74] Mario Forni and Marco Lippi. Aggregation of linear dynamic microeconomic models. *Journal of Mathematical Economics*, 31(1):131 – 158, 1999.
[75] Jay W. Forrester. *Industrial dynamics*. Pegasus Communications, Waltham, MA, 1961.
[76] Jay W. Forrester. System dynamics and the lessons of 35 years. In Kenyon B. de Greene, editor, *The Systemic Basis of Policy Making in the 1990s*. MIT Press, 1991.
[77] Jay W. Forrester. System dynamics, systems thinking, and soft or. *System Dynamics Review (Wiley)*, 10(2/3):245 – 256, 1994.
[78] Jay W. Forrester. The beginning of system dynamics. *McKinsey Quarterly*, 4:4 – 16, 1995.
[79] The Framingham heart study, 2002. `http://www.nhlbi.nih.gov/about/framingham`.
[80] J. F. Fries, C. E. Koop, J. Sokolov, C. E. Beadle, and D. Wright. Beyond health promotion: reducing need and demand for medical care. *Health Affairs*, 17:70–84, 1998.
[81] World Cancer Research Fun. Food, nutrition, physical activity and the prevention of cancer: A global persective. Technical report, World Cancer Research Fund, 2007.
[82] Mauro Gallegato and Alan Kirman, editors. *Beyond the representative agent*. Edward Elgar, Cheltenham, UK, 1999.
[83] Mauro Gallegato, Alan P. Kirman, and Matteo Marsili, editors. *The Complex Dynamics of Economic Interaction: Essays in Economics and Econophsyics*, volume 531 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, 2004.
[84] M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27:1641–1653, 2001.
[85] Jon Gjerde, Sverre Grepperud, and Snorre Kverndokk. On adaption and the demand for health. *Applied Economics*, 37:1283–1301, 2005.
[86] F. Glover and M. Laguna. *Tabu Search*. Springer, New York, 1997.
[87] J. B. Goldberg. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, 1(1):20–39, 2004.
[88] A. Goldbeter. A model for the dynamics of human weight cycling. *J. Biosci.*, 31:129–136, 2006.
[89] B. González-Busto and R. García. Waiting lists in spanish public hospitals: a system dynamics approach. *System Dynamics Review (Wiley)*, 15(3):201 – 224, 1999.
[90] Leo A. Goodman. Statistical methods for the Mover-Stayer model. *Journal of the American Statistical Association*, 56:841–868, 1961.
[91] Florin Gorunescu, Sally I. McClean, and Peter H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307 – 312, 2002.
[92] L. V. Green and S. Savin. Providing timely access to medical care: a queueing model. *Operations Research*, preprint:1–42, 2008.
[93] Michael Grossman. On the concept of health capital and the demand for health. *The Journal of Political Economy*, 80:223–255, 1972.
[94] Martin Grötschel, Sven O. Krumke, and Jörg Rambau, editors. *Online optimization of large scale systems*. Springer-Verlag, Berlin, 2001.
[95] Louise Gunning-Schepers. *The Health Benefits of Prevention*. Elsevier, Amsterdam, 1989.
[96] Jahn Karl Hakes and W. Kip Viscusi. Dead reckoning: Demographic determinants of the accuracy of mortality risk perceptions. *Risk Analysis*, 24:651–664, 2004.
[97] Randolph W. Hall. Patient flow. *OR/MS Today*, 33(3), 2006.
[98] Marianne Hanning. Maximum waiting-time guarantee – an attempt to reduce waiting lists in sweden. *Health Policy*, 36(1):17–35, 1996.
[99] Marianne Hanning and Ulrika Winblad Spangberg. Maximum waiting time – a threat to clinical freedom?: Implementation of a policy to reduce waiting times. *Health Policy*, 52(1):15–32, 2000.
[100] L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[101] H. Hare, W. & Dodd. Non-publicly funded accommodation environments in British Columbia: survey and analysis. technical report, 2008.

[102] W. L. Hare and G. Tanoh. Recovery rates for knee and hip surgery patients, 2007. CSMG Technical Report prepared for the Fraser Health Authority.

[103] Robert Haveman, Barbara Wolfe, Brent Kreider, and Mark Stone. Market work, wages, and men's health. *J. Health Econ.*, 13:163–182, 1994.

[104] F.J. He and G.A. MacGregor. A comprehensive review on salt and health and current experience of worldwide salt reduction programmes. *Journal of Human Hypertension*, 23:363–384, 2009.

[105] J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320, 1984.

[106] S. G. Henderson and A. J. Mason. Ambulance service planning" simulation and data visualisation. In M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, editors, *Operations research and health care: A handbook of methods and applications*, International series in operations research and management science, pages 78–102. Kluwer, Boston, 2004.

[107] Gary Hirsch and C. Sherry Immediato. Microworlds and generic structures as resources for integrating care and improving health. *System Dynamics Review (Wiley)*, 15(3):315 – 330, 1999.

[108] Gary B. Hirsch. System dynamics modeling in health care. *SIGSIM Simul. Dig.*, 10(4):38–42, 1979.

[109] Marna Hoard, Jack Homer, William Manley, Paul Furbee, Arshadul Haque, and James Helmkamp. Systems modeling in support of evidence-based disaster planning for rural areas. *International Journal of Hygiene and Environmental Health*, 208(1-2):117–125, 2005.

[110] Michael Hoel and Erik Magnus Saether. Public health care with waiting time: the role of supplementary private health care. *Journal of Health Economics*, 22(4):599–616, 2003.

[111] Jack Homer and Bobby Milstein. Optimal decision making in a dynamic model of community health. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, 2004.

[112] Jack B. Homer and Gary B. Hirsch. System dynamics modeling for public health: Background and opportunities. *American Journal of Public Health*, 96:452–458, 2006.

[113] Amanda A. Honeycutt, James P. Boyle, Kristine R. Broglio, Theodore J. Thompson, Thomas J. Hoerger, Linda S. Geiss, and K. M. Venkat Narayan. A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050. *Health Care Management Science*, 6:155–164, 2003.

[114] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin. Systems biology and new technologies enable predictive and preventive medicine. *Science*, 306:640–643, 2004.

[115] David W. Hosmer, Scott Taber, and Stanley Lemeshow. The importance of assessing the fit of logistic regression models: A case study. *American Journal of Public Health*, 81(12):1630, 1991.

[116] Darrel Huff. *How to Lie with Statistics*. W. W. Norton & Company, New York, 1954.

[117] J. S. Ivy and L. Maillart. Mathematical modeling of dynamic breast cancer screening policies. In *INFORMS Optimization Society Conference: Optimization and Health Care*, San Antonio, Texas, February 3–5 2006.

[118] T. E. Novotny et al. J. P. Pierce, M. C. Fiore. Trends in cigarette smoking in the united states, projections to the year 2000. *Journal of the American Medical Association*, 261:61 – 65, 1989.

[119] Nancy K. Janz and Marshall H. Becker. The health belief model: A decade later. *Health Education Quarterly*, 11:1–47, 1984.

[120] M.R. Joffres and A. Alimadad. Estimates of reductions in events from ischemic heart diseases, cerebrovascular diseases and heart failure in canada following a 5 or 10% yearly reduction in sodium intake at the population level, and potential savings in hospital costs. Technical report, Faculty of Health Sciences, Simon Fraser University, 2009. report for public health agency of Canada - Sodium subgroup research committee.

[121] Andrew Jones. *Applied Econometrics for Health Economists*. Radcliffe Publishing Ltd, United Kingdom, 2007.

[122] Andrew M. Jones. Health econometrics. In K. J Culyer and J. P. Newhouse, editors, *North-Holland Handbook of Health Economics*. North Holland, 2000.

[123] Jay Ashvin Joseph. The applicability, usefulness, and limitations of the PREVENT model, as demonstrated by modeling the effects of alcohol consumption interventions on coronary heart disease mortality. Master's thesis, Department of Community Health, University of Toronto, 1997.

[124] Donald S. Kenkel. Health behavior, health knowledge, and schooling. *Journal of Political Economy*, 99(2):287–305, 1991.

[125] D. M. Kennedy, S. E. Hanna, P. W. Stratford, J. Wessel, and J. D. Gollish. Preoperative function and gender predict pattern of functional recovery after hip and knee arthroplasty. *The Journal of Arthroplasty*, 21(4):559–566, 2006.

[126] D. M. Kennedy, P. W. Stratford, S. E. Hanna, J. Wessel, and J. D. Gollish. Modeling early recovery of physical function following hip and knee arthroplasty. *BMC Musculoskeletal Disorders*, 7:100, 2006.

[127] S. Kirkpatrick, C. D. Gelatt, and Jr. M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[128] Alan Kirman. Whom or what does the representative individual represent? *The Journal of Economic Perspectives*, 6:117–136, 1992.

[129] Alan Kirman and Jean Benoît Zimmermann, editors. *Economics with Heterogeneous Interacting Agents*, volume 503 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, 2001.

[130] Marie Fanelli Kuczmarski, Robert J. Kuczmarski, and Matthew Najjar. Effects of age on validity of self-reported height, weight, and body mass index: Findings from the third national health and nutrition examination survey, 1988-1994. *Journal of the American Dietetic Association*, 101(1):28–34, 2001.

[131] Snorre Kverndokk. Why do people demand health? Technical report, Ragnar Frisch Centre for Economic Research, University of Oslo, 2000.

[132] Kajal Lahiri and Guibo Xing. An econometric analysis of veterans' health care utilization using two-part models. *Empirical Economics*, 29(2):431 – 449, 2004.

[133] D. C. Lane, C. Monefeldt, and J. V. Rosenhead. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, 51(5):518–531, 2000.

[134] David Lane, Camilla Monefeldt, and Jonathan Rosenhead. Emergency - but no accident — a system dynamics study of an accident and emergency department. *OR Insight*, 1998.

[135] Jean K. Langlie. Social networks, health beliefs, and preventive health behavior. *Journal of Health and Social Behavior*, 18(3):244–260, 1977.

[136] V Lattimer, S Brailsford, J Turnbull, P Tarnaras, H Smith, S George, K Gerard, and S Maslin-Prothero. Reviewing emergency care systems I: insights from system dynamics modelling. *Emerg Med J*, 21(6):685–691, 2004.

[137] R. M. Lewis and V. Torczon. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM J. Optim.*, 12(4):1075–1089, 2002.

[138] Stephen Lewis and Claudia Sanmartin. Managing waiting lists to achieve distributive justice. A working paper prepared for the Western Canada Wait List Project.

[139] Steven Lewis, Morris L. Barer, Claudia Sanmartin, Sam Sheps, Samuel E.D. Shortt, and Paul W. McDonald. Ending waiting-list mismanagement: principles and practice. *CMAJ: Canadian Medical Association Journal*, 162(9):1297 –, 2000.

[140] W. G. Liddell and J. H. Powell. Agreeing access policy in a general medical practice: a case study using qpid. *System Dynamics Review (Wiley)*, 20(1):49 – 73, 2004.

[141] Cotton M. Lindsay and Bernard Feigenbaum. Rationing by waiting lists. *American Economic Review*, 74(3):404 – 417, 1984.

[142] Liu, Chih-Ming, Wang, Kuo-Ming, and Guh, Yuh-Yuan. A markov chain model for medical record analysis. *The Journal of the Operational Research Society*, 42(5):357–364, 1991.

[143] Mary V. Look. *Policy Systems and their Complexity Dynamics: Academic Medical Centers and Managed Care Markets*. PhD thesis, Virginia Polytechnic Institute and State University, Virginia, 2003.

[144] A. Mark, D. Pencheon, and R. Elliott. Demanding healthcare. *Int. J. Health Plann. and Mgmt.*, 15:237–253, 2000.

[145] Brian J. Masterson, Thomas G. Mihara, George Miller, Stephen C. Randolph, M. Emma Forkner, and Andrew L. Crouter. Using models and data to support optimization of the

military health system: A case study in an intensive care unit. *Health Care Management Science*, 7(3):217 – 224, 2004.

[146] Roger McCain. *Game Theory: A Non-Technical Introduction to the Analysis of Strategy*. South-Western College Pub, 2003.

[147] McClean, S. I., McAlea, B., and Millard, P. H. Using a markov reward model to estimate spend-down costs for a geriatric department. *The Journal of the Operational Research Society*, 49(10):1021–1025, 1998.

[148] K. M. McGrail, B. Green, M. L. Barer, R. G. Evans, C. Hertzman, and C. Normand. Age, costs of acute and long-term care and proximity to death: evidence for 1987-88 and 1994-95 in British Columbia. *Age and Ageing*, 29:249–253, 2000.

[149] Beth E. Meyerowitz and Shelly Chaiken. The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of Personality and Social Psychology*, 52(3):500 – 510, 1987.

[150] L. A. Meyers, M. E. J. Newman, M. Martin, and S. Schrag. Applying network theory to epidemics: Control measures for *mycoplasma pneuminiae* outbreaks. *Emerging Infectious Diseases*, 9:204–210, 2003.

[151] Quentin Michard and Jean-Philippe Bouchaud. Theory of collective opinion shifts. *The European Physical Journal B — Condensed Matter and Complex Systems*, 47:151–159, 2005.

[152] Olli S. Miettinen. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology*, 99:325–332, 1974.

[153] Bobby Milstein and Jack Homer. Background on systems dynamics simulation modeling, with a summary of major public health studies. Syndemics Prevention Network, May 2006.

[154] Y. Mizuno, D. Wilkonson, S. Santibanez, C. Dawson Rose, A. Knowlton, K. Handley, M. N. Gourevitch, and INSPIRE Team. Correlates of health care utilization among hiv-seropositive injection drug users. *AIDS Care*, 18(5):417–425, 2006.

[155] John Morecroft and Stewart Robinson. Comparing discrete-event simulation and system dynamics: modelling a fishery. In *Proceedings of the 2006 Operational Research Society Simulation Workshop, 28-29 March, 2006,*, pages 137–148. UK Operational Research Society, 2006.

[156] Oskar Morgenstern and John von Neumann. *Theory of Games and Economic Behaviour*. Princeton University Press, 1944.

[157] N.R. Campbell M.R. Joffres and K. Tu B. Manns. Estimate of the benefits of a population-based reduction in dietary sodium additives on hypertension and its related health care costs in canada. *Can J Cardiol*, 6:437 – 443, 2007.

[158] P. Mullen. Waiting lists in the post review nhs. *Health Service Management*, 7(2):131 – 145, 1994.

[159] Mark Murray and Donald M. Berwick. Advanced Access: Reducing Waiting and Delays in Primary Care. *JAMA*, 289(8):1035–1040, 2003.

[160] John F. Nash. *Non-cooperative Games*. PhD thesis, Princeton University, 1950.

[161] Vicente Navarro, Rodger Parker, and Kerr L. White. A stochastic and deterministic model of medical care utilization. *Health Services Research*, 5(4):342 – 357, 1970.

[162] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[163] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):16128 – 16139, 2002.

[164] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167 – 256, 2003.

[165] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[166] Tom Noseworhty. Top priority. *Canadian Healthcare Manager*, 11(6):43–44, 2004.

[167] Ted O'Donoghue and Matthew Rabin. Risky behavior among youths: Some issues from behavioral economics. In Jon Gruber, editor, *Risky Behavior Among Youths*, pages 29–67. University of Chicago Press, Chicago, 2001.

[168] Public Health Agency of Canada. Hiv/aids epi updates, november 2007. Technical report, Surveillance and Risk Assessment Division, Centre for Infectious Disease Prevention and Control, Public Health Agency of Canada, 2007.

[169] Rogelio Oliva. Model structure analysis through graph theory: partition heuristics and feedback structure decomposition. *System Dynamics Review (Wiley)*, 20(4):313 – 336, 2004.

[170] Panel on Dietary Reference Intakes for Electrolytes and Standing Committee on the Scientific Evaluation of Dietary Reference Intakes Water. *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate*. National Academies Press, 2004.

[171] World Health Organization. Fact sheet number 297: cancer. Technical report, WHO, 2006.

[172] World Health Organization. Reducing salt intake in populations. Technical report, World Health Organization, 2007. report of a WHO forum and technical meeting, 5-7 October 2006, Paris, France.

[173] A. Signhal P. W. Vaughan, E. M. Rogers and R. M. Swahele. Entertainment-education and hiv/aids prevention: a field experiment in tanzania. *Journal of Health Communication*, 5(Supplement):81–100, 2000.

[174] Mari Palta, Ronald J Prineas, Reuben Berman, and Peter Hannan. Comparison of Self-Reported and Measured Height and Weight. *Am. J. Epidemiol.*, 115(2):223–230, 1982.

[175] Bernice A. Pescosolido. Beyond rational choice: The social dynamics of how people seek help. *American Journal of Sociology*, 97:1096–1138, 1992.

[176] Carl V. Phillips. Quantifying and reporting uncertainty from systematic errors. *Epidemiology*, 14(4):459–466, 2003.

[177] Carl V. Phillips. Publication bias in situ. *BMC Medical Research Methodology*, 4, 2004.

[178] Carl V. Phillips and Richard Zeckhauser. Communicating the health effects of consumer products: The case of moderate alcohol consumption and coronary heart disease. *Managerial and Decision Economics*, 17:459–470, 1996.

[179] K. A. Phillips, K. R. Morrison, R. Andersen, and L. A. Aday. Understanding the context of healthcare utilization: assessing environmental and provider-related variables in the behavioral model of utilization. *Health Services Research*, 33(3 Pt. 1):571 – 596, 1998.

[180] Paul E. Plsek and Trisha Greenhalgh. Complexity science: the challenge of complexity in health care. *British Medical Journal*, 323:625–628, 2001.

[181] B. Pourbohloul, L. A. Meyers, D. M. Skowronski, M. Krajden, D. M. Patrick, and R. C. Brunham. Modeling control strategies of respiratory pathogens. *Emerging Infectious Diseases*, 11:1249–1256, 2005.

[182] Nicolaas P. Pronk, Michael J. Goodman, Patrick J. O'Connor, and Brian C. Martinson. Relationship between modifiable health risks and short-term health care charges. *Journal of the American Medical Association*, 282:2235–2239, 1999.

[183] H. N. Psaraftis. Dynamic vehicle routing problems. In B. L. Golden and A. A. Assad, editors, *Vehicle Routing: Methods and Studies*, volume 16 of *Studies in Management Science and Systems*, pages 223–248. North-Holland, Amsterdam, 1988.

[184] B. Ramadanović, A. van der Waall, A.R. Rutherford, L. Vertesi, Y. Wang, and I. Rongve. Performance Metrics and Service Discipline in a System-Scale Model of Surgical Wait List. In *Proceedings of the 35th International Conference on Operational Research Applied to Health Services (ORAHS)*, Leuven, Belgium, July 2009.

[185] B. Rockhill, B. Newman, and C. Weinberg. Use and misuse of population attributable fractions. *American Journal of Public Health*, 88:15–19, 1998.

[186] Thomas Rohleder, Diane Bischak, and Leland Baskin. Modeling patient service centers with simulation and system dynamics. *Health Care Management Science*, 10(1):1 – 12, 2007.

[187] C. Roos, T. Terlaky, and J.-P. Vial. *Interior point methods for linear optimization*. Springer, New York, 2006. Second edition of *Theory and algorithms for linear optimization* [Wiley, Chichester, 1997; MR1450094].

[188] Carla Rossi. Operational models for epidemics of problematic drug use: the Mover-Stayer approach to heterogeneity. *Socio-Economic Planing Sciences*, 38:73–90, 2004.

[189] ML Rowland. Self-reported weight and height. *Am J Clin Nutr*, 52(6):1125–1133, 1990.

[190] Geoff Royston, Ayesha Dost, Jeremy Townshend, and Howard Turner. Using system dynamics to help develop and implement policies and programmes in health care in England. *System Dynamics Review*, 15(3):293–313, 1999.

[191] Claudia Sanmartin. Acceptable waiting times for medical services: A review pf the evidence and proposed methods. A Working Paper prepared for the Western Canada Wait List Project.

[192] Claudia Sanmartin, Samuel E.D. Shortt, Morris L. Barer, Sam Sheps, Steven Lewis, and Paul W. McDonald. Waiting for medical services in canada: lots of heat, but little light. *CMAJ: Canadian Medical Association Journal*, 162(9):1305 – 1310, 2000.

[193] Sisira Sarma and Wayne Simpson. A microeconometric analysis of Canadian health care utilization. *Health Economics*, 15:219–239, 2006.

[194] M. W. P. Savelsbergh and M. Sol. The general pickup and delivery problem. *Transportation Science*, 29:17–29, 1995.

[195] Alexander Schrijver. *Theory of linear and integer programming*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1986. A Wiley-Interscience Publication.

[196] James P. Sethna, Karin A. Dahmen, and Christopher R. Myers. Crackling noise. *Nature*, 410:241–250, 2001.

[197] James P. Sethna, Karin A. Dahmen, and Olga Perković. Random field Ising models of hysteresis. arXiv:cond-mat/0406320, 2005.

[198] S. Shechter. The optimal time to initiate HIV therapy. In *INFORMS Optimization Society Conference: Optimization and Health Care*, San Antonio, Texas, February 3–5 2006.

[199] Brian Skyrms and Robin Pemantle. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences*, 97:9340–9346, 2000.

[200] Kaiser staff. Health care spending in the united states and OECD countries. online, `http://www.kff.org/insurance/snapshot/index.cfm`, 2007.

[201] Inc. StatSoft. Electronic statistics textbook. Technical report, StatSft, inc., Tulsa, OK, 1997.

[202] J. A. Stein, S. A. Fox, and P. J. Murata. The influence of ethnicity, socioeconomic status, and psychological barriers on use of mammography. *Journal of Health and Social Behaviour*, 32:101–113, 1991.

[203] Ken Stein, J. Dalziel, Andrew Walker, B. Jenkins, Alison Round, and Pam Royle. Screening for hepatitis C in genito-urinary medicine clinics: a cost utility analysis. *Journal of Hepatology*, 39(5):814–825, 2003.

[204] Andrew Steptoe and Jane Wardle. Health behaviour, risk awareness and emotional well-being in students from eastern europe and western europe. *Social Science & Medicine*, 53(12):1621–1630, 2001.

[205] John D. Sterman. Learning from evidence in a complex world. *American Journal of Public Health*, 96:505–514, 2006.

[206] Andrew Street and Stephen Duckett. Are waiting lists inevitable? *Health Policy*, 36(1):1–15, 1996.

[207] A. H. Studenmund and Henry J. Cassidy. *Using Econometrics: A practical guide*. Little, Brown and Company, 1987.

[208] L. Tabár, B. Vitak, H-H. T. Chen, M-F. Yen, S. W. Duffy, and R. A. Smith. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer*, 91(9):1724–1731, 2001.

[209] Antuela Tako and Stewart Robinson. Towards an empirical comparison of discrete-event simulation and system dynamics in the supply chain context. In S. Robinson J. Garnett, S. Brailsford and S. Taylor, editors, *Proceedings of the 2006 Operational Research Society Simulation Workshop, 28-29 March, 2006*. UK Operational Research Society, 2006.

[210] Y. W. Tan. First passage probability distributions in Markov models and the HIV incubation distribution under treatment. *Mathematical and Computer Modelling*, 19(11):53–66, 1994.

[211] C. Tarrant, T. Stokes, and A. M. Colman. Models of the medical consultation: opportunities and limitations of a game theory perspective. *Quality and Safety in Health Care*, 13:461 – 466, 2004.

[212] Gavin Turrell, Brian Oldenburg, Ingrid McGuffog, and Rebekah Dent. Socioeconomic determinants of health: towards a national research program and a policy and intervention agenda, 1999. School of Public Health, Queensland University of Technology.

[213] UNAIDS. Report on the global HIV/AIDS epidemic 2008: executive summary. Technical report, Joint United Nations Program on HIV/AIDS, 2008.

[214] National Research Council (U.S.). *Network Science*. National Academy Press, Washington, 2005.

[215] M. Utley, S. Gallivan, T. Treasure, and O. Valencia. Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. *Health Care Management Science*, 6:97–104, 2003.

[216] S. T. Vadaparampil, V. L. Champion, T. K. Miller, U. Menon, and C. S. Skinner. Using the health belief model to examine differences in adherence to mammography among african-american and caucasian women. *Journal of Psychosocial Oncology*, 21(4):59–79, 2005.

[217] Ann Van Ackere and Peter C. Smith. Towards a macro model of national health service waiting lists. *System Dynamics Review (Wiley)*, 15(3):225 – 252, 1999.

[218] A. van der Waall, A. Bakhtiari, B. Ramadanović, A. Rutherford, Y. Wang, and L. Vertesi. Analysis and modelling of hip, knee and cataract surgeries in british columbia during 2001–2008 on a health authority level. Report, The Complex Systems Modelling Group, IRMACS, SFU, April, 2009. Prepared for the British Columbia Ministry of Health Services.

[219] F. Vanden Berghen and H. Bersini. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: experimental results and comparison with the DFO algorithm. *J. Comput. Appl. Math.*, 181(1):157–175, 2005.

[220] L. Vandenberghe and S. Boyd. Applications of semidefinite programming. In *Proceedings of the Stieltjes Workshop on High Performance Optimization Techniques (HPOPT '96) (Delft)*, volume 29, pages 283–299, 1999.

[221] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, 1996.

[222] Jan M. H. Vissers. Health care management modelling: a process perspective. *Health Care Management Science*, 1(2):77 – 85, 1998.

[223] Adam Wagstaff. Econometric studies in health economics : A survey of the british literature. *Journal of Health Economics*, 8:1–51, 1989.

[224] A. F. Widmer. Replace hand washing with use of a waterless alcohol hand rub? *Clinical Infectious Diseases*, 31:136–143, 2000.

[225] Rainer Winkelmann. *Econometric analysis of count data*. Springer, 2003.

[226] E. F. Wolstenholme. Towards the definition and use of a core set of archetypal structures in system dynamics. *System Dynamics Review (Wiley)*, 19(1):7 – 26, 2003.

[227] Eric Wolstenholme. A patient flow perspective of u.k. health services: exploring the case for new "intermediate care" initiatives. *System Dynamics Review (Wiley)*, 15(3):253 – 271, 1999.

[228] Eric Wolstenholme. Using generic system archetypes to support thinking and modelling. *System Dynamics Review (Wiley)*, 20(4):341 – 356, 2004.

[229] M. H. Wright. Direct search methods: once scorned, now respectable. In *Numerical analysis 1995 (Dundee, 1995)*, volume 344 of *Pitman Res. Notes Math. Ser.*, pages 191–208. Longman, Harlow, 1996.

[230] Z. Yang, E. C. Norton, and S. C. Stearns. Longevity and health care expenditures: The real reasons older people spend more. *Journal of Gerontology*, 58B(1):S2–S10, 2003.

Index