Contents

Preface

Modelling is good, modelling healthcare is better. Read me and I'll tell you why.

Acknowledgments

It has been a great pleasure writing this book, and the authors would like to thank the many people involved in this project.

To begin, we would like to thank the complex systems modelling group (CSMG) at the center for interdisciplinary research in mathematics and computer science (IRMACS), for their continual contributions to the research and modelling techniques explored in this book. A list of CSMG members involved in this project can be found on page XXX, and a consistently updated list of CSMG members can be found on the CSMG web-pages.

The creation of this book is largely the result of a contract from the British Columbia Ministry of Health to examine options for modelling healthcare demand. The author would like to thank the BC Ministry of Health for funding this original project, and their continued support during the completion of this book.

The original report consisted of a comprehensive survey of a wide variety of techniques used to model and explain the various dimensions of the demand for healthcare. In most cases the models examined in the original report focused on the question of when and why individuals access healthcare. Since the completion of the original report, the CSMG has continued to study, research, and develop complex models of use to the healthcare community. Our current and future research has expanded greatly from our previous focus on "demand" to encompass many more topics of interest in healthcare. For example, epidemiological models of the determinants of obesity, queuing models of surgery waitlists, decision analysis models for disease diagnosis, and social network models of HIV/AIDS contagion. We would like to thank all of the funding sources for these projects as they have allowed us to extend our knowledge of modelling, and to bring together a strong group of researchers dedicated to the exploration of modelling both inside and outside the healthcare system.

In addition to this, the authors would like to thank... XXX

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)*
> A good model can advance fashion by ten years. Yves Saint-Laurent (1936-)

1.1 Why model in healthcare

XXX START WITH A HISTORY OF MODELLING AND WHY ITS SO COOL, 1 page

Once apon a time London was dying of cholera. Everybody thought it was the miasma which spread the disease. But John Snow had a different theory. John Snow made a map of London and noticed that the closer you were to the Broad Street Pump, the more likely you were to get sick. He had the pump shut down, and the city was saved. This is a first example of epidemiological modelling, and (in a way) network modeling in healthcare.

1.2 What is a Model

In todays world of pop culture the word model conjures images of beautiful people sporting the latest fashions on the front of magazines. This book is not focused on how these people can help in healthcare by posing for charity calendars. 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 always be those which arise in the healthcare industry.

There are two important aspects in the definition of a model. First, models are designed to answer specific questions. Models tend to answer the questions they are designed to answer, and as such, designing a model with no particular question in mind often results in a model which provide no insight into the situation it models. This may a useful exercise for a young academic student, but for healthcare policy-maker the result is generally just a wasting of resources.

The second important aspect of a model is that it is a *simplified* representation of the real situation. Consider the scientific endeavor of modelling a collection of building designs in order to determine which 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 could be preformed.

When simplifying the real situation for the proposes 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 serves. 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 to 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 which are applicable to solving questions in the field of healthcare, and there is no such thing as a unique "right" model for a given problem. In fact, in most cases, more than one model discussed in this book is 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.

1.3 When to Use Modelling in Healthcare

Before we discuss the various aspects of healthcare which have been impacted by modelling, expectations of what models can deliver must be tempered. 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 them. 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 blinded used, but validated both mathematically and by the solicitation of experts from the field they are modelling. A model that is in contradiction of the real situation should be held in doubt and its conclusions examined carefully.

Despite these limitations, modelling techniques stand posed to make a significant impact in the field of healthcare. XXX talk about budget explosions and how modelling might help.

Some common problems which arise in healthcare and are approached by models discussed in this book include:

- XXX list to be made later
- PROBLEM, type of models applicable (with chapter refs)

Modelling can be an invaluable tool to aid health care management, as long as it is used appropriately with awareness of its limitations. It is most useful to think of modelling in health care 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 health care experts to arrive together at appropriate approaches to problems in health care.

Chapter 2

How to Use this Book

author (-)

This book can be viewed and used in a number of different manners. Primarily it is an encyclopedia of modelling techniques with an emphasis on how to apply them to current issues in healthcare, however it could also be used as an introductory (undergraduate) text on the subject. An excellent (undergraduate) thesis project would be to select a chapter from this book, read it and its references, and then preform a literature search for new examples of the model's application in healthcare.

As an encyclopedia of modelling techniques, each chapter has been written to be entirely self-contained. That is, any given chapter can be read without previous knowledge from any other chapter. There are three chapters which focus on non-modelling issues, while the remaining chapters each focus on one specific type of model. The three non-modelling chapters are:

- Chapter 1, which discusses modelling in healthcare as a whole,
- Chapter **??**, which discusses the issues around collecting data for analysis, and
- Chapter 9, which discusses some general issues about constructing and analyzing models.

Each of the remaining chapters discusses one particular model that can be applied in the field of healthcare. Each of these chapters is given a artistic title which provides some insight as to where the modelled discussed might be used, and a scientific title which provides the standard naming for the model the chapter examines.

In order to ease reading, the layout for each modelling chapter is the same. Each modelling chapter is divided into five sections, entitled: "Model Overview," "Common Uses," "Mathematical Details," "Examples," and "Related Reading." In the Model Overview section we give a brief (usually one page) description of the model. These Model Overview sections are written using no mathematical language and should be readable by anyone with a high-school background in science. The next section, Common Uses, provides a list of questions which the model is often used to approach. These lists are not complete, but hopefully provide a strong idea of what kind of problems the model is capable of answering. In the Mathematical Details sections we give a more detailed description and analysis of the model. When possible we provide all the necessary scientific background to read these chapters (again the reader is assume to have completed a high-school education), however in some cases the models are

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

complicated enough that this is impossible. Sections which require more than a high-school level background are:

- Section 15.3, Markov and Supply Chain Models, which requires an understanding of matrix multiplication,
- XXX to be filled in later

In each of the Examples sections we provide two or three examples of how the model 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 which 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 model in healthcare, and others demonstrate how the model can fail if it is used inappropriately. The final section of each modelling chapter, Related Reading, provides details for each reference used in the chapter, as well as some references which provide more detailed reading on the model discussed.

Regardless of the model discussed, the "Model Overview" section is written assume the reader has only a high-school education.

This book also contains three appendices which should be of use to the reader. The first is a comprehensive glossary of terms used in the book. The second is a collection of data sources we have found useful in our experience modelling the healthcare system. Most of these data sources are freely available, perhaps requiring a brief e-mail contact to initiate the data exchange. The final glossary is a list, and review, of some of the modelling software that we have come across during our research. This glossary can be quite technical at times, and is intended for those who are interested in using software to producing models.

2.1 The Language of Modellers

The world of modelling and researchers whom work in it have existed for a long time. Over this time a certain vocabulary has been built around the subject which may not be common outside of it. As such, this book includes a glossary of terms (Appendix A) which should be of use to many readers. We highly recommend using it whenever a "foreign" word presents itself. For now we would like to highlight some words which are frequently used throughout this book and comment on their meaning with regards to modelling in healthcare.

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

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

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

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

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

- XXX more added as necessary

Chapter 3

The Modelling Process

> 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)*
> QUOTE 2 *name (date-date)*

From computing the optimal staff schedule for a hospital emergency room to exploring how global airline industry impacts the spread of disease, models are now impacting almost every area of the healthcare industry. Yet, to many healthcare policy makers, the development, tuning, testing, validation, and eventual application of a model is considered a foreign art. In this Chapter we provide a very broad stroke outline of the modelling process with specific emphasis on modelling in healthcare.

It is impossible to provide an efficient 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 outline some general steps that one should expect do during the modelling process. We begin with a quick overview of some guiding principles to modelling.

Guiding Principles to Modelling

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

Models should be simple and transparent: In building models, one of the most difficult tasks is to select the relevant details. Once this relevant details are uncovered, the model designed should be a simple as possible to incorporate these details.
On a related note, there are many software applications which 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 which are not. These assumptions must be clearly stated, and when possible tested. Assumptions which are not testable should be discussed with experts from the field.

Variables and measures should be clearly defined: A quantative model is useless if the result numeric 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. Although the model may be designed and tested with most data, final implementation and results should always use the best quality data available.

Interpret result 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 correlation between two events, not explaining the causality. (This is discussed more in Chapter 5, and other common errors are discussed as they arise within this book.)

In Figure 3.1 we provide our view of the modelling process. Notice it is a long process with many "feed-back" loops. This suggests that in modelling any given problem, many initial approaches to the problem will be unsuccessful. With practice and experience the number of unsuccessful approaches tried will decrease, but one should never expect the first attempt to modelling a system to work flawlessly.

XXX

Figure 3.1: Our 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.

3.1 Selecting a Modelling Approach

Almost every question (be it health care related or not) can be solving in more than one manner. Similarly, for most problems more than 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 what modelling technique to apply. The conclusion of this 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 answer desired. In general, the most convincing results are obtained when multiply modelling techniques are applied, and the results support each other.

Broadly speaking, modelling techniques fall into two categories: "Qualitative Models" and "Quantitative Models". We now discuss each of these in turn.

3.1.1 Qualitative Models

Many models in healthcare are not designed to provide specific numerical results pertaining to a given situation, but instead are designed to provide insight about why a given situation exists and what are its driving factors. Such models are generally referred to as *Qualitative Models*.

Qualitative models come in many forms, sometimes they rely on psychological analysis of a situation, while other models focus on examining how various aspects of a company interact. However, all qualitative models have the common factor that they do not attempt to produce a quantified output as a solution to a problem. Instead, they attempt to determine the factors which impact a given problem in order to provide guidance on how the situation can be adjusted.

Qualitative Models detailed in this book include

- XXX list to be made later
- MODEL NAME, chapter ref,

Consider for example the advertising industry and its continual goal to 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 there after. The reasons for these trends may appear clear (people are interested in buying gifts for their children 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 are people 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 which decrease the overall burden on the healthcare budget. By developing a qualitative model of the factors which effect an individual interactions with the healthcare system we can better understand why certain groups of people are less likely to maintain a regular schedule of mammography. In understanding this we can develop interventions which are better designed to solve the problem.

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 by the use of scientific experiments. That is, before applying the results implied by a qualitative models, one should always use quantitative modelling to confirm its validity.

3.1.2 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.

Quantitative Models detailed in this book include

- XXX list to be made later
- MODEL NAME, chapter ref,

It is instructive to consider a model of something as simple as a queue in a bank. This demonstrates some interesting characteristics of models and modelling. On the surface, a model of a bank line-up is very simple, but even at this level surprising subtleties are revealed in attempts to make a bank queue efficient or fair. In banks, usually a single queue of customers served by several bank tellers. The assumption is that the "first come first served" principle is fair and optimal in this case. In fact, whether the queue is fair depends on what is to be optimized. Are we interested in having tellers work most efficiently, in order to increase productivity, or should we optimise the time customers are waiting in line to maximize customer satisfaction? Perhaps some combination of the two is of interest. Is a single queue equally fair to all customers in line? If all bank tellers are suddenly occupied by customers with complicated requests that take a long time to resolve, the waiting time for those remaining in the queue should increase substantially compared to customers served before the block occurred. Ultimately we are forced to conclude that there is no such thing as a fair queue that is optimal in all circumstances and by every measure.

Clearly, even an apparently simple problem such as a bank queue masks a great deal of complexity. This complexity increases manyfold in common situations encountered daily in the health care system, such as a queue in an emergency department. The concerted operation of multiple health care 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 incorporates 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 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 linear extrapolation, and therefore limited in its accuracy. However, static models are often sufficient and generally easy to construct. In contrast, in*dynamic* models the state of variables changes with 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 variables is discrete or continuous. *Discrete* variables are variables which can only take values for 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 an 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 its 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 analysis 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 so small that continuity is essentially preserved.

3.2 Forming a Conceptual Model

Once a modelling approach is selected, the modeller next proceeds to 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 around[1].

When a conceptual model is formed it becomes a theoretical construct that represents, often visually, the processes, relationships, and variables considered to be important in a system. This construct should be examined by experts in the area to determine a first level of validity. After all, unless the experts believe the model it will remain unused regardless of its accuracy.

The conceptual model is both drives and is driven by which variables are considered important in the system. Since which variables are considered important may change as data analysis is preformed, 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.

3.3 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 in 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 in a surgical waitlist.

Data Collection

The classic *GIGO* axiom of modelling stands for “Garbage In, Garbage Out”. What GIGO captures is 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 accomplish 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 these data can be easily and accurately obtained.

However, 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 the confidentiality issue causes data collection to become troublesome. 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.

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

Data Processing (Cleaning)

Ideally, *data collection* is carried out with a specific modelling problem in mind. 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 another purpose. As a result, data may be biased and contain errors or inaccuracies. Another potential problem in health care modelling is that initially, the question may be too difficult to define. 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* involves checking for errors, identifying sources of bias, removing duplicates, and merging, linking, or inputing databases.

Statistical Analysis

Once data of adequate quality is available, the next step is to study the system through statistical analysis. *Statistical Analysis* may include the use of descriptive statistics (see Chapter 5), regression analysis (see Chapter 6), or risk analysis (see Chapters 7 and 8), or some combination thereof. (Many other forms of statistical analysis may also be employed, here we only list the forms detailed in this book.)

The results of the analysis of the data is used to determine which variables are most important for the problem, test a models validity, and 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 is significant or vice versa.

3.4 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. Implementation may involve a computer or may proceed using more analytical techniques.

Computer simulation is a software-based method of implementation. By simulating a system, it is possible to find a model solution without understanding how the system actually works. In this regard, simulation is often referred to as a *black box*; input and output are visible, but how the output is generated is understood. There are both advantages and disadvantage 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.

Alternately, one may chose to approach a model via the collection of tools provided by *mathematical analysis*. If the modeller is able to described 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 expensive software. Furthermore, analytic methods often provide deep insights into the workings of a system. However, analytic solutions for complex models are often difficult or even impossible to achieve, especially if the modeller is not mathematical by nature.

A third approach which lies somewhat between simulation and analysis is *numerical analysis*. In numerical analysis, the modeller uses mathematical technique to develop equations to represent the model and simplify these equation as far as possible. The modeller then turns to computers to numerically approximate solutions to the equations. These technique retains some of the robustness of mathematical analysis without requiring the modeller to have as strong of a mathematical background.

After a model is completed (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 not to use the same data set test the model as to tune the model. Doing so does not validate the model, but only shows the model works on this data set, because it was designed to work on this data set. In practice, there are many methods for validation of models. For example, in one review, a taxonomy of 77 verification and validation techniques has been compiled for conventional simulation models[2].

One must never use the same data set test the model as to tune the model. Doing this only shows the model works on the data set, because it was designed to work on the data set.

3.5 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 tested, tuned, and implemented a model is it can only examine the aspects of the system it is designed to study. A seating chart of an airplane is the perfect model to allocated seats in an airplane, but no matter how accurate 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 determine where to make interventions and where to focus further data collection efforts (parameters which make little impact on the system do not have to be quantified as accurately as parameters which have major impacts on the system). Analysis which focuses on determining which parameters have the greatest impact on the system is generally called *sensitivity analysis*.

Another popular use of a model is to determine some sort of optimal behaviour. For example, if a health care ministry has a budget to hire only a fixed number of physicians, they may wish to know the where to hire the physicians to achieve optimal patient care. In general *optimization* refers to applying this type of question to the model. Often the question can be written as “what selection of parameters minimize the cost such that the desired results occurs?” Finding the answers to such questions has become a field in itself, and can be accomplished by a number of different means. Chapters **??** and **??** of this book are devoted to some optimization problems in healthcare.

3.6 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.

3.7 Example: Modelling Healthcare Demand

Predicting the future demand for healthcare is of upmost importance to many healthcare policy-makers for the purpose of setting budgets and developing future coping strategies. In this example we use the question 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 worldwide. Data from the United States indicate that preventable illness constitute approximately 70% of the burden of illness 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[4]. This represents a huge challenge for finding ways to effectively promote lifestyle modification and prevent disease[3]. By reducing the number of people advancing from the healthy population to the at risk population, the overall demand for health care resources may be reduced.

Examples of population models in healthcare can be found in Examples XXX of this book.

Behavioural Models

Behavioural models address how people interact both 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 12. Behavioural models aim to understand the social dynamics that contribute to fluctuations in service utilisation. Thus, like population models, behavioural models can be used to ask how demand may be reduced before it is generated.

Patient-doctor interactions have been studied at the individual level using game theory[5][6][7] (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[8] (see Chapter 13 Example XXX). Further examples of behavioural models in healthcare can be found in Examples XXX 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 14) and discrete event simulation (Chapter **??**) have been largely the method used 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 10). Examples of operational models in healthcare can be found in Examples XXX of this book.

Global Models

Global models are complex system models that may incorporate any of the previous three types in order to study how multiple components of a healthcare system interact. 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 be positive or negative in a global sense.

Cost-containment is perhaps the toughest problem facing the healthcare system. Health care is absorbing a growing proportion of government budgets, but demographic explanations fail to 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 nonlinear interactions and feedback loops.

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 10)[9][10]. Examples of operational models in healthcare can be found in Examples XXX of this book.

3.8 Related Reading

For another description of the modelling process see [11].

1. C
2. B
3. P
4. F
5. diamond86
6. dowd04
7. tarrant04

8. pescosolido92
9. homer04
10. Homer06
11. carson04
12. cochran00
13. merzel03

Part I

Data Collection and Interpretation

Chapter 4

Issues of Data

> He used statistics 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 one is modelling, or what modelling technique one is applying, at some level every model should be grounded in reality. Sometimes this grounding in reality comes from consultation with experts in the field, or from logical deductions on how things work. More commonly, this grounding in reality 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 error which can arise in data collection. We begin with some terminology to describe different types of data.

4.1 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 regards it is useful to provide some terminology on these concepts.

4.1.1 Data Collection Methods

The question of how the data was acquired is of key importance in determining the quality of the data. In the field of healthcare one generally finds that there are three common methods for collecting data: *experiments*, *health records*, and *surveys*. Table **??** summarizes the differences of each of these types of data and highlights some advantages and disadvantages of each. Following this we discuss each data type is more detail.

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 health care 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

Table 4.1: Three common methods for collecting data, and the advantages and disadvantages of each.

Experimental data

The least common, but most reliable, source of data is data which is generated via scientific experiments. By scientific experiments we refer to experiments which 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, provided the *key factors* remain constant.

In physics or chemistry this may be a very achievable goal, however in the field of healthcare this reproducibility is extremely difficult to achieve. The problem lies in the fact that key factors 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. Most of this data is created in the form of *single blind, double blind,* and *triple blind* clinical testing.

The idea in a blind clinical test 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.* The level of blindness is as follows:

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 further, and some if its implications, in Section 4.2. In particular it should be noted that only experiments which produce "interesting" results tend to get published, so if 9 double blind tests show no correlation between a certain drug and disease, but 1 double blind test shows a correlation, most people only see the one test which shows a correlation.

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.

Even with this problem, 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 avoiding experimental data is often the hypothesis one wishes to test 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 those people harm. (Imagine, for example, performing a double blind test to determine if smoking cigarettes with a filter is less harmful than smoking cigarettes with a filter.)

Two other strong reasons for avoiding experimental data is the high cost of performing the experiment and the long time the experiment would require. In healthcare the costs of experiment data come from paying the participants, supplying the drug, supplying 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 require to perform proper experimental tests is often highly impractical. When researching in terms of health time scales are generally considered in terms of life times, or at least years. For example, although theoretically one could create an experiment which tests how the consumption a vitamin supplement pill alters the probability of participants catching the flu, however the experiment would have to be run in a controlled environment for years before completion. This is clearly impractical.

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 historical data is collected. Health record data refers to data that is maintained by healthcare providers regarding when, how, and why individuals access given points of the healthcare system, while 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 contains information on individuals personal habits, such as the frequency with which a person exercises. Another difficulty with health record data is it only involves individuals whom have actually used the healthcare system, while healthy individuals are unseen.

Survey data can get around both of these problems, as anyone can be requested to participant and any question can be asked. However, survey data has often been critiqued as inaccurate as participants tend to over emphasis their "good" traits and under emphasis 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 XXX Cite. This is enhanced as an individual mass becomes farther from the norm.

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

4.1.2 Serial Data

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

Serial data may be collected in either a cross-sectional or longitudinal manner. In *cross-sectional* data a new collection of individuals is surveyed at each point in time[1]. This provides a series of "snapshots" of the population in various points in time and therefore provides some insight as to how the population dynamics are changing over time. In *longitudinal* data (sometimes called *panel data*) the same collection of individual is survey at each point in time. This type of data is considerable 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 serial data one separates the participants into a collection of cohorts. A *cohort* is a group of individuals from a given population that are defined by experiencing an common event in 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).

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 easily abstracted from health records.

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

4.1.3 Electronic Health Records

Electronic health records are one of the keys to modernising the health system and improving access and outcomes[**?**]. As electronic health records become implemented, high quality comprehensive data sets will become more and more readily available.

In Canada, a drive for standardisation of electronic health records is being headed by the *Canada Health Infoway*. Standardised health records will automatically ensure that any data originating

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

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. For more information on the Canada Health Infoway see Appendix B Section **??**.

4.2 Data Quality and Data Biases

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

4.2.1 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 quite likely would include the portion of the population which 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 ignore 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 whom 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 error CITES; unfortunately, there is very little research on how to prevent it.

4.2.2 Non-sampling Errors

Non-sampling errors are errors which 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 one creates a data set through a survey, one runs the risk on a poorly designed survey skewing the results. For example consider the following questions:

- *Do you read at least two books a month?*

- *On average how books do you read in a month?*
- *How many books did you read last month?* and
- *What books did you read last month?*

All three of the above questions essentially measure the same thing. However, the differ phrasing may led 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 that may be closer to the actually average number of books they read each month. The third and fourth question are better, in that respondents will more likely answer truly, but 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 in [**?**] it was found that survey results depended on the order in which the questions were asked, and that 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 an 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 over emphasis what society considers good. For example, self reported weight tends to be lower than actual weight, and self reported height tends to be higher than actual height CITE. Similarly people tend to under report their role in illegal activities (such as drug use) and over report their role in XXX CITES.

It should be noted that, in some cases self-reported data is more important than scientifically accurate date. For example, self-reported health status (and not actual health status) is often considered to be an important determinant of health care utilization.

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

Another challenge in implementing surveys is the characteristic low response rate. In practice most surveys receive response rates of less than 50% and many less than 25%. To combat this marketing firms have developed various methods to increase response rates within a survey. Some simple suggestions include:

- keep the survey brief,
- provide a incentive (the respondent receives cash, gifts, entry to a lottery, etc... for completing the survey),
- guarantee anonymity, and

- explaining to potential respondents how the survey can help society.

Many more nefarious, and sometimes dishonest, techniques for improving survey response rates exists, but we leave those for curious readers and marketing firms to research themselves.

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

In terms of experimentation implementation errors, the most common error is the result of an experimenter biasing the results due to not being blind to the participants status. This is best corrected 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 typos that all humans are likely to produce, there is the tendency for health record data to over emphasis 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 wait list 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 much of the health record data for hospitals is inputed by nurses who have other duties. When hospitals are busy the other duties take precedence and the data is not entered. Later it may be found that some data gets lost by the time the nurse has enough free time to enter data. Thus data accuracy is only assured if data inputers are given sufficient and appropriate times to enter data.

4.2.3 A final note

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

Finally, some researchers attempt to increase the size of their data sets by pooling data from various surveys or experiments. Studies based on pooling of data should be questioned, because of the difficulty in controlling for the varying biases within the dataset.

4.3 Related Reading

Building on Values, Report of the Romonow Commission on the Future of Health Care in Canada, p. 77 (2002).

Surveys of self-reported health status are commonly used to assess correlations between socioeconomic status and health. However, this raises the important question, “Is there a correlation between socioeconomic status and how people assess their health?” If there were such a correlation, then it would introduce a bias into surveys of this nature. This question was addressed in **?**]. They compared the results of a health status survey with mortality data in the Swedish Survey of Living Conditions for 1975–1997. They did not find a strong correlation between socioeconomic status and self-reported health status. Therefore, properly conducted surveys of self-reported health status are generally valid as a measure of the link between socioeconomic status and health.

Mention appendix on data sources

Mention Phillips work on errors in epi.

Chapter 5

The Basics

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

5.1 Model Overview

In mathematics, the fields of statistics and probability are intimately intertwined.

Descriptive stats, mean, median, mode

Types of distributions, important to pick well.

Confidence intervals

As mentioned, 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 intentionally display the data in a misleading manner. (Some examples of this are given in Subsection 5.4.1.) Policy-makers should be careful in interpreting graphs, and question any conclusions which are not clearly supported.

5.2 Common Uses

At some level statistics are used in almost every form of modelling. Any time a researcher collects data, they usually provide some level of 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?* and
- *What is the mean age for a women to first experience breast cancer?*

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

Any time a model involves random events, a researcher must select an appropriate distribution to describe how the random events are selected. Models which involve random events, and therefore probability distributions, can be found in XXX (chapter list).

5.3 Mathematical Details

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

5.3.1 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 which give the reader a sense of the data's central tendency and degree of separation. In other cases, researchers rely of charts and graphs to help describe the data.

The phrase 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 is often denoted by $\bar{x}$ or μ and is defined by the well known formula $\mu = (x_1 + x_2 + \ldots + x_N)/N$. Thus the mathematical word "mean" corresponds to what is commonly refereed to as the average. To compute the median we begin by sorting the data: $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_N$, and then simply use data element $x_{N/2}$. In the case where $N/2$ is not an integer, we use the mean of data elements $x_{(N-1)/2}$ and $x_{(N+1)/2}$. 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 multiply modes.

The phrase degree of separation refers to descriptions of how close the average data element is 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

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

where μ is the mean of the data. The standard deviation is then denoted by σ and is the square root of the variance, (hence the σ^2 in the definition 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 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 5.3.3, for now it suffices to say the consequence of this 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, 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 intentionally display the data in a misleading manner. (Some examples of this are given in Subsection 5.4.1.) Policy-makers should be careful in interpreting graphs, and question any conclusions which are not clearly supported.

5.3.2 Basic Probability

The probability of an event E is denoted $\Pr(E)$ and 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 occurs 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 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. These outcomes are listed in Table 5.1.

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 5.1: Possible outcomes for the sum of two rolled dice

From Table 5.1 we can easily count to see there are 6 ways a sum of 7. Therefore the probability of totally 7 is

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

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

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

Notice that the probability of totally 5 is equal to the probability of totally 9, these 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 totally 5 given that the green die rolls a 4. Notice since the green die is fixed at 4 there are now only 6 possible outcomes, one of which is a total of 5. Thus,

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

Similarly we find

$$\Pr(\mathit{total} = 7 | \mathit{red} = 3) = \frac{1}{6}, \quad \Pr(\mathit{total} = 5 | \mathit{green} = 5) = \frac{0}{6}, \quad \Pr(\mathit{total} = 12 | \mathit{total} \geq 9) = \frac{1}{10}.$$

Notice, in some cases (such as $\Pr(\mathit{total} = 5 | \mathit{green} = 5)$) 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 these are beyond the scope of this book, we will take a brief look at probability distributions.

5.3.3 Probability Distributions

In the previous Subsection we discussed a simple example of a random event, rolling a pair of dice. In this example we were able to develop a table (Table 5.1) which described all the possible outcomes for rolling the dice. Using this table we are able to determine the complete list of outcomes and there probabilities. We present this in Table 5.2.

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

Table 5.2: Possible outcomes for the sum of two rolled dice

Table 5.2 represents what is called a *probability distribution.* Specifically, 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 also 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 $E(X) = \sum_x xf(x)$.

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 individuals mass, it is clear that there is not a finite number of options. (True, 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 which can take on any value, continuous distributions are used.

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.

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

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).

There are many different **pdf**s which 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.

First is the *multinomial distribution*. The multinomial distribution results from having a **pdf** which is formed from a series of steps (see the leftmost graph in Figure 5.1). The use of this distribution function is simply to recreate finite distributions in the framework of **pdf**s.

Second is the *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 5.1). This distribution has two parameters μ and σ which 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 which is equally likely to be above the mean as below the mean.

Figure 5.1: The probability distribution functions for a binomial distribution (top left), a Poisson distribution (top right), a normal distribution (bottom left), and a exponential distribution (bottom right).

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

Last, 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 in healthcare is the number of newly arriving patients into a hospital in a given hour. The **pdf** for the Poisson distribution is defined by

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

(see the top right graph in Figure 5.1). This distribution has one parameter λ which represents the expected number of occurrences during one time interval.

The Poisson distribution is a natural choice to modelling arrival rates for several reasons. Foremost is the logical supposition that an individual visiting a doctor or specialist is an independently occurring event which happens at a constant average rate. That is, the event of one individual visiting a given specialist in a 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 health care utilisation. For one, a restrictive feature of the Poisson distribution is that the mean and variance of a random variable following this distribution are equal (a notion called *equidispersion*). Departures from equidispersion can occur if the variance is either greater than the mean (overdispersion) or less than the mean (underdispersion) the mean [1][2]. Overdispersion, which can be caused by a large number of zero counts in the data set, is unfortunately common in health care data, as there will be some people who never utilize services.

There may also be reason to question that events occur independently at a constant rate. Events in health utilisation data are not always independent — multiple visits to a physician by the same patient are often related, for example — nor is the probability of event occurrence always constant. It seems likely that Poisson assumptions are often violated in health utilisation 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 health care 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 health care utilisation data [2].

5.3.4 Confidence Intervals, and Statistical Significance

Once a probability distribution is selected, the mean and variance of a data collection 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 $1-\alpha$*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 to 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. Combined these provide 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 return the appropriate z value. Clearly the desired results is for the standard error of the mean, the z value to be small, and the value of α to be small. Generally an $\alpha < 0.05$ is considered statistically significant, and αs greater than this value are considered insufficiently small to develop and any conclusions.

5.4 Examples

5.4.1 "9 out of 10 Doctors Agree" – Interpreting Descriptive Statistics

Let us pretend that an unscrupulous 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 shall call the unscrupulous 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 effectiveness of each product, the company examined 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 performed a case study of 500 people in which it asked 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 reported not requiring headache medication during the month of the study were excluded. The finding of this research is found in Table 5.3.

To present their results to the public, Company A produces the pamphlet found in Figure 5.2.

Let us critically examine this figure. The first portion of the figure comprises of 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. Let us begin with the survey itself. Since this was not presented, the public is left with the question of what is meant by a value of 5, 4, 3, 2, or 1? (Note they was actually no value for 4 or 2, something the public 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

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

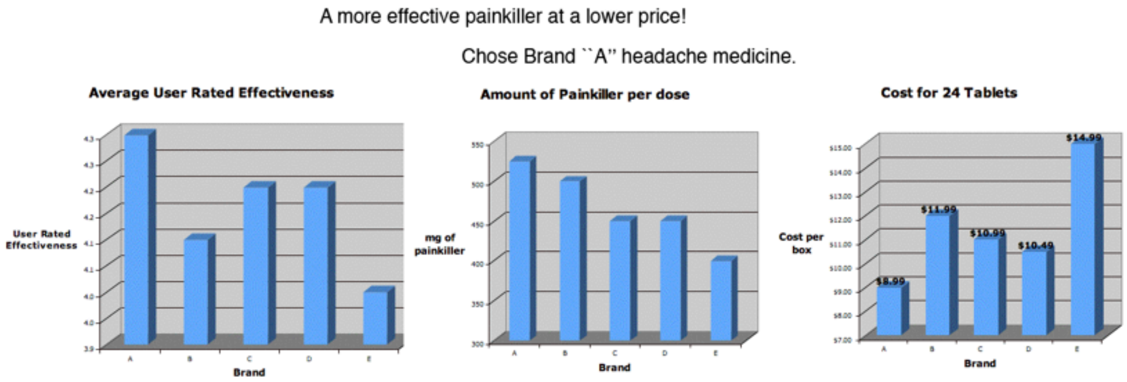


Figure 5.2: "Descriptive Statistics" developed by Company A.

$\sigma_A = 1.15$, 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 $z = 1.645$ [1]. Thus we have the confidence intervals

$$4.3 - 1.6450.115 < \mu_{A-\texttt{true}} < 4.3 - 1.6450.115 \Rightarrow 4.110825 < \mu_{A-\texttt{true}} < 4.489175,$$

and

$$4.0 - 1.6450.135 < \mu_{E-\texttt{true}} < 4.0 + 1.6450.135 \Rightarrow 3.777925 < \mu_{E-\texttt{true}} < 4.222075$$

for Brand A and Brand E respectively. Since these two intervals overlap, one cannot draw any statistical significance from the survey data.

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 require 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 appear to have it support the conclusion one desires. More information on this can be found in the popular book [4].

5.4.2 EX-TWO

5.4.3 EX-THREE

5.5 Related Reading

1. Cameron 1988
2. Cameron-book
3. A good statistics textbook
4. Huff, D. How to Lie With Statistics (book)

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 and Econometrics

6.1 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. That's 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 econometrics.

According to the Merriam-Webster dictionary, econometrics is the application of statistical methods to the study of economic data and problems, however in the field of healthcare it might be better to reverse this definition and say: *econometrics* is the application of economic theory to aid in statistical analysis.

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 econometrics of healthcare one attempts to develop a formula or equation which relates the *predictor variables* to the *response variable.*

Econometrics is the application of economic theory to aid in statistical analysis.
When econometrics is applied to something other than "cost of healthcare" it is often refered to as *regression analysis.*

It should be noted that, in econometric literature the terms *explanatory variable* and *predictor variable* are used interchangeably. We favour predictor variables to emphasize that when choosing predictor variables it is important that these variables be measurable

[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/`

before the responses variable is known. For example, suppose we are interested in determining an equation which will help provide a estimated cost for a patient undergoing chemotherapy to treat cancer. The response variable is the "cost of treatment." Explanatory variables might include things such as the initial size of the cancer discovered, the age of the patient, and the number of chemotherapy session the patient uses. The first two of these 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.

Once the response variable and predictor variables have been determined, econometrics uses statistical techniques to understand and predict how the value of the response variable will change for different values of the predictor variables. This begins by creating a hypothetical model under which one feels the data is likely to fit. (The term econometrics refers to the fact that these models are often based in economic theory.) 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 begin again.

Three possible explanatory variables for determining the response variable "time required to recover from hip replacement surgery" are:

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.

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, one 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, one generally expects body mass to increase with the square of a person 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. This is where the vast collections of economic literature become useful.

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, one would not expect the response variable to be exactly the same. Thus we assume, the response variable contains some degree of randomness. In associating a probability distribution with the data, one is quantifying how one predicts this randomness will behave. The choice of probability distribution function will depend on the nature of the data collected, so questions such as, "is the measured response discrete or continuous?" and "does one expect a skewing of the probabilities or that the measured responses will be evenly distributed about its average?" should be considered when deciding on what probability distribution to use.

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

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, and is generally done by a method called regression analysis. If a collection of parameters can be found which creates 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.

6.2 Common Uses

Econometrics is a subset of statistics in which one uses techniques developed in economic theory to help clarify the results one uncovers. Many of the questions approached by these techniques involve considering an individual's health as a personal asset which can be increased or decreased by the health choices an individual makes. As such these techniques are most 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?* and
- *How does physician density public access to healthcare?*

The ultimate goal of econometric analysis as applied to healthcare demand is to characterise the incentive structures underlying observed patterns, test the effects of incentive-altering policy, and to estimate future demand. Statistical models are often used to reveal associations between patient variables and the frequency of health care utilisation.

When econometrics is applied outside of cost and demand, it is often refereed to as regression analysis. In regression analysis one attempts to determine an equation which 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 one might try to create equations relating:

When the response variable is a risk of disease, then one should also use the specific field of statistics called epidemiological risk modelling. This is discussed in detail in Chapter 7.

- *number of cigarettes smoked per week* to *likelihood of contracting cancer*,
- *age and education* to *likelihood of attending a immunization clinic*, or
- *initial sense of pain and age* to *recovery rates after knee surgery.*

6.3 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 one seeks to develop formulae which can be used to predict what impact a change in the information will have on the probability of an event. (Henceforth we shall call the outcome we are seeking to predict the *response variable*, and information which we are using to make the prediction the *predictor variables.*) This is the primary goal's of econometrics and is generally accomplished in two parts.

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

Econometric analysis begins by creating a model under which one feels the data is likely to fit. These models are generally based in economic theory, which leads to the term econometrics (economic measurement). After creating a hypothetical model, econometrics continues by using 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 begin again.

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

A young student who wishes to know how much her car is costing her to drive. In order to develop an answer to this for several months she records her milage and her car related expenses (gas, insurance, maintenance, etc...). This hypothetical data is collected in Table 6.1. Examining Table 6.1 it is clear that the farther 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 determines that the formula

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

Table 6.1: Car related costs and milage by month.

$$cost/month = 88 + milage \times 0.25$$

gives a reasonable approximation of these numbers (see Figure 6.1).

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

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 relationship is impossible. In these cases, one must rely on economic theory to help select what 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 which are well suited to analysis in healthcare: normal linear regression, logistic regression, and generalized linear regression.

In *normal linear regression* one 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 6.3.1.

In *logistic regression* one assumes that the relationship follows an 'S' shaped curve called the logistic curve. This implies 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 collect data has a response variable that can only take on

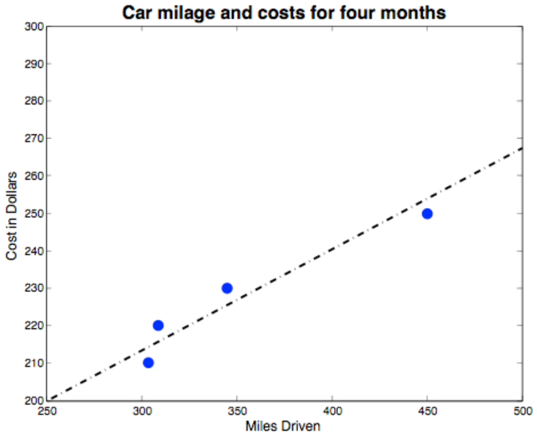


Figure 6.1: Car related costs and milage for four months. The dots are data points and the dashed line is the approximated linear regression.

one of two possible values, the common consensus amongst 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 6.3.2.

Generalized linear regression is a style of regression analysis which incorporates both linear and logistic regression. The mathematics behind generalized linear regression are quite deep, so we only provide a cursory overview of the subject (Subsection 6.3.3).

6.3.1 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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{6.1}$$

where X_i are the predictor variables, Y is the response variable, and β_i $(i = 1, 2, ...n)$ 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 $(i = 1, 2, ...n)$. Indeed in many cases one does 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 individuals 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 BMI can all be predictor variables.

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 maybe 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 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 $X_2 = X_1^2$.

The difficultly in developing a linear regression largely reduces to how to determine the coefficients β_i $(i = 1, 2, ...n)$. To do this we begin by making multiple independent observations of the response variable. If the predictor variables are something one can control, this can be done in the form of experiments which 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 which provide us with N independent observation of the response variable and N distributions of the predictor variables:

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

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

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

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

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

and we attempt to find the coefficients β_i $(i = 1, 2, ...n)$ which 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:

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

Taking squares of the differences between observed and fitted responses prevents positive and negative deviations from the line from cancelling each other when summing the errors.

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[1]. 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 $(i = 1, 2, ...n)$.

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

The normal distribution (a.k.a. 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.
The *central limit theorem* states that that the sum of the random variables with finite variance tends towards a normal distribution.

6.3.2 Logistic Regression

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

$$Y = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}} \tag{6.2}$$

where, as before, X_i are the predictor variables, Y is the response variable, and β_i $(i = 1, 2, ...n)$ are fixed coefficients used to describe the linear relationship between X_i and Y. As in linear regression, it is important to note that one 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 $X_2 = X_1^2$:

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

In order to compute the coefficients β_i $(i = 1, 2, ...n)$ for the logistic regression, we begin by considering the function

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

Notice that

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

Therefore $\log\left(\frac{P(x)}{1-P(x)}\right) = x$, and in particular, equation (6.2) tells us

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

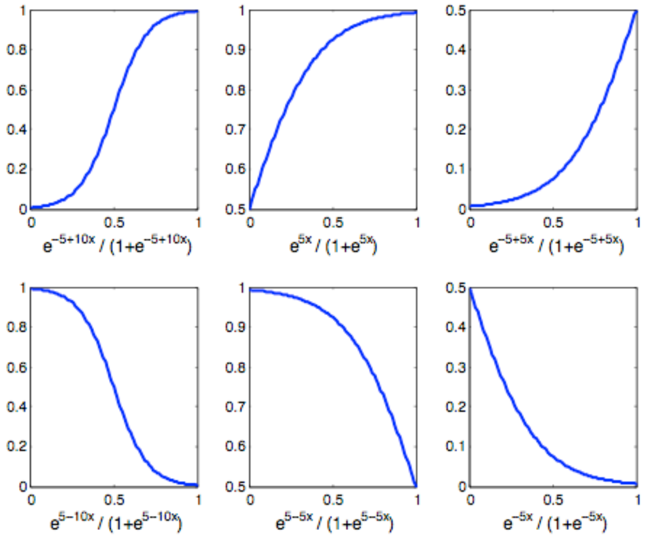


Figure 6.2: Examples of logistic curves. Logistic curves can take the shape of an 'S' *(top left)*, the shape of an 'inverted S' *(bottom left)*, or a portion of one of these shapes *(right)*

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 normal 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 to performing this estimation.

6.3.3 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 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) will be not be normally distributed, since it 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*.

In order to deal with regression analysis for non-normal distributions one usually turns to *generalized linear models*. The full mathematics of generalized linear models are beyond the scope of this book, but it is worth noting on some of their 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 distributions in the exponential family of distributions. These include (but are not limited to) the normal, binomial, Poisson, and gamma distributions.

Generalized linear models we replace equation 6.1 with

$$f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n. \tag{6.3}$$

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.

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

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, something called *maximum likelihood estimation* must be used. Most statistical software packages can perform this estimation.

An interesting final note is that if the link function f is the identify function (that is $f(Y) = Y$) then the generalized linear model reduces to normal linear regression, while if f is the logit function then the generalized linear model reduces to logistic regression.

6.4 Examples

6.4.1 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 employes 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 6.2.

# 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 6.2:

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

He begins by plotting the points $[0, 1.04]$, $[2, 1.18]$, $[5, 2.78]$, $[8, 5.31]$, $[11, 7.04]$, and $[14, 7.93]$ on a graph (see Figure 6.3). 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.

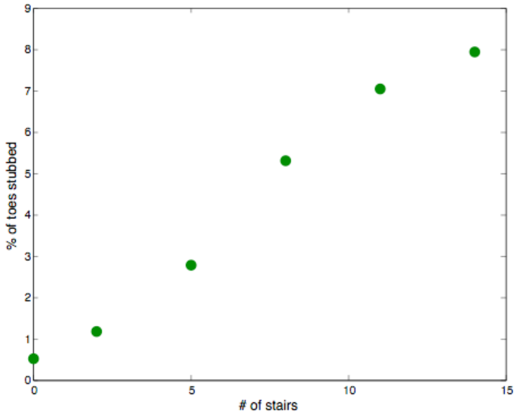


Figure 6.3: Percentage of toes stubbed by number of stairs to get to the office.

Examining Figure 6.3 he hypothesizes that a good fit might be found via logistic regression. This hypothesis is supported by several other factors. 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.

The researcher proposes the logistic curve

$$P = A\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\,\% \tag{6.4}$$

where P is the response variable representing the probability of stubbing one's 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. From here, he asks his favour statistics software package to fit the coefficients of equation (6.4) to the points $[0, 1.04]$, $[2, 3.53]$, $[5, 3.64]$, $[8, 6.19]$, $[11, 7.04]$, and $[14, 7.58]$. The software package tells him that the values $A = 9$, $\beta_0 = -3$ and $\beta_1 = 0.4$ 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/decends x stairs on the way to work. In particular he may conclude that is never more than a 9% chance of stubbing one's toe in any given month.

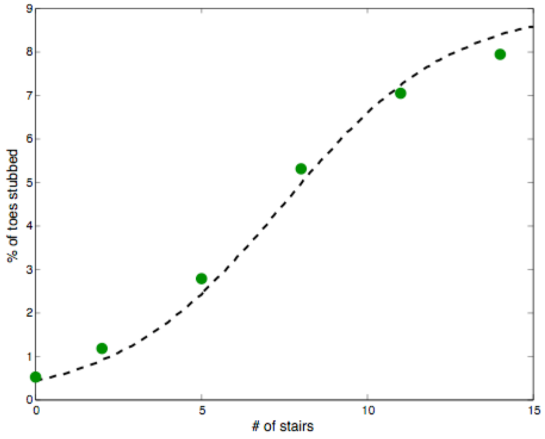


Figure 6.4: Percentage of toes stubbed by number of stairs to get to the office fit with a logistic curve.

6.4.2 Work, Wages, and Men's Health

The correlations between work-time, wages and health have been repeatedly addressed in econometric literature. One example is due to Haveman, Wolfe, and Kreider in 1994[2]. In this example we review this study and 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 (6.5)).

$$\begin{aligned} H_{it} &= \beta_0 + \beta_1 W_{it} + \beta_2 P^H_{it} + \beta_3 O^H_{it} + e_{it} \\ W_{it} &= \alpha_0 + \alpha_1 H_{it-1} + \alpha_2 R_{it} + \alpha_3 P^W_{it} + \alpha_4 O^W_{it} + u_{it} \\ R_{it} &= \pi_0 + \pi_1 H_{it-1} + \pi_2 W_{it-1} + \pi_3 R^P_{it} + v_{it}, \end{aligned} \tag{6.5}$$

In these equations H is health status, W is hours worked, R is wages, P^H is personal characteristics which determine health, P^W is personal characteristics contributing to work time, and P^R is personal characteristics contributing to wages. 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 [3] 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 however suggests that it is not how long you worked in a given job, but what that job is that determines your health status.

6.4.3 Recovery Curves for Post Surgery Physiotherapy

6.5 Chapter References and Related Reading

Econometric models have close connections to (XXX list models and chapters of note in the book XXX).

Reference XXX discusses XXX.

1. A textbook which includes Gauss-Markov theorem.

2. Haveman, R., Wolfe, B. & Kreider, B. "Market work, wages, and men's health." *J. of Health Econ.* 13: pp. 163–182 (1994).

3. Hansen, L. "Large sample properties of generalized method of moments estimators." *Econometrica* 50: pp. 1029–1054 (1982).

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

7.1 Model Overview

A recent study by Hakes and Viscusi showed that the general public has a serious problem in evaluating risky behaviour[1]. The study consisted of asking 493 adults a series of questions of the form “how many people do you think died 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 lead 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. Everyday, 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.*

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 refereed to as a *risk factor*, it should not necessarily be viewed as a negative. For example, the risk factor of “hand-washing” has been shown to reduce the risk of disease spread in hospitals[2]. On the other hand, numerous studies have linked the risk factor of “smoking” to an increased risk of contracting lung cancer[2] In the same manner it should be noted that, with a little creativity, one 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...), one could study what factors impact the chance of *avoiding*

epidemiology (”e-p&-”dE-mE-’ä-l&-jE): The study of the factors controlling the presence or absence of a disease or pathogen.

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

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

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 shall 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, one begins collecting data regarding the risk factors and diseases one wishes to examine. The data is then used 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 benificial.

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 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%.

Conversely, if the relative risk is less than 1, then one 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 could be (theoretically) reduced amongst the individuals that do not have the risk factor. For example, a prevented fraction of 0.37 means that 37% of individuals which do not currently have the risk factor would benefit from obtaining the

[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.

risk factor.

Regardless if the risk factor is beneficial or harmful, on may wish to examine the potential impact fraction. In the potential impact fraction, one assumes that they are able to make an intervention which somehow changes the prevalence of the risk factor in the population. (If the risk factor is beneficial one attempts to increase its existence, if it is harmful one attempts to decrease it.) The potential impact fraction then measures the reduction that would result from a given change in prevalence. This is a considerable more complicated concept than that of relative risk, attributable risk, and prevented fraction, so we leave further discussion to Section 7.3 and refer the reader to Example 7.4.1 for a detailed example on these concepts.

WARNING ABOUT EPI RESULTS and RESEARCH BIAS – Phillips work on errors in epi.

On a final note, in simple risk models the various statistical values that associate the risk factors to the disease may be computed analytically. However, in general the risk factors and the impacts of intervention hold complicated interrelationships which make analytic calculations difficult. In these cases one often resorts to computer based simulation to approximate the value of objects such as the potential impact fraction. An example of this is provided in Subsection 7.4.3 and more details on computer simulation can be found in Chapter **??**.

7.2 Common Uses

The goal of epidemiological risk modelling in healthcare is to develop models of the relationship bewteen 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 risk of a heart attack?* and
- *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?* and
- *Would a policy enforcing exercise programs in high-school have a significant impact on the number of heart attacks in young adults?*

7.3 Model Details

7.3.1 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 an event E occurs given factor F is present:

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

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

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

and the probability of a randomly selected female drinking beer is

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

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

$$RR = \frac{\Pr(Disease|Risk\ Factor)}{\Pr(Disease|Risk\ Factor^c)} = \frac{\Pr(D|F)}{\Pr(D|F^c)}. \tag{7.1}$$

Henceforth we shall denote the *D*isease by D and the risk *F*actor by F.

Since the probability of an event is always between 0 and 1, the relative risk is can be any number greater than zero (we assume that neither $\Pr(D|F)$ nor $\Pr(D|F^c)$ is 0, since these cases restrict the health outcome to occurring only if the risk factor is in the appropriate state).

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

$$AR = \frac{\Pr(D) - \Pr(D|F^c)}{\Pr(D)}. \tag{7.2}$$

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 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

In literature relative risk is sometimes also refereed to as the *likelihood ratio*.

$$PF = \frac{\Pr(D|F^c) - \Pr(D)}{\Pr(D|F^c)} \tag{7.3}$$

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

To compare the attributable risk and prevented fraction notice that

$$1 - AR = \tfrac{\Pr(D|F^c)}{\Pr(D)} \quad \text{and} \quad 1 - PF = \tfrac{\Pr(D)}{\Pr(D|F^c)}, \tag{7.4}$$

(provided $\Pr(D|F^c) \neq 0$). Therefore the attributable risk and prevented fraction are related by the equation

$$(1 - AR)(1 - PF) = 1 \;\text{ whenever }\; \Pr(D|F^c) \neq 0. \tag{7.5}$$

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 one 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 $0 < PF < 1$. Indeed, if the relative risk is greater than 1 then $\Pr(D|F) > \Pr(D|F^c)$, while $\Pr(D)$ must lie between these two values. This implies $\Pr(D) > \Pr(D) - \Pr(D|F^c) > 0$ and therefore $0 < AR < 1$. 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. This implies $\Pr(D|F^c) > \Pr(D|F^c) - \Pr(D) > 0$, and therefore $0 < PF < 1$.

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

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

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

Notice that these formulae for AR require not just knowledge of the relative risk, but also the probably 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 (7.5).

From the presentation given above it may appear that the calculation of relative risk, attributable risk, and the prevent fraction is straight forward. 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 straight forward 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 use to estimate the key values $\Pr(D)$, $\Pr(D|F)$ and $\Pr(D|F^c)$. The choice of statistical model to use is usually based on the type of data collect, and can result in biases in the end analysis. We refer readers to Chapter 5 for information on dealing with these issues.

7.3.2 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 one often calculates the potential impact fraction for each intervention. 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, and resulting in either an increase or decrease of 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:

$$PIF = \frac{(\Pr(D) - \Pr^*(D))}{\Pr(D)} \tag{7.6}$$

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 generalised 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

$$\Pr(D|F) = \Pr^*(D|F) \quad \text{and} \quad \Pr(D|F^c) = \Pr^*(D|F^c). \tag{7.7}$$

In words Assumption (7.7) 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 (7.7) we find

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

Next we recall that the definition of relative risk (equation (7.1)) and the fact that $\Pr(F^c) = 1 - \Pr(F)$ to simplify this to

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

Finally, applying Equation (7.4) we conclude (under Assumption (7.7))

$$PIF = (RR-1)(1-AR)\left(\Pr(F) - \Pr^*(F)\right). \tag{7.8}$$

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

$$PIF = (1-RR)\left(\tfrac{1}{1-PF}\right)(\Pr^*(F) - \Pr(F)). \tag{7.9}$$

On the surface Assumption (7.7) 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 is no, but if we look deeper we see that the surface answer is flawed. 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 preformed which (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 which make Assumption (7.7) false. In these cases it is often easiest to resort to computer based simulation to approximate the value of $\Pr^*$ under given interventions. An example of this is given in Subsection 7.4.3 and more information on computer simulation can be found in Chapter **??**.

7.3.3 Analysis using multiply risk factors

To be written. XXX

7.4 Examples

7.4.1 An Artificial Comparison between Chocolate Consumption and the Chicken-pox

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 Chicken-pox[4]. To test this she successfully contacts 3429 parents of children who had the Chicken-pox 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 Chicken-pox and F represent the factor of eating chocolate, she compiles the following table:

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

Table 7.1: Artificial data table relating chocolate consumption of Chicken-pox 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 them took four or more days to recover. If they had all consumed chocolate we would expect only 33.08% of them to take four or more days to recover, this is 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 Chicken-pox. To do this she notes that currently only $\frac{749+1515}{3429}100\% = 66.03\%$ of

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

children consumed chocolate while they had the Chicken-pox. If chocolate were freely available for children with the Chicken-pox, she feels this would increase to 98%. Thus the potential impact fraction for this intervention would be

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

That suggests that if chocolate were freely available to children with the chicken-pox, 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, it would be much more logical to group people into more than two categories with regards to chocolate consumption. (Not to mention, the data is completely fabricated.) Nonetheless, the above example does provide a simple demonstration of the prevented fraction, the potential impact factor, and their interpretations.

7.4.2 Mistakes and Misuses of Risk Analysis

Forthcoming Rockhill, and how not to use AR. XXX

7.4.3 Prevent: A Computer Simulation for Computing Potential Impact Fractions

In many cases the risk factors and impact of intervention hold complicated interrelationships which 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[3]. In this example we attempt to provide a broad stroke outline of the Prevent model. References for those who seek further information can be found in Section 7.5.

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 methodolgies to take into account two important facts:

i. 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.

ii. 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

$$PIF_t = \frac{\mathrm{Pr}_t(D) - \mathrm{Pr}_t^*(D)}{\mathrm{Pr}_t(D)},$$

where $\mathrm{Pr}_t(D)$ is the time-dependent probability of disease in the reference population and $\mathrm{Pr}_t^*(D)$ 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

$$TIF_t = \frac{\mathrm{Pr}_0(D) - \mathrm{Pr}_t(D)}{\mathrm{Pr}_0(D)}.$$

This measures the number of disease cases in the reference population that are prevented (or caused) by the autonomous time evolution of the risk factor prevalence. In Figure 7.1 we see how these values interrelate.

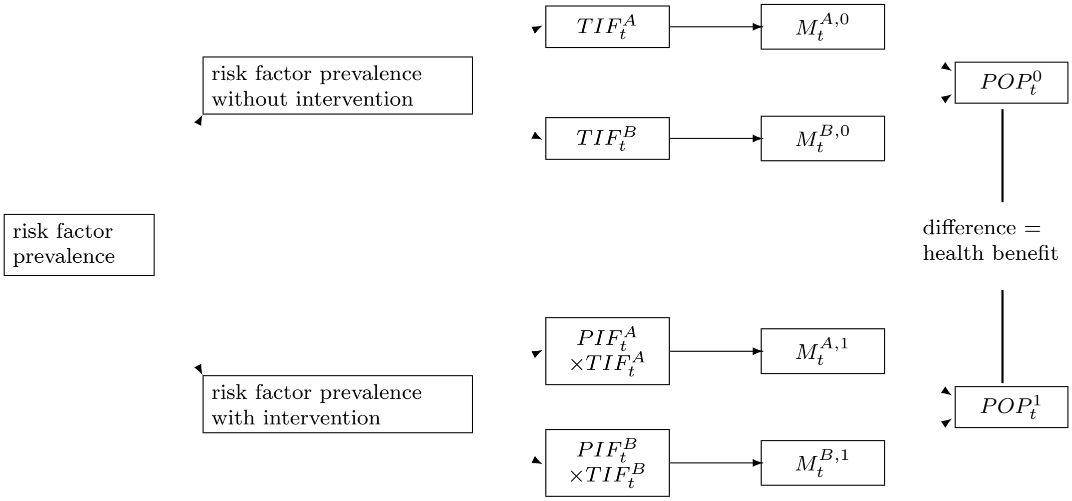


Figure 7.1: Simulation Flow in Prevent Model

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, repectively. 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 1989, Gunning and Schepers evaluated the Prevent model on Dutch population health data collected from a variety of sources[3]. The risk factors examined included cigarette smoking, hypertension, hyperlipidemia, occupational hazards, obesity, diet, and alcohol use. The diseases examined included ischemic 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.

In 2006, Bronnum and Hansen evaluated the accuracy of the Prevent model by applying it to synthetic population data generated by microsimulation[5]. 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.

7.5 Related Reading

Epidemiological Modelling has close associations with ...XXX

Reference [1] discusses the inaccuracy of public knowledge regarding risk. Reference [2] discusses the benefits and drawbacks of replacing soap hand washing with an alcohol hand rub in hospitals. References [3], [4], and [5], are examples of implementation of the Prevent Model.

1. Hakes, J. K. & Viscusi, W. K. “Dead Reckoning: Demographic Determinants of the Accuracy of Mortality Risk Perceptions.” *Risk Analysis*, 24(3): pp. 651–664 (2004).

2. Widmer, A. F. “Replace Hand Washing with Use of a Waterless Alcohol Hand Rub?” *Clinical Infectious Diseases*, 31, pp. 136–143 (2000).

3. Gunning-Schepers, L. “The health benefits of prevention: a simulation approach.” *Health Policy (Elsevier, Amsterdam)*, Jul 12(1-2), pp. 1–255 (1989).

4. Joseph, J. A. “The applicability, usefulness, and limitations of the PREVENT model, as demonstrated by modeling the effects of alcohol consumption interventions on coronary heart disease mortality.” *University of Toronto, M.Sc. Thesis*, Dept. of Community Health, (1997).

5. Bronnum-Hansen, H. “How good is the Prevent model for estimating the health benefits of prevention?” *J. Epidemiol Community Health*, 53, pp. 300–305 (2006).

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

8.1 Model Overview

In Chapter 7 we 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 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.*

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 US Surgeon General's 1964 announcement that smoking causes lung cancer[1]. 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. Most recently, in March of 2006, Scotland has banned smoking in *all enclosed public places*; this includes every pub, club and bar, and some outside shelters.

The term disease refers to any negative health outcome.
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. We use the word policy used instead of law since policy-based approaches

[1] On January 11th, 1964, the US Surgeon General Luther Terry released the 387 page report "Smoking and Health." In Chapter 4, Summaries and Conclusions, 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). A complete copy of the report can be found online at `http://www.cdc.gov/tobacco/sgr/sgr_1964/sgr64.htm`.

include examples such as the distribution of free condoms and clean needles in areas with a high prevalence of HIV.

In many cases, including smoking, policy-based interventions have had profound impact on the prevalence of a risk factor in a population[2]. 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 red wine and individual health[1]. These results suggest that moderate amount of red wine increase overall health, while large amounts decrease overall health. Clearly, enacting a policy enforcing adults to drink the appropriate 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 regards, 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 effect how they behave.

In psychosocial modelling the phrase perceived risk is used to refer to an individuals belief on how detrimental (or beneficial) a given course of action is to ones health. In this context, perceived risk can be thought of as a balancing of the answers to "what's 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. In regards to health care, 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 this knowledge is useless. 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 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 choice 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 a education-based approach to adjust risky behaviour. Even in hindsight it is often unclear which intervention has made the larger impact. One could argue that the decrease in the prevalence in smoking in the United States is not due to policy changes but instead due to a spreading of the knowledge that smoking is hazardous to one's health. Alternately one could point out that, despite this knowledge, every year over one million teenagers take up smoking[3], and argue that without laws restricting the purchase of cigarettes to minors this number would be even higher.

[2] 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.

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

8.2 Common Uses

The goal of psychosocial risk modelling in healthcare is to develop models of 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 amongst 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 the idea of 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 public school cafeterias adopt a policy of no longer serving high fat food?*

8.3 Model Details

The two most common approaches to adjusting risky behaviour lie in eduction and policy making. In education-based interventions, one attempts to increase the public's awareness of the risk factor and what they can do to reduce it. In policy-based interventions, one attempts to adjust the prevalence of the risk factor by developing laws or policies which 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 which 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*. (For more complete descriptions of models to explain human behaviour see Chapters 11 and 12.)

8.3.1 Perceived Risk and Perceived Efficacy

By *perceived risk* we refer 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 individuals perceived risk might be to educate them on the negative side effects of unhealthy eating habits. Alternately, a more policy-based intervention might be the enacting of laws forcing restaurants to provide nutritional information regarding each option on their menu.

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 with regards to this action may eat poorly based on a low perceived efficacy. For example, somebody in a corporate sale position may feel that their job requires them 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 which 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 regarding 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[2].

8.3.2 Education-based versus Policy-based interventions

Unfortunately there is often no clear cut answer regarding when one should use an education-based interventions 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 were 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[4].

Although there are no strict rules to determining what intervention will be most effective in a given situation, there are some ways to guide one's thought process. To begin it is always a good idea to examine pervious interventions. If a past intervention was particular 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 they can 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. This information can usually 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 would not be effective. In these cases more creative interventions are needed.

There are many arguments for implementing both policy-based and education-based interventions interventions to adjust health behaviour. For policy-based interventions the arguments are

[4] For more information on the Health Canada: Challenge to Youth Media Contest see `http://www.schoolfile.com/cash/HealthCanadaYouth.htm`

often along the lines of: without intervention, perceived risk and perceived efficacy are too low. New policies can increase perceived risk by enforcing a extra penalty 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 – see Subsection 8.4.3). 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 component 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 flaws in interpreting these measures[3]; we discuss this further in Subsection 8.4.2.) Similarly, journals such as the *American Journal of Epidemiology* are excellent sources of information for healthcare professionals, but 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.

8.4 Examples

8.4.1 Tanzanian soap-operas and HIV: a success story.

The country of Tanzania and its 37 million citizens lie 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[5]. This HIV infection rate is among the hightest in the world.

It has been found that the epidemic is maintained primarily through extramarital sex, with about 97% of cases occurring through heterosexual intercourse. Thus, modification of sexual behavior is potentially an important means of controlling the epidemic. In 2000, the results of a 4 years education-based intervention study were published[5]. In this example we outline the study, and highlight some of its successes.

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.

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.

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

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

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 promoted the following:

- 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.

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.

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.

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 about 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.

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.

Statistical analyses of the Twende na wakati project used ANOVA and logit loglinear models to compare the control and treatment areas. Additional analyses using logistic regression and ANOVA were carried out to account for 8 independent variables. Possible geographic effects were also tested by stepwise multiple linear regression models. (See Chapter 5 for descriptions of these models.)

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 education-entertainment can be a highly effective intervention impacting both perceived risk and perceived efficacy.

8.4.2 Interpreting AR and RR: a dangerous game

The issue of communicating information in a form the general public can make sense of has been addressed by various researchers ([1] and citations therein). As mentioned, there is a common consensus that measures such as risk ratio and attributable risk are be 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"[1]. 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 it 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[1] developed a simulation model using mortality data from the Framingham Heart Study[6]. Their simulation predicts that the potential life years gained from moderate alcohol consumption[6] 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 [1] 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 attributable risk (AR) and relative risk (RR) are developed. Equation 7.2 defines attributable risk as

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

while Equation 7.1 defines relative risk as

$$RR = \frac{\Pr(D|F)}{\Pr(D|F^c)},$$

where $\Pr(D|F^{(c)})$ is the probability of experiencing disease D given risk factor F is present (^{c} - not present).

8.4.3 SAFE injections: a successful policy change

Forthcoming, after special visit by a SAFE fellow.

8.5 Related Reading

Psychosocial Modelling is also discussed in XXX. It has close associations with ...XXX

Reference [1] ...XXX

1. Phillips, C. V., & Zeckhauser, R. "Communicating the Health Effects of Consumer Products: The Case of Moderate Alcohol Consumption and Coronary Heart Disease." *Managerial and Decision Economics*, 17, pp. 459–470 (1996).
2. Hakes, XXX
3. Rockhill
4. Pierce JP, Fiore MC, Novotny TE, et al. "Trends in Cigarette Smoking in the United States, Projections to the Year 2000." *Journal of the American Medical Association,* 261, pp. 61–65 (1989)

[6] 3 to 12 drinks per week, spread out evenly over the week

5. Vaughan, P. W., Rogers, E. M., Signhal, A., & Swahele, R. M. “Entertainment-education and HIV/AIDS prevention: a field experiment in Tanzania.” *Journal of Health Communication*, 5(supplement), pp. 81-100 (2000).
6. The Framingham Heart Study. `http://www.nhlbi.nih.gov/about/framingham/index.html`

Part II

Model Design and Interpretation

Chapter 9

Issues in Mathematical Modelling

> It's tough to make predictions, especially about the future. *Yogi Berra (1925-)*
> quote II *name (yyyy-yyyy)*

Model Selection, Development, and Implementation

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 I 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 satisfying. Alternately, if one wishes examine the complicated relationships between various societal factors and health, then the regression analysis and econometrics of Chapter 6 provide most (if not all) of the tools necessary for the project.

However, in many cases researchers, and policy-makers, are more interested in "what-ifs" than what is. That is, policy-makers are happy to see that a drug is effective, but are really 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 II of this book describes some of the modelling techniques which have been used to answer these harder "what-if" questions.

Before turning our attention to these techniques it is prudent to lay down a key observation one should keep in mind while using these models. Namely, its 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 one uses, at some level the model should be rooted in real life, which means tuning the model using real statistical data. During this stage of modelling, the ideas and mathematics discussed in Part I of this book become necessary to proceed.

With this in mind, we note that the remainder of this Part 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 is no way exhaustive, but hopefully provides a broad background for any researcher or policy-maker interested in modelling in healthcare.

Regardless of the modelling technique used, at some level all models must be tuned using real data and the statistical analysis discussed in Part I of this book.

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

9.1 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 multiply 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 **??**, Section 3.1, so we will 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 which 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 **??** Section 3.1, so we will say no more here. Instead we turn our attention to the idea of a feedback loop.

9.1.1 Feedback Loops

An important concept in both qualitative and quantitative modelling is that of a feedback loop. A feedback loop is what 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 9.1 (this figure reappears in Figure 12.1 of Chapter 12).

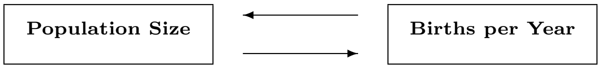


Figure 9.1: A simple feedback loop.

In Figure 9.1 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. To summarize, the first arrow states that the more people there are the more babies are born, while the second are 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 there construction and interpretation. However, as models become more complicated feedback loops can become very difficult to analysis. 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 in most circumstances 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 the entire goal of a qualitative model is simply to describe of the feedback loops in a system.

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

9.2 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 **??** we will 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 sometimes a researcher will 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 work losing focus part way through the project.

The theoretical model is incomprehensible: One of the guiding principles of modelling is transparency. Models which 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 they model and determine if they believe the underlying assumptions are correct.

The model is unbelievable: It is important to seek out experts in the field one wishes 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: If the model does not fit the data, then it is best to assume that the model is wrong. Since this sometimes means throwing out months of work, some researchers are very reluctant to do this. Instead they may continually try to tweak and tune the model to make it work. This process is seldom successful, and usually results in an incomprehensible model 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.

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 they margin of error for the model is small. In other cases, this can lead to a long series of "model corrections" which result in an incomprehensible model that works only for the very specific situation is was designed for. 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.

9.3 Implementation of Models

Supposing that 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 not 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 simply equations which describe the state of the model at a given time, given a certain initial state. This is refereed 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 multiply techniques are employed. We now discuss each technique in turn.

9.3.1 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 many provide insight into the model, and 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 document (in fact it would essentially amount to a survey of the entire field of mathematics). A Instead, we will give a brief introduction to a few of the branches of mathematics which are most often applied to modelling. In each of the remaining chapters of Part II we will discuss the tools needed to analysis 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 I of this book, the field of statistics plays a pivotal role in development of models.

Calculus: Calculus is the study mathematics enhanced by the concept of a limit. Like statistics it is a huge field of study and 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 $\int f(t)dx$ 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 there 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 is the 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 refereed 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-dimsenional light. In particular, derivatives become gradients ($\frac{d}{dt}f$ becomes ∇f) and integrals get taken over sets instead of intervals ($\int_a^b f$ becomes $\int_S f$). In modelling multivariate calculus arises when the model has multiply 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 equation are *attractors* and *equilibrium states.* An attractor is a state to which the system evolves towards over time regardless of initial conditions (or at least for a wide variety of initial conditions). Equilibrium states are states that once obtained the differential equations 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 13.

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 be best run a hospital?" cannot be approached through optimization, but questions such as "what nursing schedule minimizes the number of staff hours while maintaining a given level of service?" can be approach through optimization.

9.3.2 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, or these languages are outlined in Appendix C 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 these 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 Runga-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 work in order to use it. However, it is best if the user has a decent understand of the methods employed so that they know whether these methods are appropriate in their situation. As an analogy, consider that one does not need to know how the combustion engine works to drive a car, but every good driver should know that the tires should be inflated before going anywhere.

9.3.3 Simulation

Sometimes a model is complicated enough that even writing down the equations which represent the model becomes overly challenging. In cases such as this one often moves to simulation software to try an 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 C 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.

Finally, although in the past simulations have been termed either *Discrete Event Simulation* or *Continuous Simulations*, current practice has put Discrete Event Simulations at the forefront. In discrete event simulation, the time flow in 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 vast 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 step for the simulation is crucial to its success. The *time steps* in a simulation corresponds 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 simulation runs 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 which 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.

9.4 Related Reading

XXX some textbooks which cover the stuff in section 9.3.1

Chapter 10

Viewing the System as a Whole

> The science of healthcare has progressed more rapidly than our ability to manage healthcare as a truly integrated system. *Randolph W. Hall (-)*
> We have a lot of people revolutionizing the world because they've never had to present a working model. *Charles F. Kettering (-)*

System Dynamics and Systems Thinking

10.1 Model Overview

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 organisation of healthcare systems are failing[1]. 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 being to apply *systems thinking* to the field of healthcare, and turning to *system dynamics* to help quantify their models.

Systems thinking is more of a style of thought than an actual modelling technique. Overall, *systems thinking* can be viewed as a form of qualitative modelling which focuses on viewing the system as a whole instead of a collection of individual parts. The method is based strongly 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 new, many policy makers will already be familiar with systems thinking. Generally systems thinking models are the type of models which show up in a boardroom during a talk on "how X impacts Y." Often, but not always, they are best visualized as a collection 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 10.1, 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

New Medicines – + –¿ Lives Saved ¡– - –Medicinal Errors ¡– - – Staff Training.

Figure 10.1: 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.

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 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 an decrease (note the negative sign) in the number of lives saved. To counter this effect on must use staff training to decrease the possibility of medicinal errors. The model suggests that in order for new medicines to have there maximum impact, their introduction must be accompanied with 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*[2]. To understand the concepts of 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 now becomes quantified.

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 happy side effect is that the systems thinking beginning usually makes the model more understandable and believable later.)

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 model shifts to a quantitative model by adding the equations which describe each arrow in the model. The equations should be developed and tested using actually data, which gives the model a 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 patients entire past history. In some cases, such as modelling wait times for surgery, this is significant, in most these details 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 14. In cases where the entire history of a patient is important, Discrete Event Simulation is probably the best option, see Chapter **??**.

10.2 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 is population age demographics going to impact the healthcare system?* and
- *How 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 better implement intervention strategies for better control of disease?*
- *Would hiring more cleaning staff alleviate the hospital bed crisis?* and
- *XXX?*

Unfortunately, this versatility often makes it impossible to approach System Dynamics via analytical means, so simulation techniques become necessary.

10.3 Mathematical Details

10.3.1 Systems Thinking

Systems thinking is more of a style of thought than an actual modelling technique. Overall, it is a form of qualitative modelling which 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, but not always, systems thinking models are best visualized as a collection 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 manners than the box and influence diagrams described above. As one of the major goals of any systems thinking models is to clarify interactions between various parts of a system, any visual or descriptive model which accomplishes this suffices. If the model does not accomplish this, one should “rethink” the system.

An example of a systems thinking model which employs the “box-influence” visualization method can be found in Figure 10.1, page 88. An example of a systems thinking model which does not employ the “box-influence” visualization is detailed in Example 10.4.1.

10.3.2 System Dynamics

After a systems thinking approach is employed, and a box-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 one often turns 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* derives from the business concept of a stock, which is 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 whom require home care nursing. Returning to systems thinking, stocks measure the contents of the boxes in a box-influence diagram.

The term *flow* also derives from concepts in economics, and refers to the total value of changes to a stock during an given period. Thus in health care, 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 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 a box-influence diagram flows measure the amount of influence an arrow has on a given box.

Of course stock and flows can measure objects other than patients. In fact, one of the strengths of systems modelling is stocks and flows can measure any object which 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 a box-influence diagram created, the next step in creating a system dynamics model is to determine the equations which govern each stock. These equations generally rely on the state of time t and the state of the system $S(t)$ at that time. For example, consider the box-influence diagram in Figure 10.1. To flesh this out into a full systems dynamics model we make some small changes to the model to develop the diagram shown in Figure 10.2.

round two

Figure 10.2: A reconstruction of the systems thinking model in Figure 10.1 as a system dynamics model.

In Figure 10.2 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 can 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 number of change in the medicines employ per month is the number of new medicines introduced, we have

$$\frac{d}{dt}M(t) = N(t). \tag{10.1}$$

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 reasonable accurate.) This leads to the differential equation

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.

$$\frac{d}{dt}E(t) = \alpha N(t) - \beta T(t), \tag{10.2}$$

where α and β are constants.

Finally, the function $L(t)$ is 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 100 lives. That is

$$\frac{d}{dt}L(t) = 100M(t) - E(t). \tag{10.3}$$

Equations (10.1), (10.2) and (10.3) 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 (10.3) to obtain

$$\frac{d^2}{dt^2}L(t) = 100\frac{d}{dt}M(t) - \frac{d}{dt}E(t),$$

then use equations (10.1) and (10.2) to replace $\frac{d}{dt}M(t)$ and $\frac{d}{dt}E(t)$ to obtain

$$\frac{d^2}{dt^2}L(t) = 100N(t) - (\alpha N(t) - \beta T(t)).$$

Integrating this twice we obtain the solution

$$\begin{aligned} L(t) &= \int_0^t \int_0^t (100 - \alpha)N(t) + \beta T(t))d\tau_1 d\tau_2 \\ &= (100 - \alpha)\int_0^t \int_0^t N(t)d\tau_1 d\tau_2 + \int_0^t \int_0^t \beta T(t))d\tau_1 d\tau_2. \end{aligned}$$

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 often very useful to lean on the growing selection of system dynamics software available. Some of this software is reviewed in Appendix C.

10.4 Examples

10.4.1 Going Solid: The Danger of Lacking Wiggle Room

In 2005 Cook and Rasmussen published an exercise in systems thinking which examines how hospital operating budget and staff workload might impact the safety margins for practitioners of the healthcare system[3]. 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 that patients are admitted 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 suppose 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"?

Consider three of the major factors which effect 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 is overworked (or perceives itself 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 a operating room table long after the surgery is complete is generally considered bad for the hospital image.

Now consider these three factors as sides of triangle, and the actual operating state of the hospital as a point inside of the triangle (see Figure XXX a). 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 accepted a greater number of accidents and the public acceptance boundary is loosened without inciting public outcry (see Figure XXX b).

This simple model demonstrates how the constant pushes of budget and workload tolerance are overwhelming the standard operating proceedings, forcing hospitals into states which 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 explore by further systems thinking, or developing psychosocial models on how the public views hospital practices (see Chapter 12 for more information on psychosocial modelling).

10.4.2 EX-TWO

10.4.3 Predicting Future Usage for Home and Community Care

XXX to be inserted in Sept.

10.5 Chapter References and Related Reading

Systems Thinking produces qualitative models that have close connections to (XXX list models and chapters of note in the book XXX).

System Dynamics models have close connections to (XXX list models and chapters of note in the book XXX).

Reference XXX discusses XXX.

1. plsek-2001

2. Forrester1961

3. “Going Solid”: A Model of Systems Dynamics and Consequences for Patient Safety, Cook, R., and Rasmussen, J. *Qual. Saf Health Care*; **14** pp. 130–134 (2005).

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)*

Game Theory and Human Capital Models

11.1 Model Overview

The question of why people behave as they do is one of the oldest and hardest questions worldwide. Usually this question is examined by psychologists via a battery of creative psychological experiments. However, recently, mathematicians have begun their own assault on the question of human behaviour via the mathematics and logic of 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 the maximizing personal gain. In such models each *players*, that is the individuals being modelled, decision impacts the gain of the other players in the game. Therefore players seek 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 the *human capital model*.

In order to quantify personal gain, each decisions are 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[1][2]. However, more recently researchers have considered short term and long term health as quantifiable factors, and used these to develop what are refereed to as the human capital model.

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

Before discussing human capital, it is prudent to bring up the idea of the *Nash equilibrium*. Introduced by John Nash in his Ph.D. thesis[3], 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 which can be bought or sold for other assets.

Returning to healthcare, game theory has a clear usage in examining how policy changes regarding financial incentives will effect doctor and patients 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. 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[4][5][6][7][8].

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 players perception of how the other players will react to a given decision. For example, will providing training to one's employees encourage them to seek better paying jobs? The ideas are further complicated by ideas 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 game theory works on the assumption 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 which 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[9][10][11][12][13][14]. 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.

11.2 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 health care system, addressing questions such as,

- *How does trust effect doctor-patient cooperation and quality of care?*
- *How do incentive structures effect physician decisions?* and

- *How might user fees effect 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 with captures the idea that health is a *stock* which 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?* and
- *How can be better understand drug addiction?*

11.3 Mathematical Details

11.3.1 The Prisoner's Dilemma, a Introduction to Game Theory

Fortunately game theory models often do not require advanced mathematics to understand. Unfortunately, they often require a long and carefully thought-out series of logical thinking 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 classical prisoner's dilemma.

Possibly the most famous example of 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 results of each prisoners choice is usually called the *payoff* and captured in a *payoff table.* To demonstrate, in Table 11.1 we provide the payoff table for each combination of strategies. The table represents the jail time each prisoner will incur given their and their partner's strategy.

Next we consider the Nash equilibrium for the prisoner's dilemma. The *Nash equilibrium* is defined as XXX. 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

		Prisoner a	
		N	T
Prisoner 2	N	(1 year, 1 year)	(0 years, 10 years)
	T	(10 years, 0 years)	(6 years, 6 years)

Table 11.1: 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.

does not testify, then it is in prisoner 2's best interest to testify. However, regardless of prisoner 2's action, prisoner 1's jail time is decreased by testifying (either by 1 year or by 4 years). Hence prisoner 1's best course of action is to testify. Reversing 1 and 2, the same argument shows prisoner 2'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 mathematics 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 (see Example 11.4.1), the demand for pharmaceuticals[1], and the rising cost of hospital nursing staff[2]. In all of these cases the overall optimal solution is offset by that fact a individuals can gain (or at least not lose as much) by not playing to the communal good.

11.3.2 Zero-sum Games and the Maxi-min 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, henceforth we shall 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 they the *payoff table* is consulted to determine the outcome. If a game is played multiply 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.

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.

An alternate definition of a *zero-sum game* is: a game in which wealth is never created nor destroyed.

[1] http://abcnews.go.com/Technology/WhosCounting/story?id=98179

[2] http://www.gametheory.net/News/Items/059.html

It is easy to provide examples of zero-sum games, as most forms of 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 (that is, finding the strategy that rational players should follow) zero-sum games can be accomplished by a series of techniques. The two most common are *dominance,* the *mini-max criterion.*

The idea of dominance is based on eliminating strategies which are dominated by another strategy. 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 successful solved by dominance in provided in Table 11.2.

	A	B	C
x	(5, -5)	(10, -10)	(10, -10)
y	(0, 0)	(-25, 25)	(-10, 10)

⇒

	A	B	C
x	(5, -5)	(10, -10)	(10, -10)

⇒

	A
x	(5, -5)

Table 11.2: 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 dominants player 2's payoff, so player 2 should always select strategy A. The end result is player 1 gaining 5 per round.

Dominance solutions rely on the payoff table having strategies which 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 which maximizes 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)

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.

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 to using that strategy. However, player 2 will probably notice this, especially if player 1 uses strategy y every time, and therefore lean to using strategy A. This would 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, we propose that both players select their strategies randomly with some probability distributions.

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

$$E(p, q) = 5pq - 20p(1 - q) - 10(1 - p)q + 15(1 - p)(1 - q).$$

(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 Minimax Problems*. In 1944, Von Neumann showed that the linear minimax problems which 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 have a gradient of zero[3]. In our case this yields,

$$\begin{aligned} 0 = & \;\nabla E(p, q) \\ = & \left(\tfrac{d}{dp}E(p, q), \tfrac{d}{dq}E(p, q)\right) \\ = & \;(5q - 20(1 - q) + 10q - 15(1 - q), 5p + 20p - 10(1 - p) - 15(1 - p)) \\ = & \;(50q - 35, 50p - 25)\,. \end{aligned}$$

Which implies

$$p = 0.5, \quad \text{and } q = 0.7.$$

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 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.)

Recall, the gradient function is the multi-dimensional version of the derivative. That is, it measures the slope of a multi-dimensional function.

$$\nabla f = [\frac{d}{dx_1}f, \frac{d}{dx_2}f, \ldots \frac{d}{dx_N}f].$$

11.3.3 Human Capital Models

Although the human capital model was not developed under the guise of game theory, the ideas within bare 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.

[3] 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 [2].

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 individuals perceived health at a given time. Human capital theory supposes an individual will seek to maximize this stock over time.

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.*

The health stock at time $t+1$ is related to the health stock at time t by

$$H_{t+1} = H_t + I_t - \delta_t H_t \tag{11.1}$$

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 one's wealth can never become negative in order to buy more health stock.

One example of a wealth constraint is the *full wealth constraint*:

$$\sum_{i=0}^{n} \frac{P_t M_t + Q_t X_t + W_t(\Omega - \tau)}{(1+r)^t} = \sum_{i=0}^{n} \frac{W_t \Omega}{(1+r)^t} + A_0.$$

In the full wealth constraint Ω is the total amount of time available, τ is the time spent working, W_t is the wage rate, M_t is a vector representing all goods purchased which contribute to health, P_t is the price vector for goods contributing to health, X_t is the vector of other goods purchased, Q_t is the price vector for the other goods, r is the market rate of interest, and n is the total number of time intervals. 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 payoff 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 utility 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 is beyond the scope of this book.

11.4 Examples

11.4.1 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 interests or (whether through error, misjudgement, lack of skills, or conflicting goals) to recommend treatment that is not in the best interests 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:

Player 1:	The Doctor	Player 2:	The Patient
	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 care; patient follows the treatment plan,

(G, I): physician provides good care; patient does not follow the treatment plan,

(B, F): physician provides poor care; patient follows the treatment plan, and

(B, I): physician provides poor care; 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 patients point of view, following good advice is rewarding (payoff = +1) but following bad advice is detrimental (payoff=-1). Finally, ignoring the physician's advice causes no change in health status so we set the payoff to 0. This produces the following payoff table:

	F	I
G	(+1, +1)	(-1, 0)
B	(0, -1)	(0, 0)

This table has two Nash equilibriums, (G, F) and (B, I). 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 players 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 which allows some concept of trust and communication. Once these elements are added to the game the (G, F) equilibrium becomes stable.

11.4.2 Adaptations of the Grossman Incorporating Family and Acceptance of 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. Recent works by Bolin, Jacobson, and Lindgren and by Gjerde, Grepperud, and Kverndokk 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[15] and game theory which incorporates the role of family into the health decision process[20]. 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 the health of his wife. 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 health of his wife. 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 (e.g., 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 which are divorced are less healthy than other individuals. However, as noted in [20], this prediction is not entirely born out by the empirical evidence. 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, than 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 phenomena is addressed in [17] and [18] where the Grossman model is modified to incorporate adaption.

One method to incorporate adaption into a human capital model is to assume that the utility function depends not the objective health status, but on a definition of subjective health status. This could be done by defining the objective health function as

$$K(t) = \frac{H_0}{1+\beta} + (1+\beta) \int_0^t e^{-\beta(t-s)}\, H'(s)\, ds, \tag{11.2}$$

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 a parameter β. The larger β, the more the individual focuses on only recent changes in health. If β is smaller, then the individual takes into a greater time period when determining their subjective health status.

Gjerde, Grepperud, and Kverndokk, treat the lifetime in their model as an endogenous uncertain variable, by using a probabilistic *hazard function* to model the occurrence of death[18]. 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 adaption 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.

11.4.3 Rational Addiction

The theory of rational addiction, first developed in [21], 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 which are harmful to public health, such as smoking, alcohol use, harmful drug use, or obesity.

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 in [22]. They 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 [22]. 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 which are useful for evaluating public health policy towards addiction.

11.5 Chapter References and Related Reading

Game Theory models have close connections to (XXX list models and chapters of note in the book XXX).

In [23], we see how game theory can be used to examine wait times under various user fee scenarios.

The original application of human capital models to health is [4].

A human capital approach to modelling has been applied to pediatric healthcare[16], addiction [21][22], and to the role of the family in the demand for health[19][20].

1. Game Theory: A Non-Technical Introduction to the Analysis of Strategy. by Roger McCain (2006)
2. John von Neumann and Oskar Morgenstern's 1944 classic, Theory of Games and Economic Behavior (1944).
3. Nash Ph.D. thesis
4. becker 62
5. becker 67
6. becker 76
7. becker 92
8. becker 93
9. forni 99
10. kirman 92

11. kirman 93
12. gallegati 99
13. kirman 01
14. gallegati 04
15. grossman 72
16. grossman 78
17. kverndokk 00
18. gjerde 05
19. bolin 01
20. bolin 02
21. becker 88
22. clarke 06
23. farnworth 03

Chapter 12

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

12.1 Model Overview

For modelling purposes, it would be considerably easier if everybody acted is 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], everyone would attend their local vaccination clinic[2], and over 150,000 psychologists would suddenly be out of work[3]. Therefore, whenever one attempts to develop models 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 effect 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 which examine how social conditions effect how people make decisions about their health.

Psychosocial (sI-kO-'sO-sh&l): 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 *Behavioral Model for 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

[1] The *2006-07 Indian Gaming Industry Report* totals 2005 casino gambling revenue in the USA as $55.3 billion; Indian owned casinos accounted for $22.6 billion of this.

[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 .

one's usage of the healthcare system is impacted by one's social circle. More recently, advances in computers and mathematical data analysis have allowed these theoretical models to be examined and verified.

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 liklihood of achieving that goal. Essentially the health belief model suggests that an individuals 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 12.3.1, page 110). The goal in researching the health belief model is to determine how these elements interrelate and what factors effect them. This is described in some detail in Subsection 12.3.1.

The Behavioural Model for Healthcare is based on the concept that an individuals 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 personal, 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 which describes these feedback loops is called a *demand-access-utilization chain.* The goal in researching the behavioural model for healthcare consists of examining the interrelationships between the *physical environment, social environment, health behaviour,* and *health outcome.* Details to this are discussed in Subsection 12.3.2, and the demand-access-utilization chain of the Behavioural Model for Healthcare is given in Figure 12.2.

12.2 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 of Psychosocial modelling in healthcare is the examination of when and why people make use of healthcare. In this regards, 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 with regards to medical instruction (such as proper use of antibiotics)?* and
- *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 one attempts to model general patient behaviour on a group level, one 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?* and

- *Who should we educate to reduce unwanted pregnancies?*

12.3 Model Details

Two of the most popular Psychosocial models in healthcare are the *Health Belief Model* and *Behaviour Model for 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.

12.3.1 The Health Belief Model

The health belief model (henceforth refereed 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 an individuals likelihood of access healthcare is based on six elements: *perceived benefits, perceived susceptibility, perceived severity, perceived barriers, cues to action,* and *self-efficacy* (see summary in Table 12.3.1).

The element of perceived benefit captures the positive outcomes that an individual feels may occurs 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, or begin chemotherapy in the hope of prolonging one's life.

The element of perceived susceptibility captures the individuals 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, an 25 year old woman will perceive a lower susceptibility to breast cancer than a women of 55 years old, and therefore be less likely to have a mammography examination.

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 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 not been educated. Perceived severity also captures possible social consequences such as how a disease may affect an individuals work or family life. For example, influenza may be perceived as more severe to someone who cannot abide the possibility of missing work.

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

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

Table 12.1: Elements affecting an individuals access to healthcare according to the HBM.

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 individuals confidence in their ability to overcome the perceived barriers to accessing healthcare.	• Positive or negative reinforcement

getting the necessary time off work, would be perceived barriers to getting regular mammograms. Perceived barriers also includes 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 which causes a constant ache is more likely to prompt an individual to seek help, than a disease which does not. Another major contributor to action is media coverage.

Efficacy ('e-fi-k&-sE): the power to produce an effect.

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

As a Psychosocial model, the HBM is primarily designed to explain patient's 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 of "improve 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 one desires to improve attendance at an immunization clinic, one would like to know whether they should try 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 Subsection 12.4.1 below.

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.

12.3.2 The Behavioural Model for Healthcare

The *Behavioural Model for 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 individuals 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 increase 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.* The demand-access-utilization chain for this example can be found in Figure 12.1.

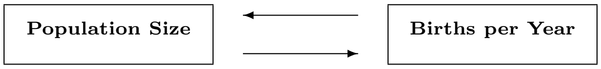


Figure 12.1: Feedback loops in a simple demand-access-utilization chain.

Figure 12.1 is a demand-access-utilization chain 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. Thus, the first arrow states that the more people there are the more babies are born, while the second are states that the more babies are born the more people there are.

[4] All statistics from the US census bureau.

Returning to healthcare, the behavioural model of healthcare can be captured in a similar demand-access-utilization chain. In Figure 12.2, we provide a version of the demand-access-utilization chain typically used in the behavioural model of healthcare (adapted from [4]). As before each arrow represents that the box it is pointing from has an impact on the box it is pointing to.

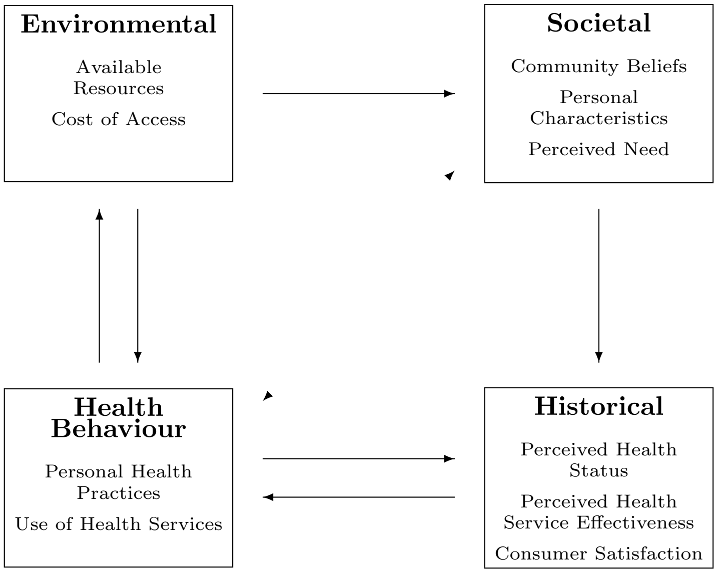


Figure 12.2: The feedback loops in the demand-access-utilisation chain for the behaviour model for healthcare.

The behavioural model for healthcare considers three main classification of factors that impact the use of health services: environmental, societal, and historical. The environmental factor largely consists of the availability of health services, and the cost 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.

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) which are not directly involved with the use of health services.

On Figure 12.2, we see that (according to the behavioural model of healthcare) 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.

The category titles given in Figure 12.2 differ in various research papers. Some common alternatives include *"Population Characteristics"* instead of Societal and *"Past Outcomes"* instead of Historical.

Displayed in Figure 12.2 we have shown one version of the behavioural model of healthcare. Figure 12.2 is actually a fairly simple version, as more modern versions include to 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 might also influence their usage 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 demand-access-utilisation chain by either adding new boxes, or by splitting pervious boxes into various parts.

Like the health belief model (Subsection 12.3.1) the behavioural model for healthcare has several drawbacks. As with the health belief model, the behavioural model for 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 in the behavioural model, one conclusion of the behavioural model for 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.

12.4 Examples

12.4.1 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)[2]. 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 in regards to the HBM.

Items 1 and 2, personal health goals and personal outcome expectations, can easily be conceived as a 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 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 seen as simply a new categorization of the HBM.

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"[2]. In order to support this claim he discusses the impact of self-efficacy on several previous studies into health behaviour.

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.

For example, in 1987 Meyerowitz and Chaiken[3] tested what style of pamphlet would have a 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 focusing 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.

Some of Bandura's more interesting example 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.

In all of these he argues that the increased self-efficacy is key to the success of the program.

Bandura's work provides several other examples along these lines. 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, future attempts to regulate health behaviour should include practices which are designed to raise an individual's self-efficacy.

12.4.2 Refining the Behavioural Model of Healthcare

Forthcoming.

12.5 Chapter References and Related Reading

Psychosocial models have close connections to (XXX list models and chapters of note in the book XXX).

References [1], [2], and [3] discuss various aspects of the Health Belief model. Reference [4] discusses the behavioural model of healthcare.

1. Janz, N. K., & Becker, M. H. "The Health Belief Model: a decade later." *Health Educ. Q.* Spring; 11(1): pp. 1–47 (1984).

2. Bandura, A. "Health Promotion by social cognitive means." *Health Educ. & Behav.* 31(2): pp. 143–164 (2004).

3. Meyerowitz, B. E. & Chaiken, S. "The effect of message framing on breast self-examination attitudes, intentions, and behavior." *J. Pers. Soc. Psychol.*, Mar; 52(3): pp. 500–510 (1987).

4. Andersen, R. "Revisiting the Behavioral Model and Access to Medical Care: Does it Matter?" *J. of Health and Soc. Behav.* Mar; 36: pp. 1–10 (1995).

Chapter 13

Modelling Social Interaction

> It is your business when the wall next door catches fire. *Horace (65 BC-8 BC)*
> Christianity teaches us to love our neighbour as ourselves; modern society acknowledges no neighbour. *Benjamin Disraeli (1804-1881)*

Network Models and Graph Theory

13.1 Model Overview

In 1967, a prominent psychologist by the name of Stanley Milgram preformed 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 Omaha knew the Bostonian on a first name basis, then they could mail the letter directly. Otherwise, the Omaha 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 people who mailed the letter was just 5.5. 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 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. Since only letters which successfully reached the target were considered in the final analysis. It is possible that the letters which did not reach the target were unconnected, or connected via very long chains.
2. Participants mailed letters based on their best guess of a shortest path, these guesses could be very wrong.
3. It is unlikely for the entire human population to be acquainted within six degrees of separation because of the existence of certain populations which have had little or no contact with people outside their own culture.

The reason we note this is that one of the primary goals of network models is to develop mathematical models which display the small-world property, and to explore how diseases travel along these models. Therefore, if one does not believe the six degrees of separation hypothesis then network models are not 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. To 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 impossible 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 which uses such a graph as its underlying structure.

The application of Network models to healthcare is a relatively new phenoma. However, a good deal of research has already been developed exploring how small-world property might impact the spread of disease. Other applications have been slower to arise.

13.2 Common Uses

Network models are models which 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 disearse. This leads to questions like,

- *How do we expect disease to spread through the network?* and
- *How can we adjust the network to better control the spread of disease?*

Other applications include,

- *Examining how social networks impact the demand for healthcare.* and [**?**].
- .

13.3 Mathematical Details

To understand network theory it is first necessary to discuss the mathematical concept of a graph. *Graphs* are mathematical objects that are composed of *nodes* (a.k.a. *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* which 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 13.1 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. To nodes are *connected* if a path exists between the two nodes. If all nodes in the graph are connected then the graph is usually refereed 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).

The concepts of paths and distances can be further complicated by the introduction of *directed edges*. A directed edge is an edge which connects nodes in one direction (the easiest analogy is a one way street.) For *directed graphs* the concepts of connected 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 10.

Given a network (or graph), one can assign to each node in the network a degree. The degree of a node is defined as the number of edges exiting the node. In the analysis of networks one of the first things done is often to bin 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.

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

$$p_k = \frac{\mu^k}{k}\exp(-\mu),$$

where the constant μ is determined from the data to fit the model. While the network follows are a power law if

$$p_k = Ck^{-\alpha}\exp\Big(-\frac{k}{\kappa}\Big),$$

where the constants C, α, and κ are determined from data to fit the network model. 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 which behave similarly to the real network. This allows for one to test interventions in a more robust manner.

In Figure 13.1 we illustrate how different probability distributions generate different types of networks.

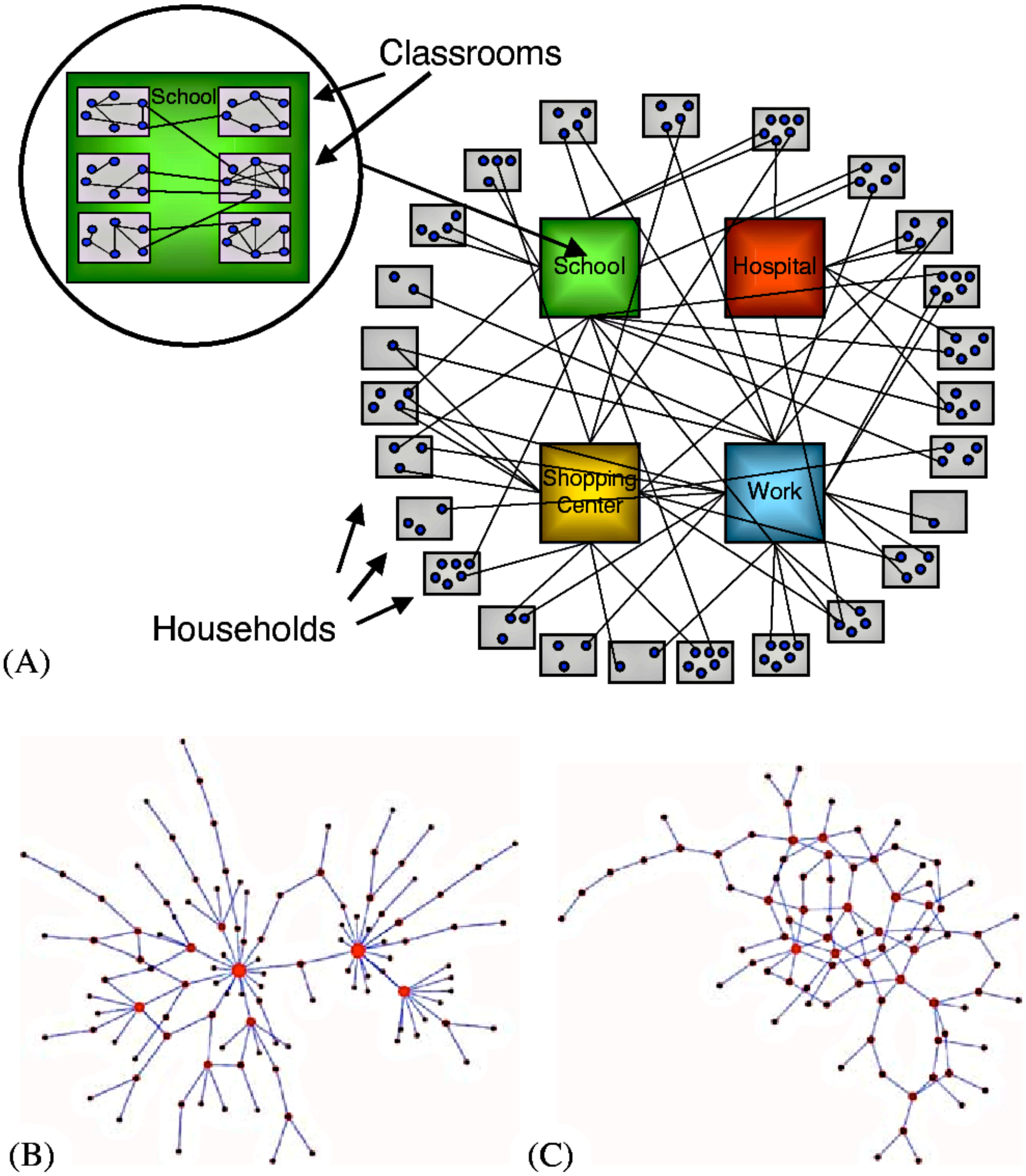


Figure 13.1: Types of Networks. (A) Poisson-law network, (B) Power-law network, (C) Hierarchical network XXX check, this doesn't look right to me. XXX

13.4 Examples

13.4.1 EX-ONE

Hmmm, maybe something to do with spread of diseases? The flu through a school?

13.4.2 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[5]. Although the reasons for this are unknown, many people have developed hypothesis 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 could 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 which demonstrated that the decrease in birthrate behaved in a manner consistent with a model which uses social pressure as one of its driving forces[**?**]. 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 opinion 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 which 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

$$S_i(t) = \text{sign}\left(\phi_i + F(t) + \sum_{j\in\mathcal{N}_i} J_{ij}S_j(t-1)\right) \tag{13.1}$$

where $\mathcal{N}_j$ 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 13.1 represents a model in which 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 (13.1), the influence of the opinions of our 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 begin by changing slowly if $F(t)$ is increased or decrease. Moreover, if $F(t)$ passes through 0 it will cause $S(t)$ to change rapidly for

[5] CITATION

several time steps, after which the rate of change of $S(t)$ will slow down. In this model, the 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[**?**]. 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 from Eurostat regarding 11 different countries (Belgium, France, Germany, Greece, Italy, Netherlands, Poland, Portugal, Spain, Sweden, Switzerland, 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 13.2.) As can be seen from Figure 13.2, 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 13.3. 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.

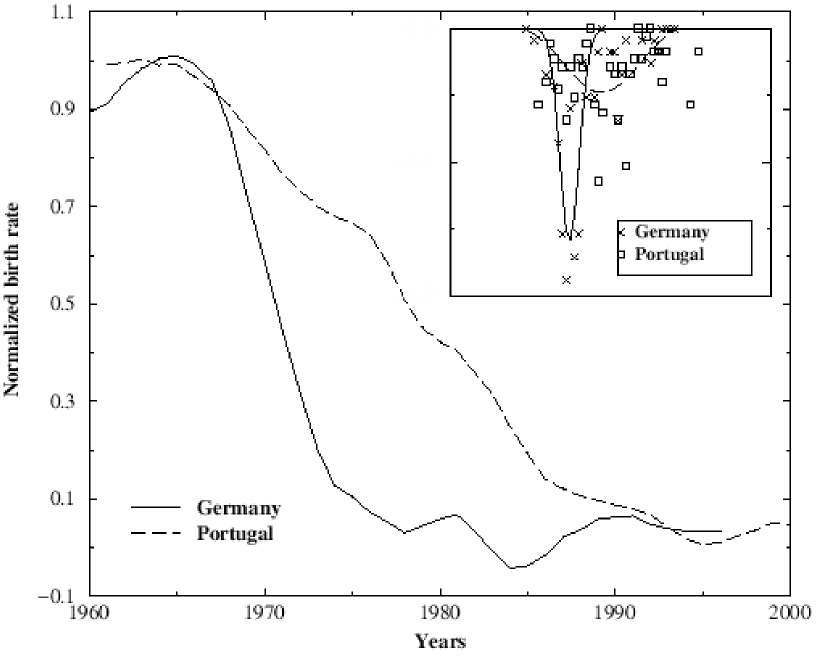


Figure 13.2: Birth rates of Germany and Portugal
These curves show the drop-off in birth rate for Germany and Portugal. The drop-off rate for the other countries is intermediate between these two examples.
Reproduced from [**?**].

The initial drop in birthrate would have been caused by an external factor, such as the availability of birth control pills. However, these results are strong evidence that once the phenomena took root,

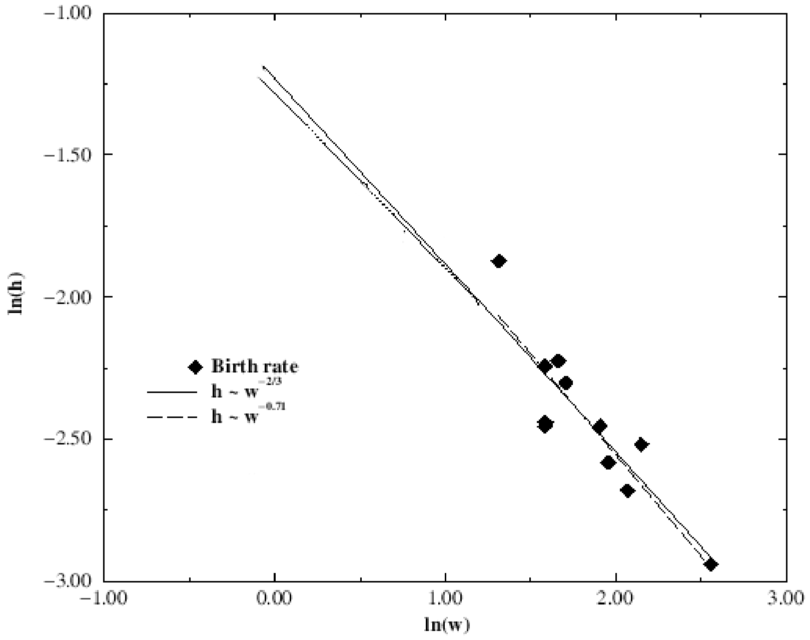


Figure 13.3: Peak of the Birth Rate Drop-off Rate versus the Width
The natural logarithm of the peak of the fecundity drop-off rate plotted against the natural logarithm of the width of the fecundity drop-off. A linear regression fit gives a slope of -0.71 ± 0.11. The RFIM predicts a slope of $-\frac{2}{3}$. Reproduced from [**?**].

social pressure became a driving force. Further study should confirm or disprove this hypothesis.

13.4.3 Control of Communicable Diseases in Healthcare Facilities

The study of social networks has a long history but until recently it has been largely descriptive. However, until recently global properties of the network, or its dynamics, have not been used to make predictions or recommendations. 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*[**?**]. In this example we summarize examine this work and summarize the results within.

Meyers et. al.'s goal was to develope a model which 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 the healthcare facility. A sample of such a model appears in Figure 13.4.

Next, the model was overlaid with a Markov state model similar to the epidemiological S.I.R.

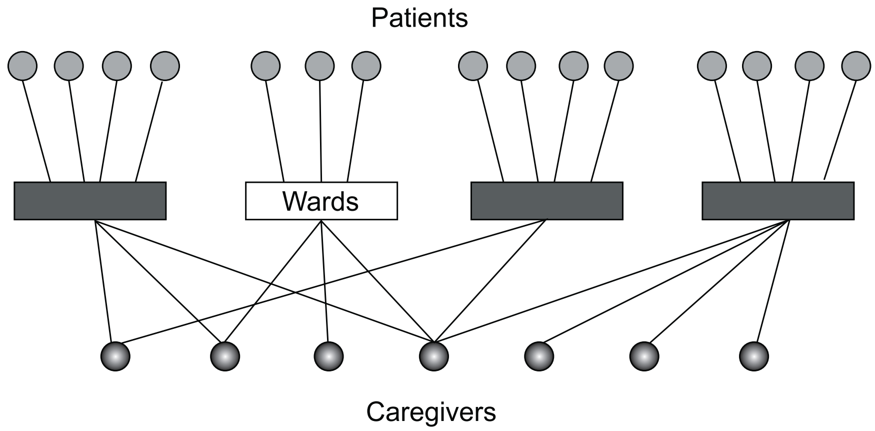


Figure 13.4: A Network of Health Care Facilities
Reproduced from **?**].

(Susceptible-Infected-Recovered) model. (For further information on Markov models and on the S.I.R. model see Chapter 15.) 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 of 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 in the epidemic started from a single case).

The model used three independent parameters 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 modelling 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.

13.5 Related Reading

Network Models have close associations with ...XXX

The science of the networks has recently became a booming field [**?**]. Its scope for possible applications is enormous. From street traffic, to epidemic spread [**?**], and biochemical networks in preventive healthcare [**?**], to name a few, countless systems can be studied as networks. The interdisciplinary nature of the network science and its importance for the economy is recognized now at the highest levels of administration. The National Research Council in the USA, for example, recommends the use of network science in the governmental departments [**?**]. Network science is promising to be a fruitful avenue to explore for understanding a variety of issues in healthcare systems as well.

Examining how social networks impact the demand for healthcare. and [**?**].

random field Ising model [**? ?**]

1. reference

Chapter 14

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-)*

Queuing and Traffic Models

14.1 Model Overview

Bank lines, telephones and pool halls. [1]

14.2 Common Uses

Queueing theory is applicable in any situation where one is 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 waitlist 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?* and
- *How does hospital emergency room admittance change throughout the day?*

In regards to capacity of the system, queueing theory in healthcare is often used to study the amount of bed space available in a hospital:

- *How does hospital bed usage vary throughout the day, week, month, or year?*
- *How do different hospital departments occupancies rates interact?*

14.3 Mathematical Details

Both queueing and traffic models address problems where objects or people move through a system. The theory behind these two modelling techniques is essentially the same, and the major difference lies in what outcome one desires to measure. In general, traffic models are focused on congestion in dynamical systems, whereas queueing models focus on understanding wait times within the system.

In either case, the model begins with a system, and a collection of objects which seek to enter and exit the system. In healthcare, two of the best examples are wait lists for surgery and bed counts in hospitals.

In the first (wait lists) 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 second (bed counts) the system one wishes 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 wait list 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 objects and the system interaction is what determines 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.

Queueing theory and traffic theory are two sides of the same coin.

14.3.1 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 which abstracts the idea of the teller that serves that line. In the case of wait list modelling, the server might represent an open operating room slot which can be used to serve one patient from the wait list. In the case of bed count modelling the server might represent a random event which sends a patient home (thus making a hospital bed available for use).

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.

To build a queue 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 often called a A / B / X / Y / Z queue.

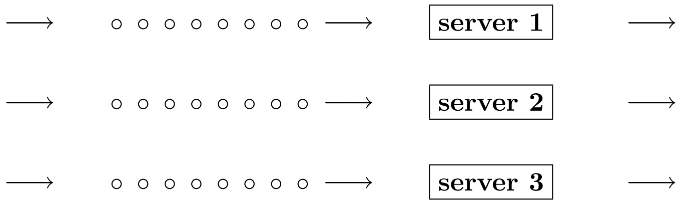
Arrival and service patterns

The arrival pattern is both the rate of how often objects enter the queue. The arrival pattern may be either deterministic or stochastic. If the arrival pattern is stochastic, a probability distributions must also be selected to describe the arrival rate (most commonly this is the poisson distribution, but occasionally others are used).

The service pattern is the rate of objects being removed from the queue. As with the arrival patterns, this may be either deterministic or stochastic, and a variety of probability distributions can be considered.

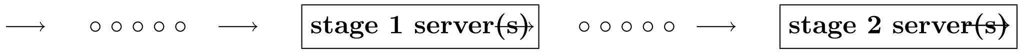
Number of service channels and multiple service stages

The number of service channels is the number of possible ways an object can exit the queue. 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.



If these independent queues are not interacting then each can be studying individually. More realistically, if one has more than one server, the queue will interact in various manners. 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 later 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 following diagram.



In multiple staged 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 server, it may be blocked from entering a later queue.

The most complicated queueing models are multistage queues take the form of a complex network with feedback loops.

MAKE A PICTURE

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 strictly random. (An example of such a queueing model is given in Example 14.4.3.)

System capacity

The system capacity refers to the maximum number of objects allowed in the system (queue). If one in studying concepts of expected wait list length, it is important to define this accurately. Conversely, if one is developing models to analyze how the capacity of the system changes, then one 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 classical 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 extra departure time will be, the doctor does cannot know the departure time of a random patient as they enter the hospital (i.e. before diagnosis occurs). More over, 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 [4]. (Perhaps this is a sign that blood banks are generally under supplied.)

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.

The importance of incorporating impatience into queueing models is discussed in XXX.

14.3.2 Analyzing a queueing model

Having built a queueing model, one now wishes 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 rate of the model is 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.

In the case of deterministic models, reaching an equilibrium state often means that the length of the queue and 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.

Not that not all queues will approach an equilibrium state.

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 of probability distributions see Chapter 5.) Since the term equilibrium is no longer sufficient, this state is usually referred to as a *statistical equilibrium*.

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 developed a closed form analytic solution that describes the equilibrium state of the queue (two examples of this are given in example 14.4.1). In more complicated cases, the queue may be solved via simulation methods (see Chapter **??**). 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.

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 and system capacities will be samplings of a time-independent probability distribution.

14.4 Examples

14.4.1 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 object 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 who's job is to clean dishes.

If one replaces the dirty dishes in this example with occupied hospital beds, and the clean dishes with available hospital beds, then one 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 the hospital cafeteria sells 10,000 meals each day, and 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

$$\begin{aligned} N(t) &= \max\{0, A\left\lfloor \tfrac{t}{24}\right\rfloor - B\left\lfloor t\right\rfloor + N_0\} \\ &= \max\{0, 12000\left\lfloor \tfrac{t}{24}\right\rfloor - 500\left\lfloor t\right\rfloor + N_0\}, \end{aligned}$$

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 which 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 $10,000$ *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 5.3.3, for more information

[1] By artificial, we do not mean that the model is unrealistic, only that the model is calibrated with artificial data.

on the Poisson distribution.) Our new queue can be visualized as follows.

$$\begin{array}{ccccc} \text{Arrival (meal consumed)} & & \text{Queue (dirty dishes)} & & Service(dishwasher) \\ \boxed{\text{Poisson: } \mu_{en} = 10000/24} & \longrightarrow & \circ\circ\circ\circ\circ & \longrightarrow & \boxed{\text{Poisson: } \mu_{ex} = 500} \end{array}$$

(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* is probability from t to $t+1$ can now be computed as follows:

$$\Pr_n(t+1) - \Pr_n(t) = \mu_{en} \Pr_{n-1}(t) + \mu_{ex} \Pr_{n+1}(t) - \mu_{en} \Pr_n(t) - \mu_{ex} \Pr_n(t). \tag{14.1}$$

Equation (14.1) can be interpreted as follows:

$$\begin{array}{rlcl} & \mu_{en} \Pr_{n-1}(t) & \leftrightarrow & \text{the chance a queue of length } n-1 \text{ increases to length } n, \\ + & \mu_{ex} \Pr_{n+1}(t) & \leftrightarrow & \text{plus the chance a queue of length } n+1 \text{ decreases to length } n, \\ - & \mu_{en} \Pr_n(t) & \leftrightarrow & \text{minus the chance a queue of length } n \text{ increases to length } n+1, \\ - & \mu_{ex} \Pr_n(t) & \leftrightarrow & \text{minus the chance a queue of length } n \text{ decreases to length } n-1, \\ \hline & \Pr_n(t+1) - \Pr_n(t) & \leftrightarrow & \text{equals the } \textit{change} \text{ in probability from time } t \text{ to time } t+1. \end{array}$$

The first two terms (of the right hand side of equation (14.1)) 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 (14.1) is replaced by

$$\Pr_0(t+1) - \Pr_0(t) = \mu_{ex} \Pr_1(t) - \mu_{en} \Pr_0(t). \tag{14.2}$$

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 (14.1) and (14.2) reduce to

$$\begin{array}{rcl} \Pr_1 & = & \frac{\mu_{en}}{\mu_{ex}} \Pr_0 \\ \Pr_{n+1} & = & \frac{\mu_{en}+\mu_{ex}}{\mu_{ex}} \Pr_n - \frac{\mu_e n}{\mu ex} \Pr_{n-1} \end{array}$$

This system is solved by the iterative formula

$$\Pr_n = \left(\frac{\mu_{en}}{\mu_{ex}}\right)^n \Pr_0 \, . \tag{14.3}$$

Recalling that the probability of something happening is always 1, we note the sum of all the probabilities $\Pr_n$ must be 1:

$$1 = \sum_{n=0}^{\infty} \underset{n}{\Pr} = \sum_{n=0}^{\infty} \left(\frac{\mu_{en}}{\mu_{ex}}\right)^n \underset{0}{\Pr} = \underset{0}{\Pr} \sum_{n=0}^{\infty} \left(\frac{\mu_{en}}{\mu_{ex}}\right)^n .$$

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

$$\sum_{n=0}^{\infty} \left(\frac{\mu_{en}}{\mu_{ex}}\right)^n = \frac{1}{1 - \frac{\mu_{en}}{\mu_{ex}}},$$

To develop equation (14.3) "correctly," one should actually begin by developing a series of differential equations, and then setting the derivatives to 0. Equations (14.1) and (14.2) provide a discretized interpretation of this mathematical technique.

which tells us $\Pr_0 = 1 - \frac{\mu_{en}}{\mu_{ex}}$, and provides (with equation (14.3)) the statical 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 which 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 which 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.

14.4.2 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, and aging population of Canada, has led to an increased demand for these procedures and, as a direct result, increased wait list length. To understand the attributes which impact wait list length, and to explore potential interventions, a research team at the IRMACS center developed a working queuing theory model for hip and knee replacement surgical waitlists[2]. In this example we outline the model, and provide some of the analytical evaluation of the model.

The basic model consisted 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 to have surgery an individual must wait until they are at the front of the queue, but an individual my drop off the queue at any time. The arrival rate and surgery rate were both assume to be continuous with rates r and s respectively (in this manner the queue can be modelled via differential equations). The drop-off rate was also continuous, but with a rate that varied 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 able to be combined and the linked cases were used for data flow analysis in preparation to the simulation.

Analytical Solution

The queue length N can be mathematically described by the differential equation

$$\frac{dN}{dt} = r - s - kN.$$

This equation states, 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

$$N(t) = \frac{r-s}{k} - \left(\frac{r-s}{k} - N_0\right) e^{-kt}, \tag{14.4}$$

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 wait list, and t_{out} is the time the patient enters surgery. We also define the size of the wait list 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 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

$$\begin{array}{rcl} N(t_{in}) & = & -\frac{s}{k} + \frac{s}{k} e^{-k(t_{in}-t_{out})} \\ N(t_{in}) & = & -\frac{s}{k} + \frac{s}{k} e^{kW} \\ e^{kW} & = & \frac{k}{s}\left(N(t_{in}) + \frac{s}{k}\right) \\ kW & = & \ln\left(\frac{k}{s} N(t_{in}) + 1\right) \\ W & = & \frac{\ln\left(\frac{k}{s} N(t_{in})+1\right)}{k}. \end{array}$$

Thus we have a closed form solution to the expected total wait time for P, provided we know the length of the wait list at the time P enters the queue.

14.4.3 Interrelating Hospital Capacity across Departments

Both from a humanitarian and a business prospective, it is of interest to reduce inefficiencies in the healthcare system. One approach to this is to examine the problem of bed allocation in hospitals. In this example we summarize work of Cochran and Bharti on designing and implementing a queueing theory model for hospital bed allocation[3].

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 which described possible patient flows in the hospital. A (very) simplified version of this network can be found in Figure 14.1.

Figure 14.1: Make this picture complicated enough to look good, but easy enough to follow.

The model was tuned using statistics obtained from economic data stored by the hospital. In particular, they required admittance rates, probability of being a 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 analytically solved. 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 appropiately 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 exponential 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 Queues can be analytically solved to find an exact closed form statistical 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 created and 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 trouble shooting step, and should be preformed 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 which 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 work proceeded to develop optimization strategies for improving hospital efficiency. First they noted that there were large discrepencies in the bed loads between different departments. Using both the analytic and simulation model 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 was 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.

14.5 Related Reading

Queueing theory models have close connections to (XXX list models and chapters of note in the book XXX).

Reference [2], contains the details omitted from Example 14.4.2.

1. "Patient Flow: The new queueing theory for healthcare." Hall, R. W. (2006)
2. Hip and Knee Wait list report
3. Cochran, J.K. and Bharti, A. (2006) A multi-stage stochastic methodology for whole hospital bed planning under peak loading, Int. J. Industrial and Systems Engineering, Vol. 1, Nos. 1/2, pp.836.
4. BLOOD BANKS USE SIRO

Chapter 15

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

15.1 Model Overview

15.2 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 property: the next state of the object is determined by a random process dependent only on the previous state(s) of the object. In health care this assumption is well suited to modeling movement of patients through disease states (see Example 15.4.2). 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 effected by diabetes in future years?* and
- *At what disease state is treatment most suitable to prevent disease spread?*

Aside from modeling disease states, Markov models are also useful for examining if patient history is a factor in behaviour:

- *How does Doctor-Patient loyalty effect usage of the healthcare system?* and
- *To what level do individual's past BMI statuses impact their future BMI status?*

15.3 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 which can take one of two states: healthy or sick. At each time interval, model assigns every object in the system to exactly one state from among a fixed set of states. At the end of each time period, the objects move from one state to the next according to transition probabilities which 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.

15.3.1 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 individuals 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 individuals 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 $S = \{s_1, s_2, ...s_i, ...s_N\}$ be the list of states an object can take. Let X^0 be a column vector of length N which 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

$$P^t = \begin{bmatrix} \Pr^t(1 \to 1) & \Pr^t(2 \to 1) & \dots & \Pr^t(N \to 1) \\ \Pr^t(1 \to 2) & \Pr^t(2 \to 2) & \dots & \Pr^t(N \to 2) \\ \vdots & \vdots & \ddots & \vdots \\ \Pr^t(1 \to N) & \Pr^t(2 \to N) & \dots & \Pr^t(N \to N) \end{bmatrix}$$

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:

$$X^{t+1} = P^t X^t.$$

The matrix P^t is generally referred to as a *transition matrix.*

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 nonnegative. This stops objects from flowing backwards through the model. Second, each column of the transition matrix must sum to 1. This prevents objects from disappearing from the model.[1]

In particular, the state at time t can be found via the formula

$$X^t = P^{t-1} P^{t-2} ... P^1 P^0 X^0. \tag{15.1}$$

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 (15.1).

The matrix P^t is generally referred to as a *transition matrix.*

15.3.2 Higher Order Markov Models

On the surface the Markov assumption appears to create models which are extremely limited in application. For example, if one were modelling the spread of disease through a population, then one would be interested in the two states “uninfected” and “infected.” However, it is well known that patients who have recovered from a virus are unlikely 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 thought 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 to to create new states which incorporate the memory for the desired trait. For example, in the case of modelling spread of disease through a population, 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 which incorporate memory in this type of 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 which 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.

15.3.3 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 15.1.

[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.

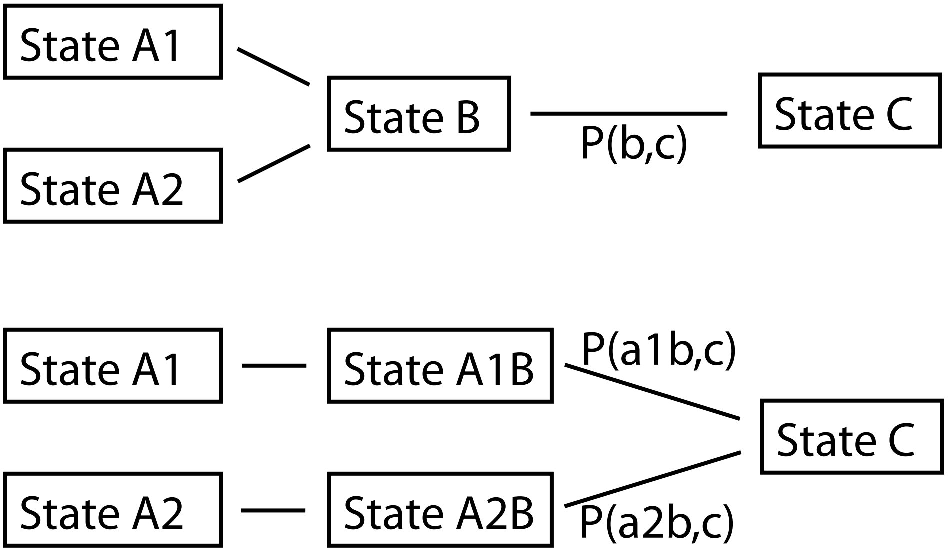


Figure 15.1: Testing the Markov Assumption
One method to test if the Markov holds is to turn a 1^{st} order Markov model into a 2^{nd} order Markov model and check if the transition probabilities are effected. In the 1^{st} order model above (top) we have two states, $A1$ and $A2$ which feed into a state 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.

15.3.4 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 15.3.1). 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 are no longer 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 which has a high probability of being occupied regardless of starting conditions. These techniques are beyond the scope of this book.

15.3.5 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 XXX citation XXX.[2] In these models, time is viewed as a continuous variable, so time steps can occurs 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, as 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.* XXX citation XXX

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 are also being applied to the modelling of chronic diseases, such as HIV. XXX citation XXX

15.4 Examples

15.4.1 A Simple Doctor-Patient Loyalty

To demonstrate the mathematics behind a simple time homogenous Markov chain consider a drop-in-clinic with three doctors. In a drop-in-clinics no appoint is necessary, so patients may not see the same doctor on every visit. However, the patient (when returning to the clinic) may request to 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 15.1. 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 15.1: Transistion probablityble

Suppose the clinic has 300 patients which return on a regular basis, and we wish to see how these patients impact each doctors work load. We begin by assume 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

$$P = \begin{bmatrix}0.72 & 0.09 & 0.21\\0.18 & 0.85 & 0.15\\0.10 & 0.06 & 0.64\end{bmatrix} \quad \text{for all } t.$$

[2] Norbert Wiener and Andrei Kolmogorov in the 1920's and 1930's.

Simple matrix-vector multiplication yields

$$X^1 = \begin{bmatrix}102\\118\\80\end{bmatrix}, \quad X^2 = \begin{bmatrix}100.86\\130.66\\68.48\end{bmatrix}, \quad X^3 = \begin{bmatrix}98.7594\\139.4878\\61.7528\end{bmatrix} \dots \quad X^20 = \begin{bmatrix}89.6613\\158.9361\\51.4026\end{bmatrix}.$$

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:

$$\lim_{t\to\infty} X^t = \lim_{n\to\infty} P \times P \times P \times \ldots \times PX^0 = \lim_{n\to\infty} P^n X^0.$$

Diagonalizing P, we obtain $P = Q\,D\,Q^{-1}$, where

$$Q \approx \begin{bmatrix}-0.583 & -0.867 & -0.481\\ 0.820 & 0.154 & -0.854\\ -0.237 & 0.713 & -0.276\end{bmatrix} \quad \text{and} \quad D \approx \begin{bmatrix}0.680 & 0 & 0\\ 0 & 0.531 & 0\\ 0 & 0 & 1.00\end{bmatrix}.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} P^n X^0 &= \lim_{n\to\infty} QDQ^{-1}QDQ^{-1}\ldots QDQ^{-1}X^0\\ &= \lim_{n\to\infty} QD^nQ^{-1}X^0\\ &= Q\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix} Q^{-1}X^0\\ &\approx \begin{bmatrix}0.30 & 0.30 & 0.30\\ 0.53 & 0.53 & 0.53\\ 0.17 & 0.17 & 0.17\end{bmatrix}\begin{bmatrix}100\\100\\100\end{bmatrix} = \begin{bmatrix}90\\159\\51\end{bmatrix}\end{aligned}$$

Thus, mathematically confirming 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

$$\lim_{n\to\infty} P^n X^0 = \begin{bmatrix}0.30 & 0.30 & 0.30\\ 0.53 & 0.53 & 0.53\\ 0.17 & 0.17 & 0.17\end{bmatrix}\begin{bmatrix}x_1^0\\x_2^0\\x_3^0\end{bmatrix} = \begin{bmatrix}0.30(x_1^0+x_2^0+x_3^0)\\0.53(x_1^0+x_2^0+x_3^0)\\0.17(x_1^0+x_2^0+x_3^0)\end{bmatrix} = \begin{bmatrix}0.30(300)\\0.53(300)\\0.17(300)\end{bmatrix} = \begin{bmatrix}90\\159\\51\end{bmatrix}$$

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 Dr. One, 53% will see Dr. Two, and 17% will see Dr. Three.

15.4.2 Slowing the spread of HIV/AIDS via earlier treatment

Azadeh and Krisz’s happy model.

15.4.3 A Mover-Stayer Model for Problemic 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 change over time by means of equations describing the disease dynamics. A common approach to compartmental modelling is to use a mixed Markov processes, which consist 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[3].

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 consist 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 [**?**], movers can become drug users either by coming contact with other drug users or by contact with drug dealers. The diagram in Figure 15.2 is a compartmental representation of the model. In this figure the straight line represent possible state changes of the movers, and the curves represent the possible interactions between drug users and people susceptible to drug users (movers) in the population.

As shown in the diagram, 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 health care 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 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 15.2 to write a set of eight coupled difference equations that describe the evolution of the system (see Table 15.2). In evolving the system, it is assumed that the lengths of stay in each of the compartments are exponential distributed. To implement the model, a computer program is written to evaluate the time evaluation of the difference equations. Since the computer program evaluates the difference equation using a series of predetermined time steps, this mathematically is a Markov model.

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 problemic 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 [3], values for these parameters are estimated for heroin use in Italy from 1980 to 2000.)

Using this model, the authors explored the effect of both primary and secondary preventive interventions[3]. A primary intervention is one which is applied directly to the susceptible popu-

$$
\begin{aligned}
X(t+\Delta t) &= (1+\mu_{01}-\mu_{10}-\pi_{17})X(t) \\
&\quad - [1-S(t)][\mu_{12}+\nu_{12}Y_1(t)+\nu_{13}Y_2(t) \\
&\quad + \nu_{15}W_1(t)]X(t)+\mu_{61}W_2(t) \\
Y_1(t+\Delta t) &= (1-\mu_{23}-\mu_{26}-\pi_{27})Y_1(t)+[1-S(t)][\mu_{12} \\
&\quad + \nu_{12}Y_1(t)+\nu_{13}Y_2(t)+\nu_{15}W_1(t)]X(t) \\
Y_2(t+\Delta t) &= (1-\mu_{34}-\pi_{37})Y_2(t)+\mu_{32}Y_1(t) \\
Z(t+\Delta t) &= (1-\mu_{45}-\mu_{46}-\pi_{47})Z(t)+\mu_{34}Y_2(t)+\mu_{45}W_1(t) \\
W_1(t+\Delta t) &= (1-\mu_{54}-\mu_{56}-\pi_{57})W_1(t)+[\mu_{65} \\
&\quad + \nu_{26}Y_1(t)+\nu_{36}Y_2(t)+\nu_{56}W_1(t)]+\mu_{45}Z(t) \\
W_2(t+\Delta t) &= (1-\mu_{61}-\pi_{67})W_2(t) \\
&\quad - [\mu_{65}+\nu_{26}Y_1(t)+\nu_{36}Y_2(t)+\nu_{56}W_1(t)]W_2(t) \\
&\quad + \mu_{26}Y_1(t)+\mu_{46}Z(t)+\mu_{56}W_1(t) \\
D(t+\Delta t) &= D(t)+\pi_{27}Y_1(t)+\pi_{37}Y_2(t)+\pi_{47}Z(t)+\pi_{57}W_1(t) \\
&\quad + \pi_{67}W_2(t) \\
S(t+\Delta t) &= S(t)\frac{(1-\mu_{10}-\pi_{17})X(t)}{X(t+\Delta t)}+S_0\frac{\mu_{01}X(t)+\mu_{61}W_2(t)}{X(t+\Delta t)}
\end{aligned}
$$

$X(t)$	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 15.2: Coupled difference equations represented by Figure 15.2.

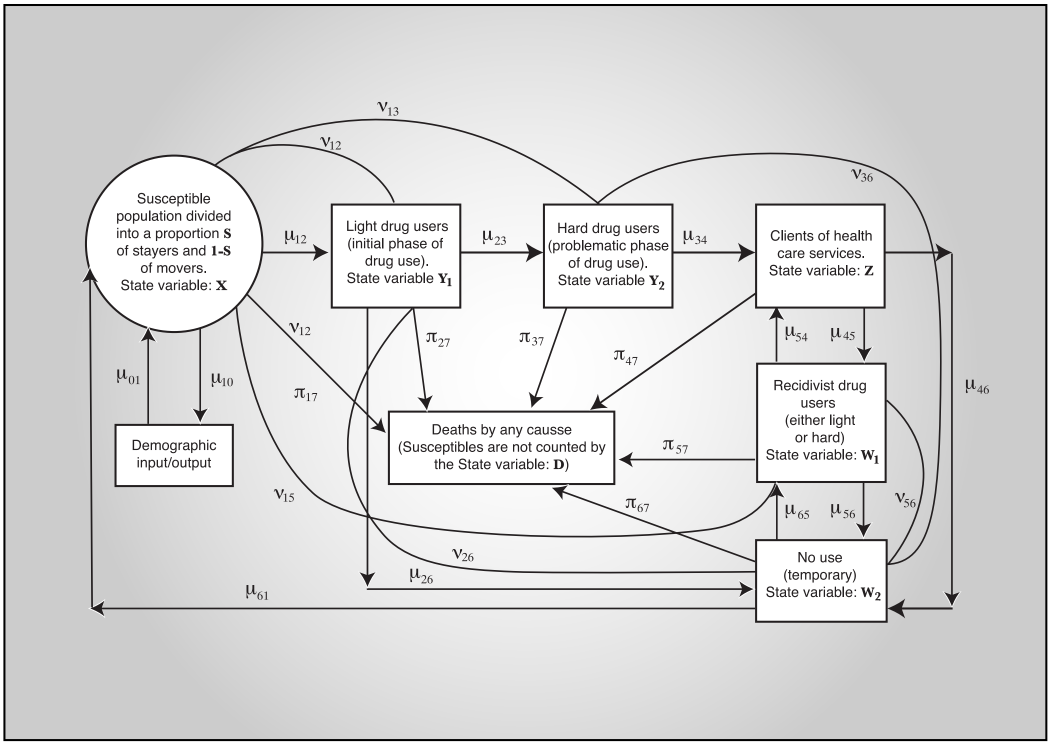

ν_{13}
ν_{12}
ν_{36}
Susceptible population divided into a proportion S of stayers and 1-S of movers. State variable: X
μ_{12}
Light drug users (initial phase of drug use). State variable Y_1
μ_{23}
Hard drug users (problematic phase of drug use). State variable Y_2
μ_{34}
Clients of health care services. State variable: Z
ν_{12}
π_{27}
π_{37}
π_{47}
μ_{54}
μ_{45}
μ_{46}
μ_{01}
μ_{10}
π_{17}
Demographic input/output
Deaths by any causse (Susceptibles are not counted by the State variable: D)
π_{57}
Recidivist drug users (either light or hard) State variable: W_1
ν_{15}
π_{67}
μ_{65}
μ_{56}
ν_{56}
ν_{26}
μ_{26}
μ_{61}
No use (temporary) State variable: W_2

Figure 15.2: 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.

lation. 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 affect on the parameters ν_{56} and μ_{56}. The impact of health care 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 and 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 be about 6 years and possible 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.

15.5 Related Reading

Markov models are closely related to XXX
[1] outlines a Markov model approach to predicting future rates of diabetes.
[2] develops mixed compartmental Markov models.
[3] contains detail of Example 15.4.3

1. honeycutt03

2. Goodman61

3. rossi 04

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

16.1 Model Overview

After modeling, one uses the model to find something out. Sometimes they use it to make predictions, other times they use it to test interventions or possible scenarios. In this later case one often wishes to determine which intervention creates the best outcome. That's optimization!

Some more babble, and then we move on to the stuff below...
XXX

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 computer science the term optimization refers to modifying a piece of software in order to make it (or some part of it) run more efficiently.

16.2 Common Uses

Strictly speaking Optimization is not a modelling technique so much as 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 combination 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?* and
- *How should be schedule nurses to minimize cost given that certain staffing constraints and hospital service constraints must be meet?*

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 minimize transportation times?*
- *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 styles of problems are easily solved by today's optimization tools, and what styles of problems are intractable using todays optimization tools.

16.3 Mathematical Details

Optimization is not so much a modelling technique as a method of examining models to determine what intervention will have the "best " effect. In order to answer this question it is clear that one must begin by defining what is meant by the word best. In order to answer this question via optimization techniques, one 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.*

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 one be interested in maximizing a function,

$$\max\{f(x) : x \in S\},$$

then one can always create a minimization problem by applying the following 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,
> $$\max\{f(x) : x \in S\} = -\min\{-f(x) : x \in S\}.$$

The addition of the constraint set can make similar optimization problems behave very differently. To see this consider the following three problems:

$$\min\{x^2 - x : x \in \mathbb{R}\}, \tag{16.1}$$

$$\min\{x^2 - x : x = -2, -1, 0, 1, \text{or } 2\}, \text{and} \tag{16.2}$$

$$\min\{x^2 - x : x = \frac{m}{3^n} \text{ for some } m, n = 1, 2, ...\}. \tag{16.3}$$

(In problems (16.1), (16.2), and (16.3), we have $S := \mathbb{R}$, $S := \{x : x = \frac{m}{3^n} \text{ for some } m, n = 1, 2, ...\}$, and $S := \{-2, -1, 0, 1, 2\}$ 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 (16.1) is solved at $x = 0.5$ and gives an objective value of -0.25. Problem (16.2) 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 (16.3) we have the strangest situation. By selecting m and n carefully one can construct points with objective function values arbitrary close to -0.25, but one 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.

In order to get around the issues arising in problem (16.3), 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 refereed to as *continuous optimization problems.* Problem (16.1) is a continous 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 (16.2) and (16.3) 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.

16.3.1 Analytically Solvable

Continuous Problems: In the simplest case, $f(x)$ is differentiable and S is $\mathbb{R}$ or a closed interval in $\mathbb{R}$, one 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 one 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 point. If S is a closed interval in $\mathbb{R}$ then one must also remember to check the endpoint of the interval as possible locations for the minimum.

For illustration 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

$$f(0.5) = -0.25, \quad f(1) = 0, \text{ and } f(7) = 42.$$

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 preformed in multiply dimensions. Basically, points are replaced with vectors and derivatives are replaced with gradients. However, in higher dimensions one must be more careful with the edges of the constraint set, as they will not be just two points.

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 one can simply preform 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. However, in practice discrete optimization problems have so many elements in the finite list that an exhaustive search would take years to complete.

16.3.2 Numerical Methods for Continuous Optimization Problems

Suppose now that although the optimization problem is continuous in nature, it is complicated enough to work with the solving the problem analytically is not an option. This might result, from working with a high number of dimensions, the objective function involving integrals, or constraint set taking a complicated form, or the objective function being non-differentiable (amongst many other possible reasons). In this case one must turn to numerical solving methods. Which method to select depends on the form of the problem.

Linear Problems

One of the simplest forms for a continuous optimization problem is that of a *Linear Program*[1]. A linear program is an optimization problem of the form

$$\min\{c^\top x : Ax = b\} \tag{16.4}$$

or

$$\max\{b^\top y : A^\top y \leq c\} \tag{16.5}$$

where b and c are column vectors and A is a matrix ($^\top$ denotes the transposition operation). For example,

$$\min \left\{ 5x_1 + x_3 : \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\},$$

is a linear program with

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad A = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and } c = \begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix}.$$

Mathematically, problem (16.4) is called the *primal problem* and problem (16.5) is called the *dual problem.*

[1] 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.

The interesting thing about linear programs is that if A, b, and c are fixed and the problem (16.4) has a solution then its solution has the same objective function values at problem (16.5). In fact, if x and y can be found where problem (16.4) and problem (16.5) 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 todays computers, linear programs with millions of variables can be solved in just hours. Commercial software for solving linear programs includes XXX, non-commerical software for solving linear programs includes XXX[2].

Semi-definite Problems

The theory and abilities of linear programming can be extended to a much broader class of problems refereed to as *Semi-definite programs*[3]. In semi-definite programming the linear problem (16.4) is generalized in a manner which 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, so 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 this setting. In particular, this allows for easy solving of problems which involve quadratic objective functions or quadratic constraints (along with many other problems).

Commercial software for solving semi-definite programs includes XXX, non-commerical software for solving semi-definite programs includes XXX[4].

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 which 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 $0 \leq \lambda \leq 1$. We shall call the objective function *convex* if by drawing the function and shading in the area above the function one creates a convex set. That is, f is convex if $\{(x,\alpha) : \alpha \geq f(x)\}$ is a convex set. (The set $\{(x,\alpha) : \alpha \geq f(x)\}$ 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 one 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, $\nabla f(x)$, represent the direction which is directly "up-hill" from x. As such, to find the minimum one can repeatedly take steps in the direction $-\nabla f(x)$. This is called *steepest descent*.
>
> **Newton's Method:** If the point x is a minimum for f then necessarily $\nabla f(x) = 0$. If one can find a second derivative of f (or even just approximate a second derivative of f), then one can apply Newton's root finding method to the function $g = \nabla f$.

[2] requires XXX to operate

[3] Like in linear programming, the word *program* is tradition. The phrase *Semi-definite Problem* is perfectly acceptable, but not often used by mathematicians.

[4] requires XXX to operate

Bundle Methods: By finding the function value f and the gradient value ∇f at a point x, one knows roughly what the function looks like very close to x. By using these approximations for a large collection of points, one can create a piecewise linear approximation to the function f which can be easily optimized. By refining the approximation function one can quickly hone in on the minimum of the function.

All of these methods are well studied, and proves 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.

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 which appears to be a minimum but is not. This problem is highlighted in Figure 16.1

XXX

Figure 16.1: In a nonconvex problem one can often locate points which appear to minimize the function but in reality do not.

There are many suggested methods for dealing with this problem, but mathematically it is impossible to guarantee optimality for nonconvex 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 in a good guess at the optimal solution.

Non-differentiable problems

If the function f is not differentiable (or the gradient of f is not readily available) then the above methods cannot be applied. In these cases one usually resorts to a form of *pattern search*. Pattern search is basically a formalized method for trial and error. One begins by selecting a number of points, and determining the objective function value for each. One then uses this information to determine where to generate a new selection of points.

One of the most famous pattern search methods is the *Nelder-Mead* algorithm CITE(1965). More recent pattern search methods include XXX.

16.3.3 Numerical Methods for Discrete Optimization Problems

When decision variables can assume only discrete values from a specified set, the problem is a called a *discrete optimization problem.* When the specified variable set is a set of integers, we deal 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 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 difficultly lay in the fact that one 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 problem. We discuss these next, and then turn our attention to some of the heuristical approaches.

Exact Solutions

One of the most powerful theorems of calculus is the fact that the minimum of a 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 which are actually proven to converge.

The most basic method that must work 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. What this means is 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 just 5, 10, or 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). To do this, the branch and bound algorithm generates subproblems by fixing the values of one variable. Each subproblem is relaxed so that a continuum of elements exist in the constraint set. These relaxed problems can then be solved by one of the above methods for continuous optimization, thus determining a lower bound for the best solution possible with the fixed variable as it is. All subproblems whose lower bound is greater than the best solution found so far may be pruned from the system. In this manner many variables can be solved without solving the original problem, thereby reducing the size of the original problem.

A great deal of research is continued to be generated regarding exact methods for discrete optimization problems. Some commercial software is available, including XXX. Non-commerical software for solving discrete optimization problems includes XXX.

Heuristics

A heuristic is a reproducible method for improving ones knowledge on a problem. In optimization this means an algorithm which, when run on an optimization problem will find a solution no worse than the best known solution. Good heuristical methods are those which 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. The disadvantage of this method is that after reaching a local optimum, local search heuristics get stuck. The stopping criteria for the algorithm is when there is no improving solution in the neighbourhood of the current solution. To overcome being stuck in a local optimum, new techniques have been developed that propose mechanisms for moving out of the local optimum. Some of these modern heuristical methods include tabu search, simulated annealing, evolutionary algorithms (also known as genetic algorithms), and ant colony algorithms.

Tabu Search: Tabu search heuristics differ from other methods .... move out from a local optimum by moving to a new solution in the neighbourhood even if this causes a move to a worse solution. In order to avoid cycling, the previously visited solution and the solutions similar to it are proclaimed tabu for a certain number of iterations.

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.

16.3.4 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 a problem where the problem changes as new data becomes available. At each time step the optimizer must provide a solution to the problem that provides a good level of optimization and allows for flexibility for 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 include the idea of solving a problem over a limited rolling time horizon[**?**].

16.4 Examples

16.4.1 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:

$$\begin{array}{lccccccc}
\text{min} & x_1 & + & x_2 & + & x_3 & & \\
\text{subject to} & x_1 & & & & & \geq & 8 \\
 & & & x_2 & & & \geq & 10 \\
 & & & & & x_3 & \geq & 12
\end{array}$$

Examining this problem it is easy to see that the optimal solution is obtained when $x_1 = 10, x_2 = 8$, 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:

$$\begin{array}{lcccccccccccccl}
\text{min} & x_1 & + & x_2 & + & x_3 & + & x_4 & + & x_5 & + & x_6 & & & \\
\text{subject to} & x_1 & + & & & & & & & & & x_6 & \geq & 4 & \text{(8am – 4pm shift)} \\
 & x_1 & + & x_2 & & & & & & & & & \geq & 8 & \text{(12pm – 8pm shift)} \\
 & & & x_2 & + & x_3 & & & & & & & \geq & 10 & \text{(4pm – 12am shift)} \\
 & & & & & x_3 & + & x_4 & & & & & \geq & 7 & \text{(8pm – 4am shift)} \\
 & & & & & & & x_4 & + & x_5 & & & \geq & 12 & \text{(12am – 8am shift)} \\
 & & & & & & & & & x_5 & + & x_6 & \geq & 4 & \text{(4am – 12pm shift)}
\end{array}$$

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 discrete optimization problem.

16.4.2 Shared Transportation 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 each a healthcare program has been put in place to help disabled and elderly people meet there 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. Assuming that there are n customers, the pickup location of customer i is labelled by i and his/her delivery location is labelled by $n+i$. Let P^+ denote the set of all pickup locations, P^- the set of all delivery locations, and $P = P^+ \cup P^-$. The set V is the set of vehicles. The starting and ending positions of vehicle v are $s(v)$ and $e(v)$, respectively. Starting and ending locations are usually called depots. The set N includes P and all starting and ending vehicle locations. The load at customer i is l_i units. The time window at location i is $[a_i, b_i]$. For each two distinct stop locations, $t_{i,j}$ and $c_{i,j}$ represent direct travel time and travel cost from location i to location j.

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 the integer program given below.

$$\text{minimise} \quad \sum_{v\in V}\sum_{i,j\in N} c_{i,j}X^v_{i,j} \tag{16.6}$$

$$\text{subject to} \quad \sum_{v\in V}\sum_{j\in N} X^v_{i,j} = 1, \qquad i \in P^+ \tag{16.7}$$

$$\sum_{j\in N} X^v_{i,j} - \sum_{j\in N} X^v_{j,i} = 0, \qquad i \in P,\ v \in V \tag{16.8}$$

$$\sum_{j\in P^+} X^v_{s(v),j} = 1, \qquad v \in V \tag{16.9}$$

$$\sum_{i\in P^-} X^v_{i,e(v)} = 1, \qquad v \in V \tag{16.10}$$

$$\sum_{j\in N} X^v_{i,j} - \sum_{j\in N} X^v_{j,n+i} = 0, \qquad i \in P^+,\ v \in V \tag{16.11}$$

$$T^v_i + t^v_{i,n+i} \le T^v_{n+i}, \qquad i \in P^+,\ v \in V \tag{16.12}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ T^v_i + t^v_{i,j} \le T^v_j, \qquad i,j \in P,\ v \in V \tag{16.13}$$

$$X^v_{s(v),j} = 1 \ \Rightarrow\ T^v_{s(v)} + t^v_{s(v),j} \le T^v_j, \qquad j \in P^+,\ v \in V \tag{16.14}$$

$$X^v_{i,e(v)} = 1 \ \Rightarrow\ T^v_i + t^v_{i,e(v)} \le T^v_{e(v)}, \qquad i \in P^-,\ v \in V \tag{16.15}$$

$$a_i \le T^v_i \le b_i, \qquad i \in P,\ v \in V \tag{16.16}$$

$$a_{s(v)} \le T^v_{s(v)} \le b_{s(v)}, \qquad v \in V \tag{16.17}$$

$$a_{e(v)} \le T^v_{e(v)} \le b_{e(v)}, \qquad v \in V \tag{16.18}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ L^v_i + l_j = L^v_j, \qquad i \in P,\ j \in P^+,\ v \in V \tag{16.19}$$

$$X^v_{i,j} = 1 \ \Rightarrow\ L^v_i - l_{j-n} = L^v_j, \qquad i \in P,\ j \in P^-,\ v \in V \tag{16.20}$$

$$X^v_{s(v),j} = 1 \ \Rightarrow\ L^v_{s(v)} + l_j = L^v_j, \qquad j \in P^+,\ v \in V \tag{16.21}$$

$$L^v_{s(v)} = 0, \qquad v \in V \tag{16.22}$$

$$l_i \le L^v_i \le Q^v, \qquad i \in P^+,\ v \in V \tag{16.23}$$

$$X^v_{i,j} \in \{0,1\}, \qquad i,j \in N,\ v \in V \tag{16.24}$$

$$\tag{16.25}$$

Constraints (16.9) and (16.10) assure that each route location starts with a pickup location and ends with a delivery location, not counting depots. Constraint (16.8) means that the number of vehicles coming to location j is equal to the number of vehicles leaving location j. Constraint (16.7) states that each pickup location is left by exactly one vehicle. Constraint (16.11), 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 (16.12), called precedence constraint, assures 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 (16.13)–(16.15) represent compatibility between routes and schedules and constraints (16.16)–(16.18) are time window constraints assuring that each location is served within its own time window. Constraints (16.19)–(16.21) represent compatibility between routes and vehicle capacity and constraints (16.22)–(16.23) are capacity constraints assuring that no vehicle is filled above capacity.

The above problem is extremely complicated, and clearly cannot be solved by hand. Research by XXX has shown that branch and bound methods can been successfully employed to solve this problem[**?**].

16.4.3 Dispatching Ambulance Vehicles

The development of new telecommunication and computer technologies now allows for the collection of real-time data and allows for the solving of dynamic vehicle routing and dispatching problems of practical dimensions. The positions of vehicles are always available through a Geographic Positioning System (GPS) and can be reported on a computerized map managed by a Geographic Information System (GIS).

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 depends on decisions made in solving allocation problems and redeployment problems. The allocation problem consists of determining which ambulance should be sent to answer a call, while the redeployment problem consists of relocating available ambulances to potential location sites when calls are received.

This problem is studied in [**?**]. In this example we provide a broad stroke overview of this work.

The redeployment problem differs from the standard ambulance location problem in several respects. While location problems are usually solved at the strategic level, redeployment problems are operational and are solved dynamically in real time, as the emergency medical service managers must make almost instantaneous and simultaneous decisions regarding allocation and redeployment. Some of the problem constraints include:

- a limited number of ambulances can be positioned at each site;
- only a limited number of ambulances can be moved when a redeployment occurs;
- vehicles moved in successive redeployments cannot be always 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;
- at the end of a shift, the ambulance has to be moved closer to the central service point where the vehicles are based; and
- the breaks of paramedic teams have to be taken into account.

This problem has been solved by 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 confirm the efficiency of the proposed approach. XXX this example needs considerable more flesh XXX

16.5 Related Reading

XXX

Appendix A

Glossary

Agent-based simulation
Models behaviours specific to individuals or agents rather than the average behaviour of a population or system.

Algorithm
A sequence of steps for solving a problem.

Analytical method
Solution method derived from pure or applied mathematics used for mathematical models.

Attractors
A set of physical properties to which a dynamical system tends to evolve.

Attributable risk
The fraction of the incidence of a disease in the population that can be attributed to a specific risk factor. Also called the aetiological fraction.

Bayes' theorem
A theorem in statistics which relates conditional probabilities to marginal probabilities. Specifically, it states that

$$\frac{\Pr(A \mid B)}{\Pr(B \mid A)} = \frac{\Pr(A)}{\Pr(B)}\,.$$

Boolean logic
A branch of mathematics in which variables may take only one of two possible values: true or false, which are sometimes also denoted as 1 or 0. The primary operations of Boolean logic are AND, OR, and NOT.

Branch and bound method
An exact method for solving a discrete optimisation problem.

Branch and cut method
An exact method for solving a discrete optimisation problem.

Case-control study
This is a type of epidemiological study in which suspected aetiological factors in the history of patients with a disease is compared to the history of control patients, who do not have the disease.

Catastrophe theory
Deals with those dynamical systems that respond with sudden, large changes in dynamics to small changes in the state of the system.

Causality
The relationship of causes to effects they produce.

Centipede game
A game in which two players take turns choosing either taking part in a slightly larger share of a slowly increasing pot, or passing the pot to the other player.

Chaos
The unpredictable behaviour of deterministic and often nonlinear, dynamical systems.

Chronic disease
A disease that develops and lasts over a long period of time.

Cohort study
A cohort study is a study which collects data about a given set of individuals over a period of time.

Communicable disease
A disease that is contagious and which can be transmitted from one source to another by infectious bacteria or viral organisms.

Compartmental model
A mathematical model used in epidemiology, which divides hosts into different compartments, according to their infectious state.

Complex system
Systems consisting usually of a large number of components, that are interrelated through non-linear relationships.

Complexity theory
An approach to understanding the behaviour of systems that exhibit non-linear dynamics, or novel, unexpected behaviours produced by adaptive systems.

Conceptual model
A mental image of an object, system or process that describes the functional relationships among components.

Conditional probability
The probability that an event happens, given the occurrence of another event.

Congestion
The state of a transmission system when a constraint on the system's transfer capacity is reached so that no further transactions can be carried out.

Constraint

A binding limit which is expressed as an equality or inequality within an optimisation problem to define the set of allowed values for a given variable.

Consumer theory

A theory of economics, which relates preferences, indifference curves and budget constraints.

Count data

Data representing counts or the number of observations in each category.

Decision variable

A variable within an optimisation problem. Different values of a decision variable define different solutions to the optimisation problem.

Deterministic

Having no random or probabilistic aspects, but proceeds in a fixed and predictable fashion.

Difference equation

An equation that describes how something changes in discrete time steps.

Differential equation

The mathematical formulation corresponding to a continuous model, involving derivatives.

Discrete event simulation

A method of simulation based on discrete time-steps used to observe the dynamic behaviour of systems.

Discrete variable

A variable that takes values in a discrete set such as the integers.

Dynamic

A theory or model that describes the time-dependent effects of forces on a system.

Dynamical system

A mathematical model describing the changing state of a system.

Econometrics

The application of statistical and mathematical methods in the field of economics to describe the numerical relationships between key economic forces such as capital, interest rates, and labour.

Econophysics

The application of computational methods from theoretical physics in economics. This approach is mostly commonly used for models with strongly interacting agents.

Eilenberg-MacLane space

An Eilenberg-MacLane space $K(G,n)$ is a topological space whose homotopy groups are $\pi_n(K(G,n)) = G$ and $\pi_i(K(G,n)) = 0$, for $i \neq n$.

Elective surgery

Any surgery that is not considered as an emergency case.

Endogenous variable

A variable which is determined by the internal dynamics within the model. This terminology is most common in the economics modelling literature.

Epidemiology

The scientific study of factors affecting the health and illness of individuals and populations.

Equilibrium state

A stable balanced situation where forces cancel out each other.

Event independence

Events are independent if the probability that one event occurs is not influenced by the occurrence of the other event.

Exact methods

Analytical or numerical methods for solving equations that produce an exact solution rather than an approximation.

Exclusion process

A non-equilibrium model in statistical physics describing the movement of particles as they hop along on a lattice of discrete sites, which has been applied to the study of congestion in transportation and other fields.

Exogenous variable

A variable in a model which is determined externally to the model. Such variable are usually determined through data analysis and then inputted into the model. This terminology is most common in the economics modelling literature.

Fecundity index

A measure of the birthrate in a country. It is defined as the annual number of births per woman of child-bearing age.

Feedback loop

A process whereby some proportion of the output signal of a system is passed (fed back) to the input.

Fuzzy logic

A mathematical discipline that deals with fuzzy sets and operations defined on fuzzy sets.

Game theory

A branch of applied mathematics that studies strategic situations where players choose different actions in an attempt to maximise their returns.

Generalised method of moments

The classical method of moments estimates the probability distribution function for a sample by equating the first m moments of the probability distribution function with the moments calculated from the data. The generalised method of moments is an extension of this method which incorporates conditions on the moments that follow from the statistical model.

Graph theory

A branch of mathematics that studies graphs and networks.

Group theory

A branch of mathematics dealing with symmetries and how they are classified in terms of objects called groups.

Hamiltonian cycle

A cycle in a graph that visits each vertex exactly once.

Health belief model

This is a psychosocial model of the role played by beliefs, motivations, and perceptions in an individuals decision to seek healthcare.

Health care burden

The impact of a disease or condition on the health care system.

Health care system

The organisation by which health care is provided.

Health care utilisation

The amount and rate at which patients/consumers use health care services.

Heuristics

A inexact solution method that seeks to finds a good solution for an optimisation problem. Efficient heuristics are able to find solutions close to optimal.

Homotopy group

The homotopy group, $\pi_i(X)$, of a topological space, X, is a mathematical group constructed from all maps from the i-dimensional sphere to X.

Household production theory

A theory suggesting that households purchase market goods and use them in household production processes that generate utility.

Human capital theory

A theory of the act of investment in human resources in the form of training and education with a view on raising the productivity of an individual.

Hurdle models

Statistical models that divide the response variable into classes of 0 and non-zero values and then predict the probability of 0 data and analyse the non-zero data separately.

Incidence

In epidemiology, incidence refers to the number of new events, such as cases of a disease or accidents, that occur in a population during a specified period of time.

Intelligent agent

A software agent that exhibits some form of artificial intelligence.

Lagrange multiplier method

A method for calculating local extrema for functions in the presence of constraints. The Lagrange multipliers are additional parameters incorporated into the problem to handle the constraints. In economics, Lagrange multipliers are often called shadow prices.

Latent class models

A statistical model representing subtypes of related cases, also called latent classes, in multivariate categorical data.

Linear

A model or a function where the input and output are proportional.

Linear programming

A branch of mathematics that uses linear inequalities to solve decision-making problems involving maximums and minimums.

Linear programming problem

An optimisation problem whose objective function and constraints are linear.

Long-term care

Medical, social, and personal care services, such as nursing home care, home and community based care, hospice care, or respite care, required over a long period of time by a person with a chronic illness or disability.

Management science

The discipline of using mathematics, and other analytical methods, to help make better business decisions.

Markov chain

A stochastic process with a finite number of states in which the probability of occurrence of a future state is conditional only upon the current state.

Mathematical model

A collection of equations describing the measurable quantities in a particular setting (both constants and variables) and their interrelationships.

Mean field theory

An approximation used in theoretical physics in which the behaviour of a complicated dynamical system is approximated by the average or mean behaviour of the system.

Model

An abstract or conceptual object used in the creation of a predictive formula.

Model boundary

The boundary of a model is the point or points beyond which no values or processes are considered or defined.

Model implementation

The process of generating output from input entered into a mathematical or computation representation of a conceptual model.

Moral hazard problem

In insurance theory, the moral hazard problem refers to an unintended increase in the risk of "immoral" behaviour resulting from an insurance contract. For example, the moral hazard problem of health insurance is that if insurance decreases the cost of medical care, an increase in unnecessary use of medical services may result.

Mover-stayer model

An extension of Markov chain models for dealing with two specific classes of unobserved heterogeneity in the population, including *movers* that follow a Markov process of change and the *stayers* with a probability of change of 0.

Negative binomial distribution

A distribution which is parameterised by a mean m and an aggregation parameter k which is large when aggregation is small.

Network

A set of nodes or vertices, connected by edges.

Network theory

A branch of applied mathematics and physics, involving the study of graphs as a representation of either symmetric relations or, more generally, of asymmetric relations between discrete objects.

Numerical method

Iterative methods of solving problems on a computer, which may have an analytical basis or may involve heuristics.

Object-oriented programming

A computer programming paradigm widely used in software development which views programs as a collection of individual units or objects that act on each other, rather than as a set of functions or instructions.

Objective function

The function that is to be optimised in an optimisation problem.

Operations research

The discipline of applying mathematical methods to applied optimisation problems. It has wide applications to decision-making in business.

Optimal value

This is either the maximal or minimal value of the objective function, in a maximisation or minimisation problem, respectively.

Optimal solution

The solution for which the objective function reaches the optimal value.

Optimisation

A process that searches for the optimal solution to a model or problem.

Optimisation problem

The problem of finding among all feasible solutions the best one according to some criterion.

Panal data

Panel data consists of a series of cross-sectional samplings of a population, or multiple phenomena, at a sequence of time intervals. Such a data set may be termed two-dimensional, because it has extent in both the cross-sectional direction and the time direction.

Parameter

A value which is usually unknown and has to be estimated to represent a certain population characteristic.

Poisson distribution

A one-parameter, discrete frequency distribution giving the probability that n events will occur in an interval x, provided that these events are individually independent and that the number occurring in a subinterval does not influence the number occurring in any other non-overlapping subinterval

Polytope

The generalisation to any dimension of a polygon in two dimensions, and a polyhedron in three dimensions.

Population health

An approach to health that aims to improve the health of the population as a whole.

Potential impact fraction

This is the proportional reduction in the number of incident cases of a disease in a population, resulting from a change in the distribution of a risk factor. It is also called the generalised impact fraction.

Primary health care

Essential health care made accessible at a cost which the country and community can afford, with methods that are practical, scientifically sound and socially acceptable.

Principal-agent problem

The principal-agent problem occurs when an agent is acting on behalf of a principal, in a situation where the agent possesses an information advantage over the principal. The issue is how to ensure that the agent acts in the best of interests of the principal. In healthcare, a potential principal-agent problem occurs in the patient-physician relationship, with the patient as the principal and the physician as the agent.

Prisoner's dilemma

A type of non-zero-sum game in which two players try to get rewards from a banker by cooperating with or betraying the other player.

Probability distribution

The mathematical description of a random variable in terms of its admissible values and the probability associated, in an appropriate sense, with each value.

PYLG

An acronym for the potential years of life gained by a death avoided.

PYLL

An acronym for the potential years of life lost resulting from a death.

QALY

An acronym for the quality adjusted life years, which is a measure based on the PYLG times a measure of health status during the years gained.

Queueing theory

The theoretical study of waiting lines, expressed in mathematical terms, including components such as the number of waiting lines, number of servers, average wait time, number of queues or lines, and either increasing or decreasing probabilities of queue times.

Random field Ising model

A model from physics that was originally developed to explain certain properties of magnetism. It has since been applied to modelling the effect of social pressure on decision-making.

Random variable

A variable characterised by random behaviour in assuming its different possible values.

Regression

A form of statistical analysis assessing the association between two variables.

Relative risk

In epidemiological risk analysis, the ratio of disease incidence among those exposed to the risk to the incidence among those not exposed. Also called the incidence density ration.

Representative agent approximation

This approximation is commonly used in economic models. It assumes that the macro-economic behaviour is approximated by a single agent, whose utility function encapsulated all of the micro-economic behaviour in the system. Although widely used, this approximation may not accurately reflect the system, especially when there are strong interactions between agents or when the system is strongly heterogeneous.

Response latency

The measure of time elapsing between the onset of a stimulus and the beginning of the response to it.

Self-organisation

The ability of certain non-equilibrium systems to develop structures and patterns in the absence of external control or manipulation.

Sensitivity testing

A testing method for estimating continuous parameters that can not be measured in practice.

Shadow price

For a constrained optimisation problem in economics, the shadow price is the rate at which the optimal value of the objective functions changes under relaxation of the constraint. Mathematically, this corresponds to the Lagrange multiplier associated with the constraint. The shadow price is also called the marginal value.

Simplex method

A general technique for solving linear programming problems by an iterative process.

Simulation

The use of a mathematical model to recreate a situation, often repeatedly, so that the likelihood of various outcomes can be estimated.

Social capital

This is the capacity of individuals to command scarce resources by virtue of their membership in networks or broader social structures.

Social networks

A map of the relationships between individuals, indicating the ways in which they are connected through various social familiarities.

Soft operations research

A qualitative method related to system dynamics for thinking about complex systems.

Social organisation strategy

This is a framework in sociology which uses a dynamic social network approach to understanding human behaviour.

Solution space

This is the set of all solutions to a given mathematical problem.

Static

The opposite of dynamical; not changing.

Statistical association

The effect arising from the relationship between variables but which does not imply causation.

Statistical model

A statistical model is a tree-structured model format that contains and persists arbitrary statistical data.

Stochastic

A process with an indeterminate or random element as opposed to a deterministic process that has no random element.

Strange attractor

An attracting subset of the solution space of a dynamical system, which has zero volume, but fractal dimension greater than 0.

Strategic planning

The process of developing long range strategies to reach a defined objective.

Symmetry

An attribute of a shape or relation characterised by the exact reflection of form on opposite sides of a dividing line or plane.

Synergetics

An interdisciplinary science explaining the formation of patterns and structures in natural and social systems.

System

An assemblage of elements comprising a whole with each element related to other elements.

System dynamics

A field of study that includes a methodology for constructing computer simulation models to achieve better understanding of social and corporate systems.

System thinking

A school of thought that focuses on recognising the interconnections between the parts of a system and synthesising them into a unified view of the whole.

Tabu search

A heuristic method used in operations research. It is a local search method that uses memory structures to improve efficiency.

Tactical planning

The process of developing medium-term strategies.

Tertiary medical care

Medical and related services provided at specialist hospitals or regional centres equipped with advanced diagnostic and treatment facilities.

Travelling salesman problem

This is a combinatorial optimisation problem, in which a "travelling salesman" must find the shortest route which visits each city in a given set of cities exactly once.

Utility function

A mathematical expression that assigns a value to all possible choices.

Validation

Verification that something is correct or conforms to a certain standard.

Variance

A measure of the average distance between each of a set of data points and their mean value; equal to the sum of the squares of the deviation from the mean value.

Wait list

A list of patients who are waiting for a medical procedure.

Appendix B

Sources of Healthcare Data

The focus of this chapter is to review a number of large-scale data projects that have been or can be used for mathematical modelling. Two of the projects considered — POHEM and ARCHIMEDES — are integrated with microsimulation software. A third, MORBIDAT, is an extensive Belgian database system, which is a good example of a large repository of health and disease information. Plans for broad implementation of electronic health records in Canada offer an unprecedented opportunity for obtaining health information, and it is included in this review for this reason. The final section describes data sources from surveys that have been used for health care studies, particularly for the study of health care utilisation and demand.

B.1 The POHEM Population Health Model

Statistics Canada assembles large linked socio-economic and health data sets. These can be analysed to assess the health status of the population [**?**]. To complement its existing programme, Statistics Canada developed a Population Health Model (POHEM) which interfaces the data with a microsimulation component.

An important motivation behind the POHEM project was a need for complete and comprehensive information on health outcomes. While available information on resource utilisation was deemed adequate, data on individual and population health status was lacking or considered insufficient. Illness care and medical interventions were heavily emphasised in the data, and this did not reflect the growing trend in public health approaches in health care, which focus primarily on the determinants of health and involve disease prevention and health promotion interventions [**?**]. Furthermore, a general lack of standard definitions for components across data sets impeded coherent analyses and interpretations of the available information.

Although POHEM is considered to be a model, it represents a different approach from those discussed so far in this report. The method bears some resemblance to risk modelling but it is essentially a "modelless" simulation, based on data, and its primary output is synthetic data. In other words, POHEM, and similar microsimulation methods, represent models of *data*. POHEM does not specify the structure of the system explicitly, nor does it provide insight into the mechanics of the system. In contrast, system models discussed in other parts of this report are implementations of conceptual models for the study of principles governing the behaviour of complex systems. They rely on high quality data for calibration, validation and forecasting but the basic model structure is not rooted in data.

One of POHEM's main objectives is to generate longitudinal data sets and thereby fill in gaps in existing data. Simulations can also be used to evaluate alternate policy scenarios, to assess the impact of public health interventions or the burden of disease, and to make cost projections for treatment options or longterm management of disease.

POHEM has proven to be a useful tool to policy makers. There is a growing body of literature on applying POHEM to evaluate population health and health care issues. The integration of data with microsimulation has also resulted in improved data quality. POHEM and other similar projects showcase the importance and potential utility of extensive, high quality data in informing health policy decisions.

The microsimulation component of POHEM is built on a higher-order Markov process, which incorporates continuous time in the simulation. The virtual, synthetic, longitudinal data produced represents the full life cycle of a birth cohort. Synthetic individuals are generated and allowed to age using a Monte Carlo simulation. This is achieved using a random number generator to simulate specific events through a comparison of the results of random draws to known distributions. The process integrates information on demographics, risk factors, disease onset and progression, resource utilisation, and cost of care, for example, to create cohorts that match real counterparts in detail. The resulting virtual population data have been analysed primarily to evaluate cost-effectiveness of intervention alternatives and to study population-level health outcomes [**?**].

The last published use of POHEM involved an evaluation of the national colorectal screening program which informed policy decisions by the National Committee on Colorectal Cancer Screening. As of September 2006, developmental work is underway for an osteoarthritis model and a diabetes model is in its planning stages. A less detailed but more comprehensive model for estimating and projecting the population health impacts of disease, injury and risk factors is in its beginning to be developed in the program of The Population Health Impact of Disease in Canada. This model will incorporate much less detail on individual disease but includes many more diseases to study overall population impact. The project is an extension of WHO research on the burden of disease [**?**].

The following overview provides representative examples involving lung cancer, breast cancer and colorectal cancer, where POHEM has been employed.

B.1.1 Lung Cancer

Several studies investigated cost and treatment effectiveness for lung cancer. In one of these, cost effectiveness of alternative chemotherapies was evaluated [**?**]. Stage-specific survival data was extracted from clinical trials and incorporated into POHEM. Various chemotherapy interventions were simulated and cost-analysis was carried out on the simulated results. The cost of treatment for toxicity, in addition to the actual cost of the treatment, was incorporated in the analyses to provide a more comprehensive estimate of costs. The therapy options could then be ranked based on cost-effectiveness.

In another model a hypothetical cohort of people with demographic and labour force characteristics, risk factor exposures, and health histories typical of Canadians was generated. The lung cancer sub-model assigned individuals to a particular histological cell type and stage based on data collected from the Canadian Cancer Registry, the Alberta Cancer Board and the Ontario Cancer Registry. Treatment, disease progression, and survival characteristics were assigned to the cohort based on clinical data, information from national physician surveys and expert opinion. Finally, costs were allocated to the various components of care appropriate for cell type and stage of disease, from the time of initial diagnosis to terminal care. Direct medical cost associated with chemotherapeutic treatment of metastatic non-small-cell lung cancer (NSCLC) was assessed and compared to the costs of best supportive care from the perspective of a provincial government payer in a universal health care system [**?**].

A different study of lung cancer [**?**] used the distribution of cancer stage at diagnosis for the reference year of 1992 in the POHEM model to demonstrate that lung cancer treatment in Canada is effective from an economic point of view.

B.1.2 Breast Cancer

In a model of breast cancer, the potential savings in the acute care setting were evaluated for a strategy to reduce length of hospital stay (LOS) for breast cancer surgery. The strategy combines reduced LOS with increased investment in the home care setting [**?**]. Again, stage of breast cancer, treatment type, disease progression and survival statistics were assigned to simulated individuals. Appropriate Canadian costs were incorporated in the model. The reduction in hospitalisation costs was estimated to be about 30%. However, the authors point out that the cost-analysis does not consider any change in the quality of care for patients in such an eary-discharge program, implying that results of the model – and other similar models – have to be considered in a broader context in reaching policy decisions.

Healthcare costs associated with lifetime management of breast cancer were estimated in a model of a cohort of over 170,000 women who were diagnosed with breast cancer in 1995 [**?**]. Another study using POHEM evaluated a strategy of administering preventive Tamoxifen therapy to women at high risk for breast cancer [**?**]. The model, which incorporated data on breast cancer prevention trials, was used to show that no overall population health benefit would be gained from preventative Tamoxifen therapy.

B.1.3 Colorectal Cancer

Colorectal cancer is the second leading cause of cancer-related mortality among Canadians. A Canadian colorectal cancer (CRC) model of incidence and progression was implemented in POHEM in 2000, including disease incidence by age, sex, and site (colon or rectum), disease progression to local recurrence, metastasis, death, treatment options and cost [**?**]. POHEM was used to simulate a screening programme with two identical cohorts, including a reference cohort and a screen cohort. Life histories for the cohorts were identical. The screen cohort was subjected to the modelled screening protocol and the impact of screening was evaluated by comparing outcomes for the two cohorts. The model results showed that biennial screening of 67% of individuals aged 50 to 74 in the year 2000 resulted in an estimated 10-year CRC mortality reduction of 16.7%.

In another study direct lifetime costs were estimated for colorectal cancer (CRC) using POHEM [**?**]. The POHEM model in this study was calibrated using Statistics Canada's person-oriented hospital discharge database and the Canadian Cancer Registry, along with some provincial and regional databases. POHEM was used to simulate the disease progression in a year 2000 cohort of patients newly diagnosed with colorectal cancer. The result of the simulation are shown in Table B.1.

Disease Site	**Stage at Diagnosis**	**Number with Diagnosis (2000)**	**Percentage Developing Local Recurrence over Lifetime**	**Percentage Developing Metastatic Recurrence over Lifetime**	**Percentage Dying of CRC within 5 years**	**Percentage Dying of CRC**	**Average Number of Years Alive After Diagnosis**
colon	I	988 (9%)	9.0	17.5	12.5	20.4	12.9
	II	3966 (35%)	14.0	23.7	22.6	28.8	10.7
	III	2970 (26%)	9.7	31.3	34.3	41.4	9.2
	IV	3506 (31%)	N/A	100	93.9	94.4	1.2
	all stages	11430 (100%)	8.2	48.6	46.7	51.6	7.6
rectal	I	1243 (23%)	29.1	23.5	18.7	33.8	11.1
	II	1592 (29%)	9.5	45.2	36.9	45.5	9.3
	III	1581 (29%)	9.4	52.1	53.2	62.3	6.8
	IV	1010 (19%)	N/A	100	94.9	95.2	1.2
	all stages	5426 (100%)	12.2	52.5	48.4	57.0	7.4

Table B.1: The results of a POHEM simulation applied to an initial year 2000 cohort of 16856 patients with colorectal cancer [**?**]. The result is a simulated lifetime disease evolution for the cohort.

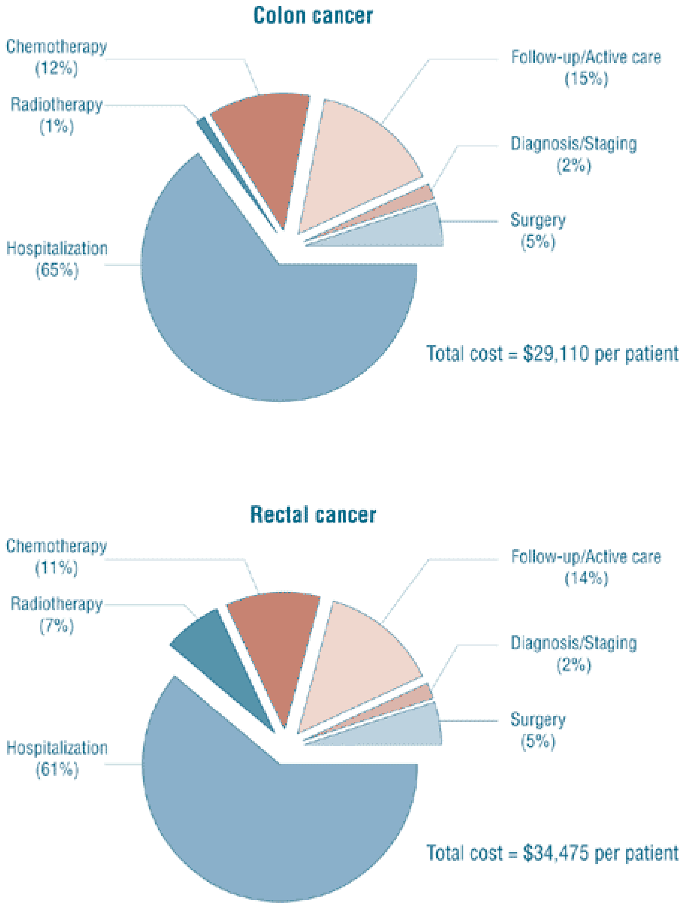


Figure B.1: The POHEM Simulation of the Average Breakdown of Future Lifetime Healthcare Costs for Patients Diagnosed with Colorectal Cancer. Note that this simulation carried out by **?**] predicts that hospitalization would account for 65% of the cost for colon cancer patients and 61% of the cost for rectal cancer patients.
Reproduced from **?**].

The POHEM simulation of the future disease course for patients with colorectal cancer may be used to determine the future demands that would be placed on the healthcare system by these patients. As an example, the breakdown of average lifetime healthcare costs for patients diagnised with colorectal cancer is given in Figure B.1.

B.2 ARCHIMEDES

ARCHIMEDES is a proprietary stochastic microsimulation model developed by the US company Kaiser Permanente [?]. It is used to address complex questions in health care. The model is a discrete event simulator that creates detailed virtual models of the physiology of diseases, patients and and health care systems. As POHEM, ARCHIMEDES simulates individual life histories rather than modelling dynamical systems of populations, but unlike POHEM, ARCHIMEDES incorporates an extraordinarily high number of variables at all levels and uses a Monte Carlo method to assign events to individuals. The large and complex model is accessible through the consulting services offered by the company.

The Diabetes PHD (Personal Health Decisions) risk assessment tool is available online from the American Diabetes Association and demonstrates the level of detail used in the model [?].

One of the main uses of Archimedes is in conducting simulated clinical trials. David Eddy, founder of the Archimedes company, has evaluated the model by comparing 74 actual and simulated trials. In 71 out of the 74 cases no statistically significant differences were found between the results of the clinical trials and their simulated counterparts [?].

B.3 MORBIDAT

MORBIDAT[1] is a sophisticated database system of morbidity and health-related behaviours and the corresponding regulations in Belgium. It is a product developed, in collaboration with the Flemish and the French Communities, by the Scientific Institute of Public Health's Unit of Epidemiology at the Centre of Operational Research of Public Health (CORPH). The main objective of CORPH is to optimise the management of health information to evaluate and to follow the state of health of the population. In addition to the MORBIDAT project, CORPH is responsible for collection of vital statistics, gathering information on health indicators, carrying out health interview surveys and integrating the informations in review documents to support public health policy development.

Currently MORBIDAT consists of three inventories

- the Inventory of Existing Databases on Morbidity and Disability;
- the Inventory of Existing Databases on Lifestyle Variables;
- the Database of Morbidity Situation Sketches for Several Disorders Important from a Public Health Perspective.

For the **Inventory of Morbidity Databases**, a list of the main morbidity categories was established based on the WHO *Health for All in the Year 2000* campaign. There are 17 categories of disorders or events, including

1. The elderly
2. Occupational diseases
3. Cardio-vascular disorders
4. Digestive disorders
5. Endocrinological disorders
6. Infectious diseases
7. Disabilities
8. Cancer
9. Mental disorders
10. Mother-child pathology

[1] The MORBIDAT website [?] has an introduction in English. The rest of the information in this section was translated from Dutch by Alexa van der Waall (IRMACS, SFU).

11. Oral and dental disorders
12. Primary morbidity
13. Musculoskeletal disorders
14. Accidents
15. Eye disorders
16. Respiratory disorders
17. Urological disorders

Each category contains anonymous individual data on patients and results of various analyses. For example, the category of endocrinological disorders currently contains the following registries and programs: Diabetes (Belgian Diabetes Registry; Epidemiological Survey of Medically Treated Diabetes Melitus; Registry of Diabetes Cases (from GPs)); Pathology of the Thyroid Gland (Pathology of the Thyroid Gland (from GPs)). For each program, additional information is recorded (such as contact details for administrators, objectives and registered parameter, and available data and/or publications).

The **Inventory of Lifestyle Databases** includes 11 categories, also based on the WHO *Health for All by the Year 2000* strategy. It is organised similarly to the Inventory of Morbidity Databases and its categories are:

1. Tobacco use
2. Alcohol abuse
3. Illegal drug abuse
4. Abuse of prescription drugs
5. Nutrition
6. Physical activity
7. Leisure activity
8. Social behaviour
9. Sexual behaviour
10. Health supporting knowledge and activities
11. Work environment

The **Database of Morbidity Situation Sketches for Several Disorders** of MORBIDAT is devoted to describing current knowledge on a selection of important disorders. Overviews are provided with short summaries, a list of the determinants and an epidemiological description with international comparisons. The overview includes prevention options and statistical data when available. Complete sketches are available for the following disorders:

1. Asthma and air pollution
2. Screening of breast cancer in women between 40 and 49 years of age
3. Cerebro-vascular disorders
4. Cervical cancer
5. Chronic aspecific respiratory disorders
6. Colon- and rectum cancer
7. Dementia
8. Depression
9. Diabetes mellitus
10. Ischemic heart disease
11. Cancer – general
12. Lung cancer
13. Accidents in the home
14. Osteoporosis and hip fractures
15. Prostate cancer
16. Tuberculosis
17. Traffic accidents
18. Suicide attempts

Information in this database is collected from existing Belgian and international sources. Birth and death information is supplied by three main registries: The national Institute for Statistics (NIS), The Ministry of the Flemish community, and the Health Observatory of Brussels Gewest.

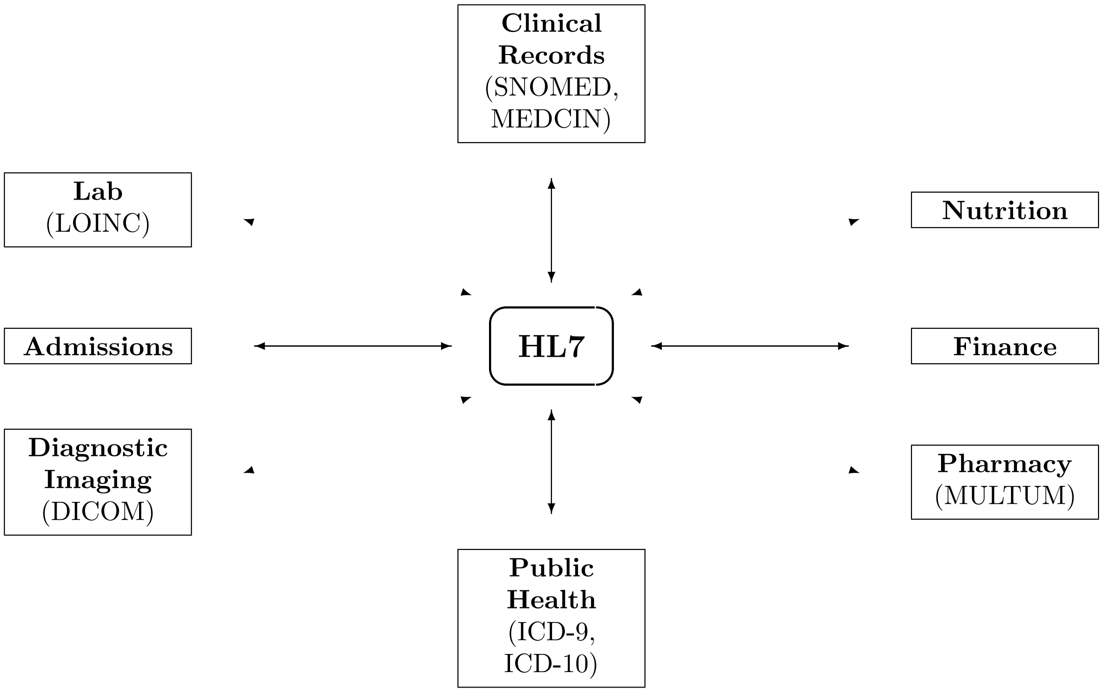

Clinical Records (SNOMED, MEDCIN)
Lab (LOINC)
Nutrition
Admissions
HL7
Finance
Diagnostic Imaging (DICOM)
Pharmacy (MULTUM)
Public Health (ICD-9, ICD-10)

Figure B.2: The Structure of an Electronic Health Record system
The HL7 suite of protocols provides a common interface for all components in the system.

B.4 Electronic Health Records

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

[

8]quotecolour3.6inIn attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purpose of comparison. If they could be obtained, they would enable us to decide many other questions...
— *Florence Nightingale* — Health Level 7 (HL7) is the broadest EHR standard under development and plans are in place to implement this standard in British Columbia. The HL7 standard has been accredited by the *American National Standards Institute* (ANSI). “Level seven” refers to the highest level communications model defined by the *International Organisation for Standardisation* (ISO).

There are a number of other EHR standards, which address the specialised needs of a specific department. Many of these standards predate HL7. Examples of such standards include *Digital*

Imaging and Communications in Medicine (DICOM) for diagnostic imaging, *Systematised Nomenclature of Medicine* (SNOMED) for clinical records, and *Logical Observation Identifiers Names and Codes* (LOINC) for laboratory results. HL7 provides an umbrella standard that incorporates more specialised standards, creating the concept of a *vocabulary of coding systems*, in other words a meta-vocabulary. Some existing coding systems that are compatible and can be integrated into HL7 include Systematised Nomenclature of Medicine (SNOMED) for clinical terms, the MEDCIN clinical coding system, the Logical Observation Identifiers Names and Codes (LOINC) for laboratory results, the International Classification of Diseases, version 10 (ICD-10). ICD-10 is a coding system used for epidemiological and many health management purposes and include terms for the analysis of population health status, monitoring of the incidence and prevalence of diseases and other health problems in relation to socioeconomic and other variables.

In an integrated EHR system (with HL7 version 3 at its centre), a system can be envisaged which provides fully standardised data system spanning all areas within the health care system (Figure B.2). Such a system would not only vastly improve efficiency of the health care system, but would tremendously facilitate its study through the availability of high quality data to address a wide range of questions.

B.5 Survey Data

In health care research surveys are commonly used to collect information from a group through interviews or questionnaires. Surveys may be cross-sectional or longitudinal. In cross-sectional surveys observations on subjects are collected at a single point in time. In longitudinal surveys, observations are collected at several points over a period of time. The sample of subjects in longitudinal studies will be representative of the same population, but they may not necessarily be the same individuals. Panel data is a special subset of longitudinal data, where the same group of individuals is followed over the study period.

Cross-sectional data is easier to collect but will provide limited insight into the population dynamics which tends to evolve over time. The time component in studies of health care demand and utilisation is important because it is desirable to study changes in preferences influenced by patient characteristics.

The accuracy of survey data is often a source of major concern. It is extremely difficult to control for biases. For example, it was noted in **?**] that there are differences between the way that people respond to written surveys versus oral surveys. This is generally attributed to people being more candid about sensitive questions in an written surveys. Furthermore, it was found that the survey results depended on the order in which the questions were asked. This was attributed the people “learning” about their health status through the course of the questionnaire.

In spite of the difficulties in collecting accurate survey data, there is clearly a role for surveys in understandings attitudes and perceptions about healthcare demand and other aspects of the healthcare system. However, great care should be taken to use as large a sample size as possible and to collected the data in a consistent fashion throughout the survey. Studies based on pooling of data from different surveys are questionable, because of the difficulty in controlling for the varying biases within the dataset.

Surveys of self-reported health status are commonly used to assess correlations between socioeconomic status and health. However, this raises the important question, “Is there a correlation between socioeconomic status and how people assess their health?” If there were such a correlation, then it would introduce a bias into surveys of this nature. This question was addressed in **?**]. They compared the results of a health status survey with mortality data in the Swedish Survey of Living Conditions for 1975–1997. They did not find a strong correlation between socioeconomic status and self-reported health status. Therefore, properly conducted surveys of self-reported health status are generally valid as a measure of the link between socioeconomic status and health.

Appendix C

Mathematical Programming Packages for Computers

C.1 Mathematical Software

There are a large number of software packages which are designed to aid in mathematical analysis. A few of the more popular ones are listed below.

XXX rewrite in more generic terms (get away from the ODE kick) XXX

XXX add at least three stats packages XXX

Berkeley Madonna: Berkeley Madonna is an all purpose numerical differential equation solver. It is able to solve systems of differential equations rapidly and graphically display the system evolution at run time. It was developed at the University of California at Berkeley and is available for both Microsoft Windows and Mac OS X.

Maple: Originally developed at the University of Waterloo, Maple is now one of the pre-eminent software packages in mathematics. It is capable of both symbolic and numerical analysis of systems of differential equations. It is also a general purpose computer algebra system, with support for most branches of mathematics. Maple is available for Microsoft Windows, Linux, and Mac OS X.

Mathematica: Mathematica from Wolfram Research is a full featured suite for the mathematical sciences. It has extensive capabilities in both computer algebra calculations and numerical analysis. It is available for Microsoft Windows, Mac OS X, Linux, Solaris, and most other unix operating systems.

MATLAB: MATLAB is one of the most widely used software packages for numerical analysis in applied mathematics, science, and engineering. MATLAB has superior numerical algorithms and contains Maple as its symbolic manipulation tool. It has a large number of toolkits supporting a wide range of mathematical specialisations. It is developed and supported by MathWorks and is available for Linux, Solaris, Mac OS X, and Microsoft Windows.

C.2 Simluarion and Modelling Codes

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 (DML), which was developed as part of the Scalable Simulation Framework (SSF) and the Extensible Markup Language (XML). ANML was developed at the University of Calgary. Their ANML processor is open source and freely available.

dML: dML (deX Modelling Language) is an object-oriented 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 representations, which represents the process flow in the simulation. It is particularly well-suited to modelling queueing problems 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 to 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 U. S. Air Force and released in 1962. This initial version was implemented as a FORTRAN preprocessor, producing FORTRAN code which 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 it 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 utilises a layered approaching, 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-lever 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. SLX syntax is C-like and it is not an object-oriented language. However, it does have object-based features.

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 is an object-oriented framework targeted at programmers developing simulation models in Java. It provides support for building models with a graphical user interface. The acronym DESMO-J stands for "Discrete-Event Simulation and Modelling in Java". 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 Licence.

SSF: SSF (Scalable Simulation Framework) is an open standard for a discrete-event simulation application program interface (API). SSF 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 is a high-performance discrete event simulation engine that runs over a standard Java virtual machine. JiST is an acronym for "Java in Simulation Time". 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 parallelisation 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 characterised 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 specialised 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 2D and 3D animation, as well as optimisation. 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 organised 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 98, Windows Me, Windows 2000, Windows Server 2003, and Windows XP. In addition to Arena, Rockwell Automation also offers OptQuest, an optimisation suite for Arena.

AutoMod: Automod is a graphical modelling package that provides 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 2000 or XP.

eM-Plant: This package focuses on the modelling and simulating of production systems and processes. It provides the capability to optimise material flow, resource utilisation and logistics. It takes an object-oriented approach to model design and has extensive features for visualisation and animation of models. eM-Plant is developed and supported by UGS. It 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 two-dimension flowchart animation and three-dimensional models. Enterprise Dynamics runs under Microsoft Windows 98, 2000, and XP.

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 graphical interface utilising 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 NT family (Windows XP, 2000, NT 4.0+) and Mac OS X.

iThink and Stella: iThink and Stella from ISEE Systems use take 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 more towards education and research. Both packages run under Microsoft Windows and Mac OS X.

JSIM This a Java-based simulation and animation environment, which focuses on web-based simulation. It is 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.

MATLAB and SimuLink: These software packages are products of MathWorks, which is one of the leaders in developing software for scientific and engineering research. Employed extensively on large scale numerical research projects, MathWorks software is noted for its reliability and minimal number of bugs. MATLAB is a high-level language and interactive environment for performing computationally intensive modelling. Simulink is a graphical environment for developing models of dynamical systems. It essentially provides a graphical interface to MATLAB. SimEvents is an extension to SimuLink, which is specialised for discrete event simulation. A large number of other special purpose extensions for both MATLAB and SimuLink are also available. MATLAB and SimuLink are available for Linux, Solaris, Mac OS X, and Microsoft Windows.

MicroSaint: MicroSaint is 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 two-dimensional animation of the model. Support for optimisation is also provided by the OptQuest module, which is included with the package. The package runs under Microsoft Windows 98, 2000, ME, XP, and Server 2003.

OMNeT++: This is a discrete event simulation package with strong GUI support. It's utilises 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 which have concurrently executing components, which interact with each other. Ptolemy has a graphical interface for constructing simulation models using block diagrams. It supports 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 followup Java-based system called Ptolemy II is under development.

SansGUI: This is 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 which 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. It runs under Microsoft Windows 95, 98, ME, NT 4.0, 2000, or XP.

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 95, 98, NT, 2000, and XP.

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 which is compiled and executed. Simile runs under Microsoft Windows (98, ME, NT, 2000 and XP), 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, allowing it to be easily accessed by other programs or web interfaces.

Traffic: This is a suite of programs for 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 the modelling in the service sector and the other focused on the manufacturing sector. The models may be displayed in two-dimensional animation. A variety of extension modules are available, supporting extra capabilities such as three-dimensional animation, CAD design, and optimisation. 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 2000/XP, Linux, Solaris, and Mac OS X.

Ps-i: This is a Tcl/Tk based environment for agent-based simulations. It has built-in routine optimisation 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. SeSAm agents consist of a body, that contains a set of state variables and a behaviour that is implemented in the form of uniform modelling language (UML) style diagram. The package is able to deal with complex multi-agent systems (MAS) 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 for marketing scenarios and other types of modelling social interactions.