CHAPTER 1

The Whys, Whats, and Whens of Modelling in Healthcare

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

1. Why Model in Healthcare?

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

Healthcare is one of the oldest and largest professions in the world. Paintings discovered in the Lascaux caves in France, radiocarbon dated at over 15,000 years old, are interpreted to show the use of plants as healing agents. The Edwin Smith papyrus, dated between 3,000 and 4,000 years old, describes the examination, diagnosis, and treatment of numerous trauma injuries. Traditional Chinese medicine has origins dating back to the 5th century BC, and it is still in use today.

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

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

[1]All numbers are based on the 2003/2004 fiscal year.

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

2. What Is a Model?

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

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

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

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

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

There are many different models that are applicable to solving questions in the field of healthcare, and there is no such thing as a unique "best" model for a given problem. In fact, in most cases, more than one model discussed in this book 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.

3. When to Use Modelling in Healthcare

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

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

- predicting and adjusting the future demand for healthcare (see Subsection 7.1 of Chapter 3, Subsection 4.3 of Chapter 14 and Subsection 4.3 of Chapter 15);
- examining demographic factors which relate to health (see Subsection 4.2 of Chapter 5, Subsection 4.2 of Chapter 6, and Subsection 4.3 of Chapter 7);
- understanding and adjusting patient health behaviour (see Subsection 4.1 of Chapter 8, Subsection 4.1 of Chapter 10, and Subsection 4.3 of Chapter 11);
- decreasing wait times and understanding bottlenecks for healthcare access (see Subsection 4.1 of Chapter 14, Subsection 4.3 of Chapter 15, and Subsection 4.2 of Chapter 15);
- understanding and controlling the spread of communicable diseases (see Subsection 4.2 of Chapter 8, Subsection 4.2 of Chapter 12, and Subsection 4.2 of Chapter 13);

- optimizing healthcare delivery (see Subsections 4.1, 4.3, and 4.2 of Chapter 16).

There are many more problems which could fall under the heading of modelling in healthcare than those listed above. Some of these are covered in this book, and others are not. Many of those which are not covered are problems which are highly specialized in nature (for example, the mathematical modelling employed for the delivery of radiation therapy or analysis of MRI data). Such problems generally require a level of mathematics well beyond the scope of this book.

4. Related Reading

Detailed information on changes in national healthcare expenditures can be found in reference [**199**].

CHAPTER 2

How to Use This Book

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

Every man is wise when attacked by a mad dog; fewer when pursued by a mad woman; only the wisest survive when attacked by a mad notion.
Robertson Davies (1913–1995)

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

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

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

The remaining chapters discuss particular models that can be applied in the field of healthcare. Each of these chapters is given an artistic title that provides some insight as to where the models discussed might be used and a scientific title that provides the standard name for the model examined in the chapter. A table of the models considered in this book can be found in Chapter 3 (Table 3.1, page 13).

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

In order to ease reading, the layout for chapters on specific models is uniform. Most chapters are divided into five sections, entitled *Model Overview, Common Uses, Model Details* or *Mathematical*

Details, Examples, and *Related Reading.* In the *Model Overview* section we give a brief description of the model. These overview sections avoid mathematical language and should be readable by anyone with a solid high school background in science. The next section, *Common Uses*, provides a list of example questions that the modelling technique could be used to address. These lists are not complete, but they are intended to provide an idea of what kind of problems the type of model is capable of answering. In the *Model Details* or *Mathematical Details* section we give a more detailed mathematical description and analysis of the model. Wherever possible, we provide all the necessary scientific background to read these chapters; however in some cases the models are complicated enough that this is impossible. Sections that may require a substantial undergraduate level of knowledge are

- Chapter 11 (Game Theory), Section 3, which requires differentiation,
- Chapter 12 (Network Theory), Section 3, which discusses Graph Theory,
- Chapter 13 (Markov Models), Section 3, which requires matrix manipulation,
- Chapter 14 (System Dynamics), Section 3, which requires the use of differential equations, and
- Chapter 16 (Optimization), Section 3, which requires differentiation and matrix manipulation.

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

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

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

1. The Language of Modellers

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

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

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

Qualitative Models: Models designed to provide insight about why a given situation exists and what its driving factors might be. Such models do **not** provide numerical results pertaining to a given situation.
Disease: any *negative* health effect (for example, viral and bacterial infections, genetic disorders, increased chances of accidents causing harm, etc.).
Risk (Factor): any action or situation, be it beneficial or detrimental, that affects the probability of experiencing disease.

CHAPTER 3

The Modelling Process

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

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

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

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

Guiding Principles in Modelling.

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

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

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

All assumptions should be clearly stated: All models are built on a set of assumptions, some of which are testable and others that are not. These

assumptions must be clearly stated and, whenever possible, tested. Assumptions that are not testable should be discussed with experts from within the field.

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

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

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

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

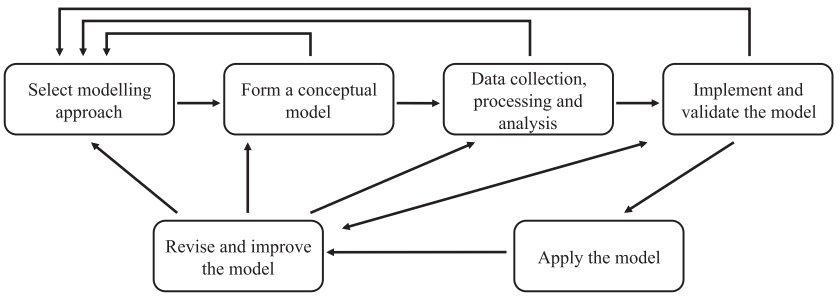

FIGURE 3.1. **The modelling process:** One View of the Modelling Process

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

1. Selecting a Modelling Approach

Contrary to what is commonly taught in high school, very few problems have a unique solution. Indeed, in our experience, most questions (be it healthcare related or not) can be solved in more than one manner and can have more than one reasonable solution. Likewise, for most problems, more than one modelling approach is possible, and each will have advantages and disadvantages. Therefore one of the first concerns a modeller will have to deal with is selecting which modelling

technique to apply. This selection process will generally be driven by many factors, including the type of data available, the nature of the situation to be modelled, and the type of question posed. In general, the most convincing results are obtained when multiple modelling techniques are applied and their results support each other.

In Table 3.1 we list the models discussed in this book and provide a brief explanation for the main usage of each with respect to healthcare. We further split the models into the two broad modelling categories of qualitative models and quantitative models. We describe these categories below.

TABLE 3.1. Categorization of the models covered in this book, along with locations.

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

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

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

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

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

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

A list of qualitative models discussed in this book can be found in Table 3.1.

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

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

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

are we interested in efficient use of resources? Is a single queue equally fair to all patients, or should some prioritization be employed? If all physicians are occupied with complicated cases that take a long time to resolve, the waiting time for those remaining in the queue should increase substantially. The complexity of this example increases manyfold as greater detail is brought into the model. The harmonic operation of multiple healthcare services, all relying on the same pool of resources, for example, can be even more complex. One task of mathematical models is to make complex situations more manageable.

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

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

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

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

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

of computation. However, the time step may be specified to be so small that continuity is essentially preserved.

A list of quantitative models discussed in this book is displayed in Table 3.1.

2. Forming a Conceptual Model

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

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

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

3. Data Collection, Processing, and Analysis

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

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

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

In many problems in healthcare, data collection is a limiting factor in model development and analysis. This is beginning to change as computerized patient tracking is developed and implemented, but even then confidentiality issues cause

data collection challenges. Ignoring the issue of patient confidentiality, data collection in healthcare remains a resource-intensive undertaking that often requires conducting surveys or population studies. Such surveys can be extremely expensive and time consuming to complete, and even on completion the data may be corrupted by survey bias.

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

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

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

4. Implementing and Validating the Model

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

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

A second approach to implementing a model is provided by the tools of *mathematical analysis*. If the modeller is able to describe the system in terms of equations,

then analytic or numerical solutions may be sufficient to "solve" the model without the need for simulation. This provides several strong advantages over simulation. For example, analytical methods produce exact reproducible solutions without the need for (often expensive) software. Furthermore, analytical methods often provide deep insights into the workings of a system. However, analytical solutions for complex models are often difficult or even impossible to achieve.

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

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

5. Applying the Model

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

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

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

6. Revising the Model

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

7. Example

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

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

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

One of the major focuses of population models is the growing rates of chronic diseases. Diseases such as cardiovascular disease, diabetes, and cancer are becoming of great concern worldwide. Data from the United States indicate that preventable illness constitutes approximately 70% of the illness burden and the associated costs. Preventable causes, such as cigarette smoking and obesity, represent eight of the nine leading causes of death in the United States [**79**]. This represents a huge challenge in finding ways to effectively promote lifestyle modification and prevent disease [**181**]. By reducing the number of people advancing from the healthy population to the at-risk population, the overall demand for healthcare resources may be reduced.

Examples of population models in healthcare can be found in Subsection 4.2 of Chapter 5, Subsection 4.2 of Chapter 6, Subsection 4.3 of Chapter 7, Subsections 4.1 and 4.2 of Chapter 8, Subsection 4.2 of Chapter 12, and Subsection 4.2 of Chapter 13.

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

Patient-doctor interactions have been studied at the individual level using game theory [**62**], [**67**], [**210**] (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 [**173**] (see Subsection 4.2 of Chapter 12).

Examples of behavioural models in healthcare can be found in Subsections 4.1 and 4.2 of Chapter 8, Subsection 4.1 of Chapter 10, Subsections 4.1, 4.2, and 4.3 of Chapter 11, and Subsection 4.1 of Chapter 13.

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

Operational models are most frequently applied within hospitals, clinics, and other healthcare facilities and measure demand in terms of wait times or blocked patients (patients who sought healthcare but could not access it). Queueing theory (Chapter 15) and discrete event simulation (Subsection 4.3 of Chapter 9) have largely been the methods of choice for such problems, since random arrivals and queueing heavily influence the demand for service. Another method that is becoming more common in this field is system dynamics (Chapter 14).

Examples of operational models in healthcare can be found in Subsection 4.2 of Chapter 12, Subsection 4.1 of Chapter 14, Subsections 4.2 and 4.3 of Chapter 15, and Subsections 4.1, 4.2, and 4.3 of Chapter 16.

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

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

Modelling at the global level is not simple, and much work remains to be done in this area. Recently, global models have often been implemented in terms of system dynamics (Chapter 14) [**111**], [**112**].

Examples of global models in healthcare can be found in Subsections 4.3 and 4.2 of Chapter 12, Subsections 4.1, 4.2, and 4.3 of Chapter 14, and Subsection 4.2 of Chapter 15.

8. Related Reading

For another description of the modelling process see [**41**]. Reference [**51**] discusses the concept of mental models. References [**62**], [**67**], and [**210**] examine game theory and some of its applications to healthcare. Reference [**173**] investigates social network theory. Reference [**13**] provides a taxonomy of 77 verification and validation techniques

for conventional simulation models. References [**79**] and [**181**] develop models for use in health resource allocation. Reference [**41**] provides an introduction to simulation and modelling. References [**111**] and [**112**] look at system dynamics modelling in healthcare.

CHAPTER 13

The Future Starts Now

Not the power to remember, but its very opposite, the power to forget, is a necessary condition for our existence.
Sholem Asch (1880–1957)

It is singular how soon we lose the impression of what ceases to be constantly before us. A year impairs, a luster obliterates.
Lord Byron (1788–1824)

Markov Models

1. Model Overview

To begin this chapter, we consider a classical model of disease spread. In this model, we examine a collection of individuals that can take on one of three states: susceptible, infected, and recovered. If an individual in the model is susceptible, then he or she has a probability of becoming infected during the next time step. This probability is not based on any demographic factors of the individual, but instead on the number of infected individuals currently in the model. In particular, the more people that are currently infected, the more likely it is that a susceptible individual comes in contact with an infected individual and thereby becomes infected. If an individual in the model is infected, then he or she has a probability of becoming recovered during the next time step. This probability is fixed and is based on the standard recovery rate of an infected person (for whatever disease one is studying). Finally, if an individual is recovered, he or she will remain in the recovered state during the next time step. A visual of this model is given in Figure 13.1.

The simple model described by Figure 13.1 is often called the SIR model and is a classic example of a Markov model in healthcare. In particular, the model examines a collection of *objects* that are assigned a series of *states* over a period of *time steps*. Moreover, at the end of each time period, the objects move from one state to the next according to transition probabilities (also called transition rates) that depend only on the current state of the system. These are the key aspects of a Markov model. Markov models are models that examine a collection of objects in a system that move through a series of states. Moreover, Markov models work on the assumption that the future state of the object is determined by a random process dependent only on the current state of the system. This assumption is so basic to the methodology of Markov models that it is generally referred to as the *Markov assumption*. Due to the Markov assumption, Markov models are "forgetful" in the sense that a knowledge of the past states of the system is not required to predict the future. In spite of this, Markov models can exhibit deep structure through the cumulative effects of repeated stochastic events.

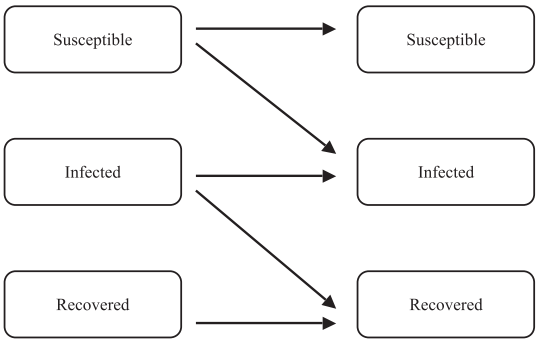

FIGURE 13.1. **The SIR model of disease spread:** The classical SIR model (Susceptible, Infected, Recovered) of disease spread can easily be viewed as a Markov model.

Before implementing a Markov model, it is important to determine the list of all possible states that an object can take. The collection of all possible states that an object can take is called the *state space*. In more complicated models, the list of possible states becomes longer, and the manner in which objects move from one state to another becomes more complex. For example, the objects may be people and the states may be the amount and type of healthcare services each person uses during a given time period. This would result in an extremely large number of possible states, and therefore a rather complicated model to implement.

The Markov assumption *is that the future state of an object in the model is determined by a random process dependent only on the current state of the system.*

The simplest type of Markov model is the *finite state Markov chain*. In such models the number of possible states is finite, and transition from one state to the next occurs at predefined points in time. Such models are relatively simple to implement but can still model some surprisingly complex systems. As such they are often a good first choice for many modelling problems.

Many people believe that the Markov assumption causes Markov models to be extremely limited in application. For example, your probability of gaining weight is partly dependent on your current weight, but it is also partly dependent on your history of weight gain (see Subsection 4.2). Even though the Markov assumption forces some level of forgetfulness on the models, it is nonetheless possible to build memory into a Markov model. The way this is done is to create new states that incorporate the memory for the desired trait. For example, we might create a state called "currently normal but was obese". Markov models that incorporate memory in this manner are sometimes referred to as *higher-order Markov models*.

Whenever we choose to approach a problem via Markov models, it is of the utmost importance to test whether the Markov assumption is valid.

If the model incorporates one level of memory, it is referred to as a second-order Markov model (or a Markov chain of order 2), and models that do not incorporate any memory are sometimes called first-order Markov models.

Whenever we choose to approach a problem via Markov models, it is of the utmost importance for testing whether the Markov assumption is valid. Fortunately there is a simple method to test the Markov assumption. Basically, we build a higher-order Markov model and check that its results agree with the original lower-order model. If the Markov assumption holds, the two models will produce the same results (see Figure 13.2, page 140).

There are many generalizations to finite state Markov models, including infinite state models, Markov processes, semi-Markov processes, and Markov decision processes. These generalizations are discussed briefly at the end of Section 3, but they are beyond the scope of this book.

2. Common Uses

Markov models explore the properties of objects in a system that move through a series of states. The most important aspect of Markov models is the assumption that the system satisfies the Markov assumption: the future state of the object is determined by a random process dependent only on the current state of the system. In healthcare, this assumption is well suited to modelling the movement of patients through disease states. In regards to disease states, Markov models are suitable to answer questions such as the following:

- *How do immunization rates impact the spread of disease through a population?*
- *How many people will be affected by diabetes in future years?*
- *At what disease state is treatment most suitable to prevent disease spread?*

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

- *How does doctor-patient loyalty affect the use of the healthcare system?*
- *To what level does an individual's past BMI status impact his or her future BMI status?*
- *Does surgeon skill affect patient progress through post-surgery recovery stages?*

3. Mathematical Details

In Markov models we begin with a collection of objects and a list of possible states for each object. For example, the object may be individual people that can take one of two states: healthy or sick. The collection of all possible states that an object can take is called the *state space*. At each time interval, the model assigns every object in the system to exactly one state from the state space. At the end of each time period, the objects move from one state to the next according to *transition probabilities* (or *transition rates*) that depend only on the current state of the system. That the transition probabilities depend only on the current state of the system is the key aspect of Markov models and is generally referred to as the *Markov assumption*.

Due to the Markov assumption, Markov models are "forgetful" in the sense that a knowledge of the past states of the system is not required to predict the future. In spite of this, Markov models can exhibit deep structure through the cumulative effects of repeated stochastic events.

3.1. Finite State Markov Chains. We begin our discussion with the simplest type of Markov models, *finite state Markov chains*. The words "finite state" mean exactly what one would suppose them to mean: that the list of possible states for an object is finite. The final word, "chain", refers to the assumption that transition from one state to the next occurs at predefined points in time. For example, in examining the spread of disease, we might decide to update each individual's state at the end of each day. Alternately, states might be updated on an irregular, but still predefined, basis. For example we might be interested in studying an individual's BMI status when he or she turns 16, 19, 25, 50, and 65. Whether time periods are evenly spread or not makes no difference in the mathematics required to analyze the model.

Let $S = \{s_1, s_2, \ldots, s_i, \ldots, s_N\}$ be the state space of the model (the list of all possible states that an object can take). Let X^0 be a column vector of length N that represents the initial state of the system. That is,

$$X^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_i^0 \\ \vdots \\ x_N^0 \end{bmatrix},$$

where x_i^0 is the number of objects in state i at time step 0. In general we shall use

$$X^t = \begin{bmatrix} x_1^t \\ x_2^t \\ \vdots \\ x_i^t \\ \vdots \\ x_N^t \end{bmatrix},$$

where x_i^t is the number of objects in state i at time step t.

Next, let $\Pr^t(i \to j)$ be the probability of an object moving to state s_j at time $t+1$ given that the object was in state s_i at time t. Creating the matrix

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

we find that the state vector for the system at time $t+1$ is the matrix multiplication of P_t and the state vector of the system at time t:

$$X^{t+1} = P^t X^t. \tag{24}$$

The matrix P^t is generally referred to as a *transition matrix* and may be dependent on the time t or the state of the system at time t.

In order for a Markov chain model to run correctly, transition matrices must satisfy several special properties. First, all elements of the matrix must be 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

appearing or disappearing from the model (models that allow entry into the model or exit from the model are discussed in Subsection 3.2).[1]

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

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

$$X^t = P^{t-1}P^{t-2}\dots P^1P^0X^0. \tag{25}$$

If the transition probabilities do not change over time, that is, if $P^t = P^0$ for all t, then the Markov model is called *time homogeneous*. Time homogeneous Markov models allow for obvious simplifications to formula (25):

$$X^t = [P^0]^t X^0.$$

3.2. Markov Model Involving Entry and Exit. In the previous subsection, we discussed Markov models that did not allow new objects to enter the model nor objects from the model to exit the model. In some cases one wishes to incorporate this idea into the model (for example to represent births or deaths). This can be accomplished in several manners. The easiest manner is to create two new states, one labelled "to-be-born" and one labelled "dead". Next set the initial number of objects in the state "to-be-born" to a very high number (say 1,000 times the size of the model), and set the initial number of objects in the state "dead" to zero. Finally, one adjusts the transition matrices P^t to allow for objects to be born into the appropriate states and for objects from the appropriate states to transition into the "dead" state.

There are other methods of incorporating birth and death into Markov models, and in general they are quite straightforward. For example, one can rewrite the state-vector equation (equation (24)) to $X^{t+1} = P^tX^t + b^t - d^t$, where b^t is a vector of objects born into each state at time t and d^t is a vector of objects exiting each state at time t. The disadvantage of this method is that one must take care that the total number of objects in a given state never becomes negative. In particular, one should never have more objects exit a state than there are objects in that state. In the first method discussed, this is taken care of automatically when we check that each column of the transition matrix must sum to 1.

An additional advantage of the first method is that it allows us to keep track of how many objects enter and exit the system over the course of the model. (The number entering the system is the difference of the initial number and final number of objects in the state "to-be-born"; the number exiting the system is the final number of objects in the state "dead".)

3.3. Higher-Order Markov Models. On the surface, the Markov assumption appears to create models which are extremely limited in application. For example, if we were modelling the spread of disease through a population, then we would be interested in the two states "uninfected" and "infected". However, it is well known that patients who have recovered from a virus are less likely to become infected again from the same disease. Therefore, if the infected state simply feeds back into the uninfected state, the model is unlikely to provide useful information.

[1] A few texts consider transition matrices as the transposition of the above approach; in this case, the sum of each row totals to 1.

Even though the Markov assumption forces some level of forgetfulness on the models, it is nonetheless possible to build memory into a Markov model. The way this is done is to create new states that incorporate the memory for the desired trait. For example, in the case of modelling spread of disease through a population, one could create states labelled "susceptible", "infected", and "recovered". The recovered state now effectively contains the memory that the individual was once infected.

Markov models that incorporate memory in this manner are sometimes referred to as *higher-order Markov models*. The order of the Markov model is one more than the level of memory the model attempts to incorporate. For example, if the model incorporates one level of memory, it is referred to as a second-order Markov model (or a Markov chain of order 2), and models that do not incorporate any memory are sometimes called first-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.

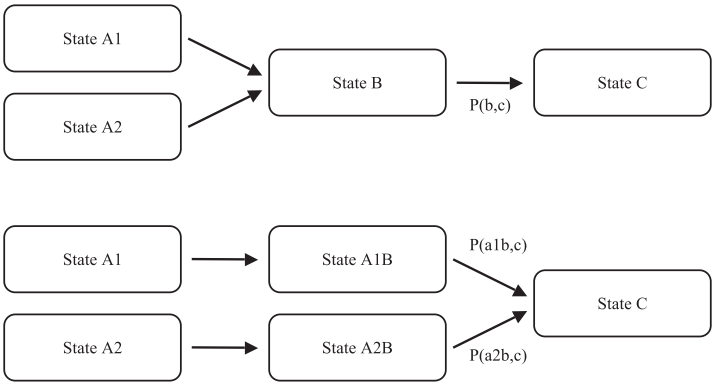

FIGURE 13.2. **Testing the Markov assumption:** One method for testing if the Markov assumption holds is to turn a first-order Markov model into a second-order Markov model and check if the transition probabilities are affected. In the first-order model (top) we have two states, $A1$ and $A2$, which 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 first-order model suffices. Otherwise, one should test if the second-order model satisfies the Markov assumption.

3.4. Testing the Markov Assumption. Higher-order Markov models provide us with insight into 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 13.2.

3.5. Infinite State Markov Models. For Markov chains with a finite number of states, the transition probabilities may be represented as a transition matrix (see Subsection 3.1). If the number of states is infinite, then this property does not apply. Instead, the transitions must be described in terms of functions. Recall that for finite chains we used $\Pr^t(i \to j)$ to represent the probability of an object moving to state s_j at time $t+1$ given that the object was in state s_i at time t. If there are an infinite number of states, the indices i and j may no longer be integers, and so building the matrix P^t is no longer possible. Instead one creates a function

$$f(i, j, t) = \Pr^t(i \to j).$$

Various mathematical techniques have been developed to study such functions, most of which focus on the question of whether there is a state s_i that has a high probability of being occupied regardless of starting conditions. These techniques are beyond the scope of this book.

3.6. Markov Processes and Semi-Markov Processes. The Markov models discussed above were Markov chains, meaning that all state transitions occur at fixed predefined time intervals. In the 1920s, a more general class of models, called *Markov processes*, in which transitions occur at arbitrary times was also developed. In these models, time is viewed as a continuous variable, so time steps can occur at any point. One classic example of such a process is the "random walk of a drunkard", in which a point stumbles in a random direction for a random distance. In this case, the concept of time is incorporated into distance, and so the point can be thought to be traveling in a random direction for a random length of time. In literature, the random walk of a drunkard is usually referred to as a *Wiener process* or *Brownian motion*.

Another generalization of Markov chains is *semi-Markov processes*. In a semi-Markov process, the transition probabilities depend not only on the current state of the system, but also on the time that it has spent in that state. The time that the system spends in each state is assumed to vary stochastically according to a probability distribution. Semi-Markov models have wide applicability in queueing theory, reliability modelling, and operations research. Recently, they have been applied to the modelling of chronic diseases, such as HIV.

3.7. Markov Decision Processes. Markov decision processes (MDPs) are an extension of Markov processes and Markov chains, in which the modeller is allowed to interact with the objects in the system by applying *actions* to the system. Applying an action to the system can be thought of as altering the transition matrix for a selection of time steps. MDPs are used extensively in business to help examine the effect of decision making in situations where outcomes are partly random and partly under the control of the decision maker. The basics of MDPs are not difficult, once the ideas behind Markov models are understood. However, further discussion on MDPs are beyond the scope of this book.

4. Examples

4.1. A Simple Doctor-Patient Loyalty[2]. To demonstrate the mathematics behind a simple time homogeneous Markov chain, consider a drop-in clinic with three doctors. In this particular drop-in clinic, no appointment is necessary, so patients may not see the same doctor on every visit. However, the patient (when returning to the clinic) may request a specific doctor. If the doctor is available that day, the patient's wait time increases but considerations are usually made.

We assume that a patient's preference for a doctor is completely determined by the doctor that he or she visited in his or her 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, are found in Table 13.1. Notice that some doctors inspire more patient loyalty than others.

TABLE 13.1. Transition probabilities for a hypothetical doctor-patient loyalty model.

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

Suppose the clinic has 300 patients that return on a regular basis, and we wish to see how these patients impact each doctor's workload. We begin by assuming that each doctor will see 100 of these patients, so

$$X^0 = \begin{bmatrix} 100 \\ 100 \\ 100 \end{bmatrix}.$$

The transition matrix for this Markov chain will be unchanging with time and will be equal to

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

Simple matrix-vector multiplication yields

$$X^1 = \begin{bmatrix} 102 \\ 118 \\ 80 \end{bmatrix},\ X^2 = \begin{bmatrix} 100.86 \\ 130.66 \\ 68.48 \end{bmatrix},\ X^3 = \begin{bmatrix} 98.7594 \\ 139.4878 \\ 61.7528 \end{bmatrix}, \ldots,$$

$$X^{20} = \begin{bmatrix} 89.6613 \\ 158.9361 \\ 51.4026 \end{bmatrix}, \ldots,\ X^{50} = \begin{bmatrix} 89.6414 \\ 158.9641 \\ 51.3944 \end{bmatrix}, \ldots,\ X^{100} = \begin{bmatrix} 89.6414 \\ 158.9641 \\ 51.3944 \end{bmatrix}.$$

This might lead us to conjecture that in the long run Doctor 1 will have approximately 90 patients, Doctor 2 will have approximately 159 patients, and Doctor 3

[2]The model discussed in this example also relates to Game Theory, Chapter 11.

will have approximately 51 patients. To confirm this more mathematically, we are interested in the *equilibrium distribution* of X:

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

To proceed, we must *diagonalize* the matrix P. To diagonalize a matrix P is to find an invertible matrix Q and a diagonal matrix D such that $P = Q\,D\,Q^{-1}$. This can be done quickly and easily with modern mathematical computing software. In this case diagonalizing P provides

$$P = Q\,D\,Q^{-1},$$

where

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

Therefore,

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

This mathematically confirms our predictions for this model. More importantly, if we repeat this analysis with an arbitary

$$X^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix}$$

with $x_1^0 + x_2^0 + x_3^0 = 300$, notice that

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

So regardless of initial conditions, the resulting distribution of patients will be the same.

This model illustrates a key feature of many Markov models: that the model will eventually approach an equilibrium or "steady-state". This means that the distribution of patients among the physicians will eventually approach an equilibrium distribution, which is *independent of the initial distribution.* In this case, 30% of the patients will see Doctor 1, 53% will see Doctor 2, and 17% will see Doctor 3.

4.2. Testing the Markov Assumption for a Male BMI State Model. In this example we consider a simple three-state Markov model examining changes in BMI status.[3] The portion of the survey data we use consists of following 5,316 males over the course of 18 years and collecting their BMI value ($BMI = mass\ in\ kg/(height\ in\ m)^2$) every two years. In this example we seek to check if the Markov assumption is a valid assumption when studying an individual's BMI status.

The basic model consists of the states "normal weight", "overweight", and "obese", which we shall abbreviate as nw, ow, and ob, respectively. The states correspond to BMI ranges of

$$\begin{aligned} nw &\Leftrightarrow BMI < 25, \\ ow &\Leftrightarrow 25 \leq BMI < 30, \\ ob &\Leftrightarrow 30 \leq BMI. \end{aligned}$$

(Note that underweight individuals ($BMI < 18.5$) are placed in the nw state; this is due to the lack of underweight individuals in the NLS database.) Since we will tune the model using NLS data, we will use a two-year time step between transitions. To allow for easier reading of notation, we shall create a time step index i which will correspond to the age of the individuals at that time step. For example, time step 20 will represent the time at which individuals are aged 20 or 21, time step 22 will represent the time at which individuals are aged 22 or 23, etc. As two years is a relatively long time, we will assume that any state can transition into any other state over each time step. Visually this produces the model shown in Figure 13.3.

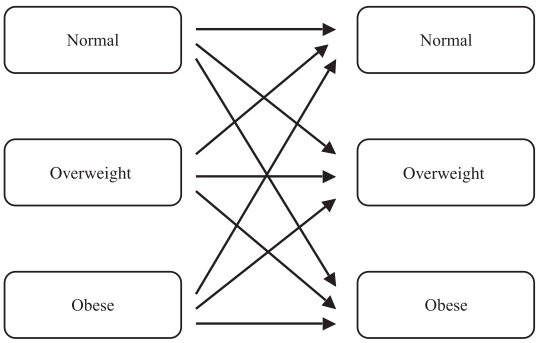

FIGURE 13.3. **A three-state Markov model of BMI status:** A visualization of the basic BMI state Markov model.

Let

$$X_i = \begin{bmatrix} nw_i \\ ow_i \\ ob_i \end{bmatrix}$$

be a vector of the number of individuals in each state at time i, where nw_i is the number of normal weight individuals, ow_i is the number of overweight individuals,

[3]The model was designed and tested against the National Longitudinal Survey (NLS) database, which is freely available at `http://www.bls.gov/nls/`.

and ob_i is the number of obese individuals. To determine the number of individuals in each state during time step $i+1$, we multiply X_i by a transition matrix T_i where

$$T_i = \begin{bmatrix} \Pr_i(nw|nw) & \Pr_i(nw|ow) & \Pr_i(nw|ob) \\ \Pr_i(ow|nw) & \Pr_i(ow|ow) & \Pr_i(ow|ob) \\ \Pr_i(ob|nw) & \Pr_i(ob|ow) & \Pr_i(ob|ob) \end{bmatrix}$$

and $\Pr_i(y|x)$ is the probability that at time step i an individual will move to BMI category y given that his or her current BMI category is x. We let these probabilities be time dependent to represent the fact that as individuals age, they will begin to behave differently. Calculating these probabilities for each time period is easily done by examination of the NLS data set. Doing so, one finds that at the 30th time step

$$T_{30} = \begin{bmatrix} 0.785 & 0.102 & 0.005 \\ 0.211 & 0.793 & 0.145 \\ 0.004 & 0.105 & 0.850 \end{bmatrix}.$$

In order to test the Markov assumption, we next produce a higher-order Markov model and check if it behaves the same as the first-order model. By this we mean that we change our three-state model into a nine-state model, where each state stores not only one's current obesity status but also one's obesity status from the previous time step. In order to facilitate discussion, we shall label the nine states $s_1, s_2, \ldots, s_9$, where

$$\begin{array}{lll} s_1 = \{nw \to nw\}, & s_2 = \{ow \to nw\}, & s_3 = \{ob \to nw\}, \\ s_4 = \{nw \to ow\}, & s_5 = \{ow \to ow\}, & s_6 = \{ob \to ow\}, \\ s_7 = \{nw \to ob\}, & s_8 = \{ow \to ob\}, & s_9 = \{ob \to ob\} \end{array}$$

and state $\{x \to y\}$ represents the condition that one time step earlier the individual was in state x and currently the individual is in state y. These nine states can transition according to the logical rule that the next state's x must be the current state's y. If

$$\widetilde{X}_i = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_9 \end{bmatrix}$$

represents the current state of the model, the transition matrix $\widetilde{T}_i$ would take the form

$$\widetilde{T}_i = \begin{bmatrix} \Pr_i(s_1|s_1) & \Pr_i(s_1|s_2) & \Pr_i(s_1|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_2|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_3|s_9) \\ \Pr_i(s_4|s_1) & \Pr_i(s_4|s_2) & \Pr_i(s_4|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_5|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_6|s_9) \\ \Pr_i(s_7|s_1) & \Pr_i(s_7|s_2) & \Pr_i(s_7|s_3) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Pr_i(s_8|s_4) & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Pr_i(s_9|s_9) \end{bmatrix}.$$

If the Markov assumption holds, then each row of $\widetilde{T}_i$ will consist of three (nearly) identical values as there should be no statistically significant difference between the

behaviour of individuals in s_1, s_2, and s_3 (for example). However, using the NLS data set, one can calculate that $\widetilde{T}_{30}$ would be

$$\widetilde{T}_{30} = \begin{bmatrix} 0.826 & 0.399 & 0.000 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.214 & 0.071 & 0.041 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.000 & 0.017 & 0.000 \\ 0.173 & 0.577 & 0.333 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.763 & 0.819 & 0.514 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.250 & 0.274 & 0.084 \\ 0.001 & 0.024 & 0.667 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.023 & 0.110 & 0.445 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.750 & 0.709 & 0.916 \end{bmatrix}.$$

As one can see, the Markov assumption does not hold (for example $\Pr(s_7|s_1) = 0.001$, but $\Pr(s_7|s_3) = 0.667$). Therefore a basic Markov model is not an appropriate approach to modelling obesity. We could continue by building higher and higher order Markov models until the Markov assumption held, but this is likely to be an unrewarding task. (On a final note, there is nothing special about examining time step 30. Examining other time steps shows the same disagreement between the data and the Markov assumption.)

4.3. A Mover-Stayer Model for Problematic Drug Use: A Compartmental Markov Model[4]. Standard Markov models assume that the same transition matrices apply uniformly to the entire population. However, it is often the case in epidemiological modelling that the population is divided into different compartments according to their susceptibility or infectiousness. The model then describes how the size of these compartments changes over time by means of equations describing the disease dynamics. A common approach to compartmental modelling is to use a mixed Markov process, which consists of a superposition or mixture of different Markov chains with independent transition matrices. In this example we review a compartmental mixed Markov model used to analyze heroin addiction [**187**].

The basic model views the population as consisting of two types of people: those who are susceptible to becoming problematic drug users and those who are "prudent" and hence not at risk of becoming drug users. Thus the compartmental Markov model consists of two subgroups, which we will call "movers" and "stayers". The *movers* represent the people who are susceptible to drug abuse and can therefore move about the various states of drug use. Conversely, *stayers* represent the people who are not at risk of becoming drug users.

In the model of [**187**], movers can become drug users either by coming in contact with other drug users or by contact with drug dealers. The diagram in Figure 13.4 is a compartmental representation of the model. In this figure, the straight lines represent possible state changes of the movers, and the curves represent the possible interactions between drug users and people susceptible to drug use (movers) in the population.

As shown in the Figure 13.4, this model uses several compartments to model the phenomenon. If a mover becomes a drug user, then he or she initially passes through a phase of light use. After a period of light use, he or she then moves on to a phase of heavy but invisible use. When usage becomes problematic, he or she

[4]The model discussed in this example also relates to System Dynamics, Chapter 14.

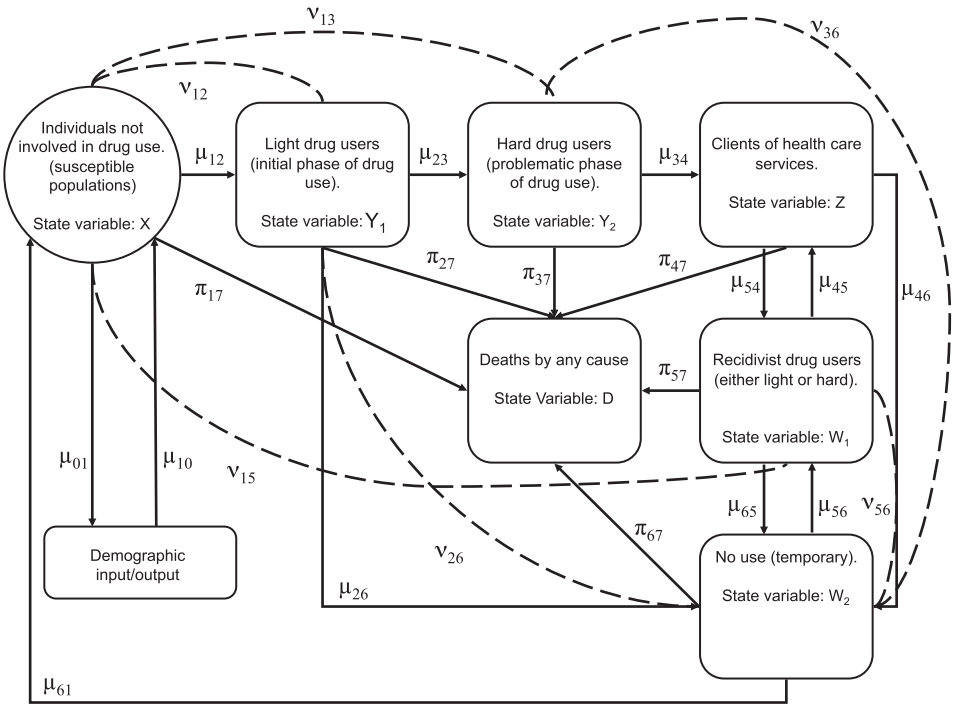

FIGURE 13.4. **System dynamics diagram of mover-stayer model epidemic drug use:** The parameters μ_{Ij} represent flow of movers from one state to other, the parameters ν_{ij} represent interactions between the different components in the model, and the parameters π_{i7} represent mortality from each of the components. Adapted from [**187**].

becomes visible as a heavy drug user and begins interaction with the healthcare system and social services. Finally, addictive use of drugs leads to possible reform and a possible recidivist phase.

In order to understand the spread of drug addiction and to 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 13.4 to write a set of eight coupled difference equations that describe the evolution of the system (see Table 13.2). In evolving the system, it is assumed that the lengths of stay in each of the compartments are exponentially distributed. To implement the model, a computer program is written to evaluate the difference equations. Since the computer program evaluates the difference equations using a series of predetermined time steps, this is mathematically 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 problematic drug use and may be obtained from studies of the incidence of drug use. The parameters π_{27} through

TABLE 13.2. Coupled difference equations represented by Figure 13.4.

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

$X(t)$	size of the susceptible population at time t
$S(t)$	proportion within the susceptible population who are "stayers" at time t
$S_0(t)$	proportion of the new population entering the susceptible population who are stayers at time t
$Y_1(t)$	population of light drug users at time t
$Y_2(t)$	population of hard drug users at time t
$Z(t)$	population whose drug use has made itself 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)

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

Using this model, the authors explored the effects of both primary and secondary preventive interventions. A primary intervention is one that is applied directly to the susceptible population. It is said to have an effectiveness P if a

proportion P of the movers in the susceptible population become stayers. The effect of secondary preventive interventions can be evaluated by modifying the ν and μ parameters. For example, the consequence of increased law enforcement would primarily be to decrease the parameter μ_{12}. Safe injection sites would primarily have an effect on the parameters ν_{56} and μ_{56}. The impact of healthcare policies on drug use would primarily be on the parameters μ_{45} and μ_{54}.

The model supports the statement that primary interventions are more effective than secondary interventions. However, there is substantial latency in the system after a program of primary intervention is initiated. That is, there would be no sign of any positive response for a significant period of time. In the case of the model applied to heroin drug addiction in Italy, this response latency would likely be about 6 years and possibly as long as 8 years. However, when the system does respond to the intervention, it does so rapidly and in a highly non-linear fashion. This result is important, as many intervention programs would be abandoned as "failures" if no improvement was seen for 5 years.

However, it should be noted although the model supports the statement that primary interventions are more effective than secondary ones, this model does not evaluate the cost of primary versus secondary interventions. For example, the model predicts a greater effect if the parameter μ_{12} is adjusted slightly than if the parameter ν_{56} is adjusted slightly. However, it does not evaluate how difficult it is to adjust each parameter. In application one must determine how much effort is required to adjust each parameter and weight this against the impact of the adjustment.

5. Related Reading

Markov models are closely related to Systems Thinking and System Dynamics (Chapter 14), as well as Queueing Theory (Chapter 15). Indeed, Markov models provide an alternate approach to developing and implementing many of the ideas in those methodologies. Markov models are often used as a base for other models to build on. In this manner, Markov models are often used in conjunction with Network Models (Chapter 12), Game Theory (Chapter 11), and many statistical models (Part 2).

Reference [**187**] contains details for the example in Subsection 4.3.

Reference [**89**] looks at the mover-stayer Markov model using various estimators and at the accuracy of these estimators. Reference [**160**] expands previous work on a Markov model for predicting future need for resources by taking varying utilization rates between age groups into account. Reference [**45**] uses a Markov model to track the movement of patients through the disease states of malaria. Reference [**142**] develops a Markov chain for the analysis of a centralized medical record system in a large hospital. Reference [**209**] derives exact HIV incubation distributions under treatment by antiviral drugs based on first passage probability distributions for some continuous-time Markov chains. Reference [**147**] uses a Markov model to describe movements of geriatric patients within a hospital system. Reference [**113**] uses Markov modelling to forecast the number of people with diagnosed and undiagnosed diabetes by age, race, ethnicity, and sex. Reference [**27**] uses Markov models to analyze various social science processes, such as bed occupancy rates, brand loyalty, occupational mobility, voter patterns, and malaria. Reference [**50**] proposes a method for analyzing surveillance data for communicable pathogens using a "structured" hidden Markov model.

Reference [**202**] uses a Markov decision process to examine the cost-effectiveness of alternative screening strategies for HCV infection in comparison with no screening. Reference [**117**] presents a Markov decision process model to evaluate different screening policies

for breast cancer. Reference [**197**] presents a Markov decision process for addressing the unresolved issue of optimal time to initiate HIV therapy.