Modelling Methods for Pharmacoeconomics and Health Technology Assessment
An Overview and Guide

James E. Stahl
MGH – Institute for Technology Assessment, Massachusetts General Hospital, Boston, Massachusetts, USA

Contents

Abstract ....................................................................................132
1. Basic Concepts ..........................................................................132
  1.1 Defining Terms .......................................................................132
  1.2 What is Technology? .................................................................132
  1.3 Why Simulation Models? ..............................................................132
  1.4 Cost Effectiveness of Simulation .......................................................133
2. Going DEEPP (Describe, Evaluate, Explore, Predict and Persuade) .......................................................134
  2.1 Steps in Building a Simulation Model ...................................................134
  2.2 Simulation Model Validation ..........................................................134
3. Sensitivity Analysis ........................................................................135
4. Simulation Models and the Kind of Evidence They Provide for Decision Making .........................................136
5. Choosing the Most Appropriate Simulation Method ........................................................................136
  5.1 Project Type .........................................................................136
  5.2 Population Resolution ................................................................138
  5.3 Interactivity/Feedback ...............................................................138
  5.4 Treatment of Time ...................................................................138
  5.5 Treatment of Space ..................................................................138
  5.6 Resource Constraints .................................................................138
  5.7 Autonomy/Freedom of Action of Simulated Entities/Populations ..............................................138
  5.8 Embedding of Knowledge ............................................................138
6. Types of Methods ........................................................................138
  6.1 Mathematical and Statistical Models ..................................................138
  6.2 Decision Trees .......................................................................139
  6.3 Markov Family .......................................................................139
    6.3.1 Mimicking Memory .............................................................140
    6.3.2 Cohort Models ..................................................................141
    6.3.3 Monte Carlo Markov Simulation .................................................141
    6.3.4 Markov Decision Processes ......................................................141
    6.3.5 Limitations of Markov Models ....................................................142
  6.4 Influence Diagrams and Causal Inference ...............................................142
  6.5 System Dynamics and Compartment Models ..........................................143
  6.6 Discrete-Event Simulation (DES) ........................................................144
    6.6.1 DES Software ..................................................................145
    6.7 Agent-Based Simulations ................................................................145
7. Conclusions .............................................................................146
Abstract

This paper provides an overview of, and guidance as to when, why and how to choose and use, different simulation modelling methods as applied to healthcare. What simulation is and why it is necessary in addressing healthcare problems are discussed. In addition, key criteria for choosing an appropriate method (project type, population resolution, interactivity, treatment of time and space, resource constraints, autonomy and how knowledge is embedded) are covered. Key concepts for each method, moving from the simplest to most complex methods, are reviewed in some detail.

The purpose of this paper is to provide an overview of, and some guidance as to when, how and why to choose and use, a given simulation modelling method; provide some basic insight into their theory and practice in the context of health technology assessment; and expose investigators to methods they might not be familiar with. During this process we need to keep in mind one of the first principles cited by PharmacoEconomics: investigators should choose the modelling methods that fit the problem, not the other way around, and this choice should be able to be described and defended explicitly according to the problem being examined. The target audiences for this paper are investigators new to simulation modelling or investigators unfamiliar with a particular method. Because of this trade off, slightly more emphasis will be devoted to the most commonly used methods. We assume that readers will be coming from a wide range of fields and that their level of experience will also range widely.

1. Basic Concepts

1.1 Defining Terms

We use simulation and modelling to help us understand the world. Unfortunately, the terms, ‘modelling’ and ‘simulation’ are frequently used loosely and interchangeably, often leading to confusion and argument. Each expert domain (e.g. epidemiology, statistics, decision science, computer science, biology, psychology, physiology, engineering, economics, physics, law, etc.) uses these terms to point to different concepts. Therefore, the first step in presenting an overview and guide to modelling and simulation methods for health technology assessment is to define terms.

For the purposes of this paper, we will define ‘model’ as referring to a simplified representation of reality that captures some of that reality’s essential properties and relationships (e.g. logical, quantitative, cause/effect). Models in turn may be grouped according to how we use them to gain understanding. In this paper, we use the term ‘simulation’ to refer to models in which an actual or proposed system is replaced by a functioning or interactive representation of the system under study as opposed to purely conceptual models such as mathematical formulae. The modelling methods we consider here are primarily those that can be instantiated with computer software rather than physical or hardware models.

1.2 What is Technology?

Just as the terms ‘model’ and ‘simulation’ are often used loosely, leading to confusion, so is the term ‘technology’. It is common to assume that we are all talking about the same thing. However, technology can refer to either tools used for industrial or commercial objectives or as the scientific method and materials used to achieve a commercial or industrial objective. Sometimes both definitions may be used in the same sentence. In the healthcare technology assessment context, we are usually comparing tools aimed at achieving some diagnostic or therapeutic purpose. For example, one may compare groups of antihypertensive or cholesterol-lowering drugs, different kinds of coronary artery stents, diagnostic imaging tools or even management strategies and health delivery systems. Each of these studies compares different technologies aiming to achieve the same goal.

The purpose of technology is to facilitate the implementation of the desires of the system’s users. To evaluate a technology, one needs a measure of its...
effect. Every technology, new or old, operates in the context of a web of other technologies. From a systems perspective, introducing any new technology into a system changes how the system behaves, by changing the effectiveness of the user in a desired or undesired direction. Outcomes are measures of system behaviour. In simulation models that seek to replicate clinical trials, these outcomes may represent clinical measures of morbidity, mortality or cost. In those that seek to represent system behaviour, we need measures of system behaviour such as flow time, wait time, throughput and resource utilization. Measures of system behaviour are particularly important when we look at methods that attempt to capture dynamic and interactive systems.

1.3 Why Simulation Models?

We create models and simulations (i) when direct experimentation is impossible; (ii) in order to better understand and predict the world or system (real or hypothetical) that we are examining in terms we can comprehend; and (iii) to aid in decision making.

The clearest indication for using simulation methods is when direct experimentation is not possible. This usually results from problems related to cost, time or ethics. For example, a randomized controlled trial (RCT) to answer a question may exceed a project’s budget once the costs of staff, recruiting and tracking patients, etc, are included. It should be noted that an RCT is a form of simulation according to the above definition. This is because RCTs are simplified yet functioning representations of the system under study, with presumably the same cause and effect relationships as the real system. A computer simulation analysis typically costs far less, often requiring only analyst time and software.

Clinical trials also typically take a substantial amount of time. It is not uncommon for an RCT or other types of prospective studies to require several years to complete, often too long for decision makers to make decisions in a timely fashion. Simulation analyses, while often time consuming and labour intensive for the analyst, typically require less time to develop evidence for decision making than clinical trials, particularly if based on an already established model.

Computer simulation also has a role when a question needs exploring, but conducting a trial is unethical. Such problems may occur because a live experiment might require denying care to a group (e.g. Tuskegee Syphilis Study, research in developing countries, etc.) or exposing a group to unacceptable risks (e.g. radiation exposure).

Simulation models are methodical frameworks for formally synthesizing our best available evidence about how the world works. Unlike many other scientific activities, which tend to be reductionist in nature, simulations are almost always synthetic, bringing together knowledge and data from many diverse sources. Unlike meta-analyses, which also synthesize data from multiple sources, simulations are not constrained to a single type of information or a single outcome. To construct a simulation model requires putting into a single framework the best available information and knowledge about the strategy or structure of the system being studied, the outcomes of interest, and the risks, rates and probabilities affecting each action. This approach is very similar to how real-world decision makers act, except that it is more formal and subject to more rigorous testing.

Simulations help generate evidence for or against our hypotheses and help investigators understand the nature of the ‘real’ or hypothetical system under study. Simulations may also suggest new theories, models and hypotheses, based on a systematic exploration of a model’s behaviour under both normal and abnormal conditions. When we vary the parameters of our simulation model it is called ‘sensitivity analysis’ when we vary elements of the model’s structure we are conducting ‘what if’ experiments.

Ultimately, we use simulation to aid our decision making by helping us to better understand the consequences of our actions when we must make decisions under conditions of uncertainty. We use it to evaluate the outcomes of different strategies, to explore the consequences of changes to the system and to predict how the behaviour of the system will change over time. Ultimately, simulation can help inform and possibly persuade decision makers to make the best decisions possible.
1.4 Cost Effectiveness of Simulation

We frequently use simulation modelling methods to evaluate the cost effectiveness of different clinical and policy strategies, but how cost effective is simulation itself? Simulation has been estimated to cost <5% of the total time and budget of a given project, and to have a very rapid and large return on investment, on the order of 1000%.\(^1\) This benefit has been seen in settings as diverse as surgical environments to national water policy making.\(^12-18\) There are several reasons for this benefit. Primarily, simulation allows designers and policy makers to make mistakes and work out design errors on the simulation model rather than in the actual system. This allows developers to identify and eliminate problems that would have gone unnoticed until implementation of the system or strategy, minimizing over-design and unnecessary capital investments, and helping discover solutions for operational inefficiencies.

2. Going DEEPP (Describe, Evaluate, Explore, Predict and Persuade)

The simulation process can be compared to a conversation between the simulationist, the system under study and the audience for the analyses. A simple mnemonic for the steps in this process is DEEPP (Describe, Evaluate, Explore, Predict and Persuade):

- **Describe** the system under study. This entails framing and characterizing the system being examined in terms appropriate for the stakeholders and in terms that can be translated into the simulation tool.
- **Evaluate** the consequences of a given strategy or set of strategies. This typically involves performing sensitivity, cost-effectiveness and -benefit analyses on the system.
- **Explore** the model to suggest new theories and identify gaps in knowledge.
- **Predict** the future and forecast behaviour of systems.
- **Persuade** decision makers through consensus building and evidence.

2.1 Steps in Building a Simulation Model

Before going DEEPP you must first build your simulation model. Building a simulation model is an iterative process that requires analytic and synthetic skills and intuition. However, the first and most crucial step is always framing the question. Framing a question forces the investigator to articulate his or her problem in a way that is relevant, focused and answerable. The most common mistake new simulation modellers make is framing their questions too broadly, often resulting in interesting but unanswerable questions. In order to make a question answerable, one must choose the appropriate outcomes (the measures by which we understand and relate the end results to the particular healthcare practices, interventions and innovations being examined). The broad criteria for choosing appropriate outcomes are as follows: (i) outcomes that are causally linked to the question; and (ii) outcomes that are or can be derived from existing measures. The framed question must then be translated or incorporated into a specific simulation method. While this sounds straightforward, it is typically a lengthy process of refinement and discovery and depends in part on the ‘vocabulary’ of the modelling system chosen. Finally, once the simulation model is satisfactory to the modeller and stakeholders in the question, it must then be verified and validated.

2.2 Simulation Model Validation

Simulation models need to be verified and/or validated before they can be believed or used for decision making. Verification and validation are the processes by which we determine the ‘truth’ of our model. This topic is too large to be considered in detail in this paper. However, these terms are often used interchangeably, leading to confusion, and should be defined. In the computer programming culture from which these terms primarily entered the healthcare simulation culture, verification means ‘does the software conform to its specification’ and validation means ‘does the software do what it is supposed to do’. From the perspective of the healthcare simulationist, the verification process asks whether or not the logic of the model is internally consistent, and validation asks whether or not the model is consistent with the real-world clinical sys-
Modelling methods for Pharmacoeconomics and HTA 135

...tem being studied. Complicating matters somewhat is the empiricist perspective, another influence on the culture of healthcare science. From this perspective, verification often refers to testing a concept against a measurable external reality, reversing the use of the terms.

Alternative and possibly more transparent definitions are internal, external and ecological validity. Internal validity refers to whether or not the simulation model satisfactorily demonstrates a causal relationship between two or more variables internal to the model. External validity refers to the ability of the model to be consistent across different ‘experimental’ settings and populations. Ecological validity refers to the ability of the model to approximate the behaviour of the ‘real’ world. Content and face validity or prima facie validity refers to the ability of the model to capture all the relevant information that is important to the stakeholders or decision makers. This often derives from intuition, expert opinion or animation and is important when trying to get stakeholder ‘buy in’. While this type of validation may be sufficient for local stakeholder buy-in, it is not sufficient unto itself for formal validation purposes. Calibration, in contrast with validity, usually refers to the process of adjusting the model inputs so that it agrees with a predefined output within a specified accuracy. Simulation model verification or internal validation typically proceeds by systematically exploring the behaviour of the model through sensitivity analysis (discussed in the following section).

3. Sensitivity Analysis

Sensitivity analysis is the procedure in which the assumptions underlying the model are challenged and the variables representing those assumptions are systematically varied. This process allows one to determine to what variables the decision strategy or system modelled is sensitive. A decision may be considered sensitive to a variable if changing it within the plausibly defined range results in a change in which strategy is favoured (e.g. changing from surgical to medical therapy).

There are three common types of sensitivity analyses in use, and each addresses different types of uncertainty: (i) first-order; (ii) second-order; and (iii) probabilistic sensitivity analysis.

First-order sensitivity analysis usually refers to testing the underlying uncertainty surrounding the model assumptions (e.g. elements such as disease prevalence or distribution in a population). Therefore, conducting a first-order sensitivity analysis usually means examining the stability of the model outcome when systematically varying the distributions or parameters defining the composition of the population being examined.

Second-order sensitivity analysis usually refers to the uncertainty surrounding the parameters within the model itself. Second-order sensitivity analysis, systematically varying the value of a parameter in a simulation model and re-evaluating the model with each incremental change, is what people usually refer to when they loosely speak of sensitivity analysis (generally, conducting both first- and second-order sensitivity analyses simultaneously will result in introducing unnecessary variance into the analysis). When one parameter is varied in this way, it is referred to as one-way sensitivity analysis; when ‘n’ parameters are varied, it is referred to as n-way sensitivity analysis. One-way sensitivity analysis describes a line between strategies, marking a threshold over which one strategy is preferred to another. When two variables are varied simultaneously in threshold analysis, the result is a two-dimensional region – and area – on a graph describing a preferred range of solutions. When three variables are varied simultaneously in threshold analysis, the result is a three-dimensional region – a preferred volume. The more variables being varied, the harder it becomes to visually describe results. The limitations of n-type sensitivity analysis are primarily in their presentation. The higher the number of variables being tested against each other, the more difficult it is for decision analysts to explain and to make summary statements, which decision makers (as against decision analysts) often desire.

Probabilistic sensitivity analysis usually refers to the simultaneous consideration of all the uncertainties surrounding the variables in the model, usually as stochastic distributions in a Monte Carlo model (see below). Probabilistic sensitivity analysis allows easier summary statements about the uncertainty and the distribution of outcomes of the strategy. However, it is less useful for determining specific variables that the strategy is sensitive to, bottlenecks
or other specific system constraints. Finally, just as in n-way sensitivity analysis, caution must be taken to ensure that the model inputs being examined are not correlated and giving a false impression of causality. Therefore, before running multi-dimensional sensitivity analysis, it is good technique to examine your model inputs using standard statistical techniques for correlation or, if the data allows, use causal analysis to identify the key variable driving correlated behaviour.

All sensitivity analytic methods examine the likelihood of an event or outcome of a modelled strategy. We can then address the question of evidence by asking under what conditions is the model stable and under what conditions is it not, providing us with a sense of the strength of the evidence generated by the model.

4. Simulation Models and the Kind of Evidence They Provide for Decision Making

Science generates evidence through observation, deduction and induction. We use deduction to derive theorems from assumptions, and induction to find patterns in observed data. Simulation, like deduction, starts with specified sets of assumptions regarding an actual or proposed system. However, unlike deduction, simulation does not prove theorems through generalization. Instead, simulation generates data suitable for analysis by induction. However, unlike typical induction, the simulated data come from controlled experiments (sensitivity analyses and ‘what-if’ experiments) rather than from direct observation of the real world. As a result, simulation acts as an intermediate way between deduction and induction for understanding systems and the world.

Generating evidence via simulation models also requires some understanding of Bayes’ theorem. From the Bayesian perspective, evidence for a hypothesis is developed when one substantially increases or decreases the posterior probability (the probability after a test) of a hypothesis relative to the prior probability (the probability before the test) [see equation 1].

\[
P(\text{posterior}) = \frac{P(\text{event given the hypothesis}) \times P(\text{prior})}{P(\text{event})}
\]  
(Eq. 1)

In other words, the posterior probability is the new probability of a hypothesis given new evidence. This new probability is directly proportional to the probability of the evidence given the hypothesis and the prior probability. It is inversely proportional to the probability of the evidence. How this relates to simulation models is that the probability of the evidence given the hypothesis is typically built directly into the structure of the simulation model itself, for example, in a decision tree’s structure. The prior probability is usually framed as the base case. The base case usually refers to the set of parameters for the models that are most commonly found. The probability of the event itself can be varied through sensitivity analysis.

5. Choosing the Most Appropriate Simulation Method

There have been several efforts in the past to categorize simulation methods and describe how the various methods relate to one another.\textsuperscript{[28,29]} In the following sections, we attempt to provide some guidance on when and why to choose a given method for your project and provide a basic understanding of the theory and practice of the methods.

When choosing a simulation method one needs to keep in mind several criteria (figure 1 provides an algorithm for choosing a modelling approach).

5.1 Project Type

Will the model be used to answer a single question one time or will it be used in a programmatic fashion (e.g. a long-term research project such as a PhD thesis, long-term independent research projects, etc.)? If it is to be used only on a single occasion, almost any simulation method the modeller is familiar with (simple trees, mathematical models, etc.) may be used. The guiding principle should be “Keep It Simple Stupid” (KISS [attributed to Apollo mission control]). Therefore, for single occasion models, the simulation model should only be complex enough to answer the specific question, for example, individual level or bedside decision analyses.\textsuperscript{[30]}

If the simulation model is to be used for programmatic reasons then models that can handle more complexity are usually required.
Fig. 1. A decision algorithm for choosing a simulation method. **DEEPP** = Describe, Evaluate, Explore, Predict and Persuade; **DES** = discrete event simulation; **KISS** = Keep It Simple Stupid.
5.2 Population Resolution

Does the question require looking at a population in aggregate or is it necessary to simulate at the individual level? Most simulation methods can handle populations in the aggregate either directly as a cohort or as a summary of the behaviour of individuals in the model. If the problem requires examining behaviour at an individual level, then methods such as Monte Carlo Markov models, discrete event simulation (DES) and agent-based models are necessary. These modelling approaches are discussed in section 6.

5.3 Interactivity/Feedback

Are interdependencies or feedback systems, or re-entrant paths important to the question? For example, modelling an epidemic might require modelling the interaction between infected, exposed and unexposed groups. In this case, consider methods such as system dynamics, DES or agent-based simulations.

5.4 Treatment of Time

Is time treated cumulatively and/or instantaneously as in simple trees or mathematical models? Are changes over time important to the study question? For example, is the time of therapy or surveillance interval important, as considered in Markov, DES or agent-based simulation models?

5.5 Treatment of Space

Is location important to the study question or may it be treated in summary? For example, what is the optimal distribution of emergency medical services in a geographic area, as considered in DES or agent-based simulation models?

5.6 Resource Constraints

Is it important to model limited resources and waiting lists, as considered in system dynamic, DES or agent-based simulation models?

5.7 Autonomy/Freedom of Action of Simulated Entities/Populations

How much freedom of action is important? Are all potential paths through the modelled system predefined (e.g. an idealized RCT)? Are absorbing states required (e.g. Markovs)? Or are multiple courses of action open to the population being modelled, as considered in DES or agent-based simulation models?

5.8 Embedding of Knowledge

Where is the knowledge best embedded? In the structure of the model (e.g. simple trees)? Or is it best to have as much of the knowledge in the model built into the population (e.g. DES or agent-based models)?

6. Types of Methods

The types of computer simulation modelling methods commonly used in health technology assessment are decision trees, Markov (cohort or Monte Carlo), DES and system dynamic and agent-based models. However, it should be noted that new simulation methods and applications are being developed all the time, so this paper cannot be all inclusive. The remainder of this paper will give an overview of the more commonly used methods, touch on their underlying theories, their pros and cons; and where they have been applied, may be applied and are best applied. We will present and organize these methods as a progression of increasingly complex figures.

6.1 Mathematical and Statistical Models

Mathematical models use mathematical language to describe the behaviour of a system. A mathematical model describes a system using a set of variables and equations that describe the relationships between the specified variables. These models may be linear or nonlinear, deterministic or probabilistic, static or time dependent. These models often treat time cumulatively but also may be used to describe a particular time interval. These models may be considered a form of a directed graph of a single dimension (see figure 2) and are most commonly used in epidemiologic studies. The literature on the use of regression and statistical models and their uses in healthcare is enormous and is beyond the scope of

Fig. 2. A directed graph of a single dimension.
this paper. Basic methods may be found in any
standard biostatistics textbook. However, over time,
as the problems become more complex, mathemat-
ic modelling is being paired with simulation meth-
ods to model problems such as cancer care and other
healthcare problems.[38-40] Thus, we will concentrate
in the rest of the paper on models involving comput-
er simulation methods.

6.2 Decision Trees

Decision trees (figure 3) are structurally the next
simplest directed graph and are among the earli-
est[41,42] and most widely used. They are simple
directed graphs without recursion. While decision
trees are primarily a formal way of describing deci-
sions, they can be instantiated as computer models
in many commonly used software packages (winDMTM, TreeAgeTM).

Simple decision trees embody the central para-
digm of decision analysis. Specifically, all decisions
may be broken down into three broadly defined
components: (i) the decision node, the formal repres-
entation of the moment in time when a decision
maker makes a choice between competing strate-
gies; (ii) the decision strategy, a specific strategy set
or programme of actions or events consequent to a
decision (in simple decision trees, this is incorpor-
at ed as a series of chance nodes and/or Boolean nodes
representing the resulting specific events conse-
quent to making a given strategic choice); and
(iii) the outcome nodes, the terminal branches of the
tree that represent the value of the outcomes of the
strategy. There is no decision if there are no values
assigned to the outcomes. In healthcare technology
assessment, the outcome node usually represents life
expectancy (LE), cost, QALYs or quality-adjusted
life-expectancy (QALE).

Classically, the expected value of a decision tree
is calculated by ‘averaging out’ or ‘folding back’ the
branches of the tree. Simple trees may also be evalu-
ated stochastically, by sending single entities down
the various branches. The expected value for the
entity is the path probability to the terminal node for
the entity, multiplied by the value of the terminal
node. The value of each strategy is the sum of all the
individual values for that strategy.

When simple trees are used, one makes the as-
sumption that the population being examined can be
modelled in the aggregate and that the aggregate
population is large enough that any fractional out-
comes reflect the same proportion of people in the
population being studied. If a decision tree is being
applied to an individual clinical choice,[30] the as-
sumption is that aggregate probabilities are relevant
to the individual. A finite time-frame, or time-hori-
zon, describing the period of time over which the
consequences of the decision are played out is also
typically assumed and is the period over which all
outcomes are aggregated.

Examining the uncertainty around decisions us-
ing these models and determining internal validity is
done through sensitivity analyses, typically second
order (see section 3).

6.3 Markov Family

Markov models (see figure 4), first developed by
Andrei Markov (1856–1922), are partially cyclic
directed graphs. They are particularly useful when a
decision problem involves exposure to risks or
events over time, ongoing exposures or situations
where the specific timing of an event is regarded as
important or uncertain, or where describing the tim-
ing of events is necessary for face validity. For
example, it is much more intuitive to assume that a
person’s cancer risk changes over time the longer
they smoke. Most Markov models used in health-
care are semi-Markov state transition models. Unlike Markov Chains,[43] where state transitions are constant over time and may be solved analytically, in semi-Markov models, state transitions may be allowed to vary or be time-variant and usually need to be solved numerically via simulation.

The major assumptions of semi-Markov models relate to state, time and memory. Patients may exist in one and only one of a finite number of health states called Markov states.[44] For example, if you were to use this method to examine diabetes mellitus, you might model the health states ‘well’, ‘pre-diabetic’, ‘diabetic without insulin’, ‘diabetic with insulin’ or ‘dead’. Simulated cohorts or individuals cannot carry a history of their disease or health. All this knowledge must be embedded in the structure of the model. Every potential health state and potential transition of interest must be explicitly modelled. Overlapping health states are not permissible. The number and choice of health states depends on what is important to your problem and what information is available to describe them. If a person carries two diseases, each permutation of the two sets of disease states describing the disease must be modelled explicitly. If it is thought that a patient treated for stroke with a thrombolytic agent carries a different immediate risk of death than one treated with angioplasty, both situations must be explicitly represented. This unitary state requirement forces simplifying assumptions about what health states are of interest or model complexity may become impractically large.

Markov models are typically discrete-time models, although continuous time versions may also be built.[39] This means that time is divided into discrete increments usually referred to as Markov cycles. At the end of each cycle, the patient may make a transition from one state to another. This may be represented by a partially cyclic directed graph (see figure 4). The arcs or edges connecting two states represent the allowed transitions at the end of a given Markov cycle. Recursive arcs represent a patient transitioning back to his/her current state; this is equivalent to remaining in a health state. It is assumed that people, individuals in a Monte Carlo simulation or fractions of a cohort in a cohort simulation, can make only one transition per Markov cycle. The directed graph is restricted in that only certain predefined transitions are allowed. For example, people do not return to ‘well’ from the ‘dead’ state. ‘Dead’ is what is called an ‘absorbing’ state, from which patients presumably cannot return. Markov models run until they either reach the time horizon of the model (e.g. 10 simulated years), or all entities are in an absorbing state.

The length of a Markov cycle is ideally chosen to be the shortest clinically meaningful time interval. For example, a model of cancer incidence may choose a cycle time of a year, whereas a model examining the consequences of thrombolytic therapy for an acute stroke may have a cycle representing months.[45]

The central Markovian assumption, and constraint, is lack of memory. The distribution of entities at all future states at time \(t+1\) is estimated solely from the distribution of entities at time \(t\). No information is needed or stored in memory. All events of interest are explicitly represented as transitions from one state to another. The net probability of making a transition from one state to another during a single cycle is called the transition probability. One may calculate the instantaneous probability of an event occurring within a fixed period of time such as a Markov cycle as \(p = 1 - e^{-rt}\) where \(r\) is the hazard rate and \(t\) is the time interval. For a Markov model of \(n\) states, there will be \(n^2\) transition probabilities. When these probabilities are constant with respect to time, they can be represented by an \(n \times n\) matrix. Probabilities representing disallowed transitions will be zero. This matrix forms the basis for the fundamental solution of Markov chains.[46]

### 6.3.1 Mimicking Memory

Two methods have been developed to mimic memory in Markov models: temporary and tunnel states and tracker variables.[44]

Temporary and tunnel states are a special class of states that only point to other states and not themselves. These states serve as a structural side-path representing a temporary change in behaviour or as a sort of predefined structural memory. This would be useful, for example, in a model examining post-myocardial infarction (MI) patients where you may want a patient or fraction of a cohort to remain temporarily at an elevated risk for a repeat MI for 6 months. Or if you wish to model the effectiveness of
treatment for influenza in a population where a fraction of the population receive a vaccine each year and the vaccine only works for that season. Patients who undergo the changed risk are shunted to different parts of the model until such time as the conditional change is over.

Tracker variables are global variables that serve as a temporary memory allowing the modeller to track (e.g. output) or modify (e.g. risk) the value of variables during individual runs of Monte Carlo Markov models.\[47,48\]

### 6.3.2 Cohort Models

Semi-Markov simulations can be used to model populations at both the aggregate and individual level (micro-simulation). In Markov cohort simulations, a hypothetical cohort/troop of patients is distributed across the starting states at the beginning of the process. At the end of each cycle, the fraction of the cohort initially in each state is partitioned among all states according to the transition probabilities. This results in a new distribution of the cohort among the various states for the subsequent cycle. The utility accrued for the cycle is referred to as the cycle sum, which is the sum of the fraction of the cohort in a health state multiplied by the utility of that health state across all health states in the model. The cycle sum is added to a running total, referred to as the cumulative utility. The model runs until all fractions of the cohort are in the absorbing state or until the time horizon of the model is reached. Time horizon refers to the time period in which the problem is framed. For example, a problem examining the prevention of MI through cholesterol reduction may have a lifetime time horizon, whereas a problem examining a new hospital guideline to reduce iatrogenic infections may have a time horizon of a few months. The expected utility is the cumulative utility at the end of the run divided by the total size of the original cohort. When performing cost-effectiveness analyses with these simulation models, a separate incremental cost may be specified for each state, representing the financial cost of being in that state for the duration of the cycle. The models are usually then evaluated simultaneously for cost and survival. This information may in turn be used for cost-effectiveness analyses.

However, it should be kept in mind that cohort models give the analyst average outcomes. Whether or not this average output, subject to sensitivity analyses, is sufficient to meet the demands of the decision makers depends on the nature of the question. The central question, as with most methods, is ‘does the simulation model capture enough of the underlying relationships to be able to discriminate between strategies with sufficient confidence?’.

Simpler models often provide greater insight into the important underlying relationships that the investigator is examining and remove some of the obfuscation inherent in over-detailed models. However, if the nature of the problem demands a more probabilistic answer, then Monte Carlo Markov simulation may be more useful.

### 6.3.3 Monte Carlo Markov Simulation

For Monte Carlo Markov simulation (first popularized in the 1940s for modelling atomic processes),\[49\] the path through the model and outcomes for an individual are determined by random draws at each chance node. The path probability is then multiplied by the outcome value and the outcome value is then calculated. This process is then rerun for large groups of individuals to obtain stable estimates of the variability surrounding each outcome and the expected utility. Metaphorically, this is similar to photons in diffraction gradient or balls moving through a Pachinko machine.

### 6.3.4 Markov Decision Processes

A Markov Decision Process (MDP) [see figure 5] is a way to describe and analyse sequential decisions under conditions of uncertainty. These are structurally similar to Markov Chains except that the transi-

Fig. 5. A Markov decision process. HS = health state; n = cycle number; t = time.
tion matrix depends on the actions or policy of the decision maker at each time increment. MDP models are also known as sequential stochastic optimizations, discrete-time stochastic control processes, stochastic dynamic programming, etc. Another way to think of MDPs is as a series of stochastic decision trees where the output of the former is the input of the latter. The decision-maker’s goal is to define a policy/strategy where the output of the series of decisions is optimized. This has been an area of intense research in the field of operations research since the 1960s. MPD has been applied over a wide range of areas such as inventory control, finance, communications networks, water reservoir management, game theory, machine learning and recently, healthcare.

In MDPs, the actions of the decision maker are specified by a ‘policy’ that specifies the response to every possible circumstance/state of the system at the end of each time increment for the entire time horizon of the problem. The choice of action, as articulated in the ‘policy’ depends on the ‘reward’ (i.e. the net benefit of the previous iteration of the decision).

Each action has a reward structure, such as the expected value of the decision as modelled in a stochastic Markov model. At time \( t_n + 1 \), the state of the system changes. This change is probabilistic and Markovian, that is, the new state of the system is based solely on the state of the system at time \( t_n \). The goal for the MDP developer is to find a policy that maximizes a reward or optimizes among several parameters such as life expectancy and cost. The MDP developer may take advantage of the fact that the decision rules can be dynamically programmed to change over time. Alternatively, the rules may be changed based on feedback loops, based on the reward output at the end of each cycle, to help in the search for optimal solutions. A typical search for an optimal policy may proceed as shown in equation 2:

\[
\text{Policy (time}_t\text{)} \rightarrow \text{Evaluation (time}_t\text{)} \rightarrow \text{Improvement (time}_t\text{)} \rightarrow \ldots
\]

\[
\text{Policy (time}_t\text{+1)} \rightarrow \text{Evaluation (time}_t\text{+1)} \rightarrow \text{Improvement (time}_t\text{+1)} \rightarrow \ldots
\]

(Eq. 2)

A special case of MDP is when the decision rule policy is stationary, meaning that the decision rules remain constant for all time. An MDP under such a Markov Policy is a Markov Chain (see section 6.3). The range of optimization methods of these sequential decision processes is a field unto itself and beyond the scope of this paper.

One of the key assumptions made in MDP is that the decision maker is expected to choose a predefined decision policy across the whole time horizon of the problem. This is not how most real-world decision makers behave. Dynamic programming can help identify optimal decision strategies/policies and are more consistent with the behaviour of real-world decision makers (i.e. adapting strategies as new information becomes available). However, these same adaptive methods, using for instance, neural nets, may obscure the rational for a given set of rules, which may defeat the purpose of using models to gain insight into the underlying problem. Finally, there is currently limited off-the-shelf software ready to use for less sophisticated users.

6.3.5 Limitations of Markov Models

There are two main limitations of Markov models. The first is that state transitions can only occur at the end of a cycle, which can create some biases. For example, patients may be given credit for living a cycle they didn’t live through or be denied credit for a cycle they did. In order to better approximate the continuous nature of time, it is good practice to incorporate a half-cycle correction in calculating utilities, no matter what the cycle length. While not perfect, this creates an assumption that transitions occur in the middle of the cycle. This has some prima facie credibility and better approximates calculated survival curves. Second, Markov cycle time may force the analyst to make simplifying assumptions regarding transition probabilities.

6.4 Influence Diagrams and Causal Inference

Influence diagrams (see figure 6) are a way of representing decision problems that are mathematically equivalent to decision trees and Markov models. These models also comprise decision, chance and outcome or value nodes. However, unlike decision trees (which are open graphs) and Markov models (which are partially cyclic graphs), influence diagrams are closed, directed graphs without recursion.
Modelling methods for Pharmacoeconomics and HTA

Influence diagrams are one method of examining causal inference. Causal inference is a method for understanding the causal relationships between factors (such as hypertension and exercise capacity) and a given outcome (such as survival). The aim of standard frequentist statistical analysis is to infer parameters of a distribution (e.g. the probability distribution describing a person’s transition from health state 1 to health state 2 per month) from samples drawn from the raw data. With such information, one can estimate the past and future likelihood of events, and assuming static conditions, infer associations among variables. Causal analysis goes one step further and aims to infer aspects of the data generation process and examine the dynamics of events under changing conditions. One can predict the effects of interventions (e.g. treatments or policy decisions), spontaneous changes (e.g. epidemics or natural disasters) and attribution (e.g. whether event x was necessary or sufficient for the occurrence of event y). One of the main strengths of causal analysis is unearthing confounding factors that may not be detectable by standard methods. Graph theory methods, such as diagrams, directed graphs, networks and structural equations are often used to both describe and analyse the relationships between causes and effects.

6.5 System Dynamics and Compartment Models

System dynamics models and compartment models are simulation methods where compartments or population pools are related to each other through a set of differential and algebraic equations. Structurally, these may be described as heavily recursive directed graphs (see figure 7). System dynamics as a field and method was initially developed for studying and managing systems with complex feedback loops in business and social systems. Inter-related differential equations allow these models to explicitly capture complex feedback systems.

In healthcare, system dynamics models have most often been used to model broad population-level questions such as trends in cardiovascular dis-

![Fig. 6. An influence diagram. QALE = quality-adjusted life-expectancy.](image-url)
Continuous time models such as those used in system dynamics models may be solved cumulatively or for any given time moment in time. Continuous time Markov chain models can address some of the same issues.

6.6 Discrete-Event Simulation (DES)

DES is a very flexible modelling method in which entities may interact or compete with each other for resources in a system. Graphically (see figure 8), DES models may be considered complex networks. Every interaction between entities (with each other or with the resources in the system) is an event. Every interaction changes the state of the entity involved and of the system as a whole. The time between events may be handled probabilistically, using fixed time increments, or both, depending on the nature of the system being modelled. There are generally four approaches for managing events in DES platforms: the process-interaction method, event scheduling, activity scanning and the three phase methods. The differences are in how the software reacts to or anticipates interactions in the system.

The key concepts in DES are entities, attributes, queues and resources.

Entities are objects. They can move or be static within the system. They have the ability to interact with other entities; they represent people, places and things and so, metaphorically, act like nouns. The types of objects represented are not constrained to physical objects. For example, entities may also represent packages of information such as phone calls, e-mails or chemical signals. DES packages have been primarily written in object-oriented computer programming (OOP) languages, and entities may be considered to represent a class of objects.

A resource is an entity or facility that provides a service to a dynamic entity. The number of entities they can serve simultaneously is the resource capacity. For example, a bank with a single cashier can serve one person at a time. A bank with three tellers
can serve up to three customers simultaneously. A mobile resource such as a motorcycle can transport one person, whereas a school bus can transport 40. Providing a service requires time. If a resource is occupied when a new entity seeks its use, the new entity must wait until the resource is free.

A queue is any place or list in which an entity waits if the resource is already occupied when they arrive seeking service from that resource. Queues have logic. For example, the line at a cashier may follow First In/First Out (FIFO) logic, getting on or off an airplane may follow Last In/First Out (LIFO) logic and the waiting room in an emergency department or the waiting list for transplant may follow Highest Value First (HVFF) logic.

DES explicitly embeds queuing theory. The simplest queuing model is the M/M/1 (Kendall’s nomenclature), which translates as Markovian inter-arrival time/Markovian process time and one server. Simple systems such as this may be solved analytically and give insight into the behaviour of more complex systems that cannot be analytically solved.

DES also explicitly allows the modeller to embed memory into the model. Attributes are variables local to the entity object. This means that each entity may carry information with them describing, for example, their age, gender, race, health state, etc. This information may be modified during any interaction within the system and may be used to determine how an entity will respond to a given set of circumstances. In DES, much of the information driving changes in the state of the model are embedded in the entities themselves in the form of attributes. This is in contrast with other modelling methods (e.g. trees, Markovs, etc.) where the information and knowledge in the model is embedded in the structure of the model itself. As a result, entities in DES potentially have many more degrees of freedom in how they transit the system being modelled.

In addition to the standard outputs such as QALYs and cost, DES can also provide operational outcome measures such as throughput, utilization, flow time and wait time.

### 6.6.1 DES Software

There are many software packages available for conducting DES. However, most of these are custom-built for a specific purpose. The Institute for Operations Research and the Management Sciences provides an extensive list of vendors on their website. However, some of the most commonly used general purpose DES packages are GPSS®, Arena®, SIMAN®, AutoMod®, Extend®, ProModel®, Simul8® and Witness®. There is also freeware available on the Internet, although this generally requires more computing skill to use.

### 6.7 Agent-Based Simulations

Agent-based simulations or models may be considered independent multi-agent discrete-event simulations. Agents in this context are entities/objects that contain information about their state, and decision rules on how to communicate and interact with other agents or their simulated environment. Agents can range from active data-gathering decision makers with sophisticated learning capabilities to passive environmental phenomena with no learning function. For example, agents have been used to represent individuals (patients, voters), biologic entities (foxes and rabbits) and environmental entities (weather and geography). This latter aspect allows agent-based simulations to be closely linked to geographic data. As in most object-oriented platforms, agents may also comprise other agents, permitting hierarchical construction such as social groupings and institutions. Agent-based modelling methods have been used to examine economics, physics, emergent behaviour, learning, markets, institutional design, networks, etc.

Modelling the influence of agents upon each other and the influence of environment on the agents is the raison d’être of agent-based simulations. From simple rules governing individual actions and communication emerges complex behaviour, which informs our understanding of complex systems.

A typical agent-based simulation framework is much closer to the OOP paradigm (familiar in C++, JAVA, etc.) than most other simulation methods. In an agent-based simulation framework, classes of objects are defined prior to building a simulation just as in an OOP language. In many packages, several basic object classes are preset and can then be modified. The most common are the ‘agent’ and ‘model’ classes.

Agents are essentially objects that can act on other objects. The agent class defines the behaviour
of agents/entities. This may include the decision rules that the entity uses when interacting with other entities or resources, the types of information shared and how it is shared. This may also include how agents adapt over time or learn from experience. Model classes define the conditions of the system and how the model is run.

There are many software packages available to create agent-based models. Two of the oldest and most widely available are Swarm®, developed at the Santa Fe Institute,[73] and RePast from the University of Michigan and the University of Chicago.[74] As mentioned above, these packages typically provide pre-built object classes for the user. However, most of the currently available packages are still fairly close to the underlying OOP language and will require real programming skills to apply.

These packages allow the modelling of agent–agent communication, so they can also be used in social network analysis problems. Social network analysis is a method for mapping and analysing inter-agent relationships and communication. Methods have evolved out of the fields of sociology and anthropology and are used in areas ranging from organizational science to information science. Traditionally, this mapping and analysis is independent of any interactions with the environment or resources, and its main aim has been to identify and define the connectedness between individuals within a network or society. These individuals may be people, groups, computers, etc. The network informs how information is spread and used. Individuals in a network are described by their centrality, which can be defined in terms of degrees (the number of direct connections a node has), ‘betweenness’ (the degree to which an individual lies between other individuals in a network) and centrality (the inverse sum of the shortest distances to everyone else in the network). Social network analysis has been used in healthcare to examine epidemiologic questions such as spread of obesity and contact tracing for hospital-acquired infections and the spread of HIV in confined populations.[75-77]

Social network analysis is highly relevant to health technology policy questions that involve the use of pharmacological or organizational interventions that are aimed at mitigating the spread of disease. This method can also help analysts better understand how the effectiveness of their proposed intervention might be modified by real-world behaviour. An example of a widely used software package solely for social network analysis is UCInet.[78]

Agent-based models provide a means of understanding why system behaviours have evolved, emerged or propagated from simpler behaviours, why some behaviours remain stable and others collapse, how the physical or social environment influences behaviour and vice versa. They also allow investigators to try out policies on self-motivated populations.

7. Conclusions

The aim of this paper has been to provide an overview and guide to simulation methods used in health technology assessment and to expose investigators to methods they might not be familiar with. This review will undoubtedly seem over-technical to some and not technical enough for others, or more or less comprehensive. More time has been spent on methods most often used. However, it is hoped that it provides a useful starting point for people new to simulation modelling, or for experienced modellers new to a particular methodology.

In the final analysis, we use simulation to aid our decision making by helping us better understand the consequences of our choices. We do this by describing the system we are interested in, evaluating the consequences of a given strategy or set of strategies (e.g. cost-effectiveness analysis), using the model to suggest new theories and identify the gaps in our knowledge, and predicting and forecasting the future behaviour of the systems in which we are interested. By using the model to build consensus and evidence, simulation models may help persuade decision makers to make the best possible choices.

When using simulation methods, care must be taken in framing the question, and in choosing outcomes and the most appropriate method. Choosing the simulation method best suited to the problem is important with regard to efficiency and transparency, and reflects our understanding of the problem. This knowledge must develop out of conversations with decision stakeholders. In turn, we can then direct how the problem is framed and how the
outcomes and the simulation methods will be chosen.

Acknowledgements

No sources of funding were used to assist in the preparation of this review. The author has no conflicts of interest that are directly relevant to the content of this review.

References

65. Taylor K, Lane D. Simulation applied to health services: opportunities for applying the system dynamics approach. J Health Serv Res Policy 1998; 3: 226-32

Correspondence: Dr James E. Stahl, MGH – Institute for Technology Assessment, Massachusetts General Hospital, 101 Meriam Street, 10th floor, Boston, MA 02114, USA. E-mail: James@mgh-ita.org