A tutorial on discrete-event simulation for health policy design and decision making: Optimizing pediatric ultrasound screening for hip dysplasia as an illustration

Sabrina Ramwadhdoebe\textsuperscript{a,*}, Erik Buskens\textsuperscript{b,c,1}, Ralph J.B. Sakkers\textsuperscript{a,1}, James E. Stahl\textsuperscript{d}

\textsuperscript{a} University Medical Center Utrecht, Department of Orthopaedics, Netherlands
\textsuperscript{b} University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, Netherlands
\textsuperscript{c} University Medical Center Groningen, University of Groningen, Department of Epidemiology, Netherlands
\textsuperscript{d} Massachusetts General Hospital, MGH-Institute for Technology Assessment, United States

\section*{A R T I C L E  I N F O}

\textbf{Keywords:}
Discrete-event simulation
Infant screening
Operations research
Regional health planning

\section*{A B S T R A C T}

\textbf{Background:} It is increasingly recognized that healthcare is a complex system with limited resources and many interacting sources of both positive and negative feedback. Discrete-event simulation (DES) is a tool that readily accommodates questions of capacity planning, throughput management and interacting resources. As a result the use of DES in informing healthcare decision making is increasing. However, understanding when and how to build a DES model and use it for policy making is not yet a common knowledge.

\textbf{Methods:} The steps in building a DES model will be demonstrated using a real-world example, i.e., pediatric ultrasound screening for hip dysplasia. The main components of a DES model such as entities, resources and queues will be introduced and we will examine questions such as referral schedule, number of ultrasound machines and type of screeners and how these entities interact. Finally a review of the statistical techniques appropriate to DES will be provided.

\textbf{Conclusion:} Discrete-event simulation is a valuable tool in the policymakers armentarium. It can be used effectively to analyze and understand complex healthcare systems and policy problems such as population screening.

© 2009 Elsevier Ireland Ltd. All rights reserved.

\section*{1. Introduction}

Simulation is any activity where an actual or proposed system is replaced by a functioning model that approximates the same cause and effect relationship of the "real" system. It is a tool of investigation that is most appropriately used when for reasons of cost, ethics or feasibility real-world trials and experiments cannot be conducted [1]. We use simulation best to generate evidence and support for decisions and policy making or to generate understanding of processes when actual experimentation is not possible. The most common types of computer simulation models used to inform health care decision making have been decision trees and Markov models [2].

Discrete-event simulation (DES) traces its origins to the field of operations research where it had been primarily used in industrial planning [5]. As the method has evolved it has been applied to a more and more diverse settings and has recently made inroads into healthcare. For example cardiovascular diseases [3,4], screening [6–8], public health and policy [9,10], cost effectiveness studies [11–14], pediatrics [16,17] and epidemiology [15]. DES differs from trees and Markov's in several ways. Most importantly it is one of
the few methods that allow entities within a system, e.g., patients to interact and compete with each other. For example, two or more end-stage liver patients may compete for a donor liver when one becomes available [18]. In addition, unlike most Markov models, the timing of each interaction can be independent of fixed length Markov cycles and completely stochastic. Each interaction between entities can create a change in the state of the system. DES is a method best used when a decision strategy or the system being modeled involves competition for resources, the timing of events is not known a priori and when examining the interdependence between events or the flow of information or entities (e.g., patients) is important.

While DES is beginning to be used for health care decision and policy making [17,19–21] when and how to apply DES, as opposed to other methods, is not well known to the healthcare community. This article describes a clinical example and explains where and why DES is used and walks through the steps from problem conception to working model to analysis. In this process we will look closely at the components of a DES model and how they are assembled for use in health care policy making.

We will proceed in 4 steps: first, the conceptual model and problem narrative, second the framing of the question, third, building the actual DES model and fourth statistical analysis appropriate to DES. Throughout this paper we will describe and illustrate each step with examples based on our ongoing study on the implementation possibilities of ultrasound screening for developmental dysplasia of the hip (DDH) in the Netherlands.

2. Step 1: conceptual model and problem narrative

A common first step in building a DES model is to create a problem narrative or conceptual model. The process of creating a narrative description of the problem lets the problem stakeholders and decision makers provide direct input into the model and (structurally) validate what the modeler proposes to build. It is important in this collaborative process to incorporate sufficient detail to satisfy the needs of the modeler (e.g., number of resources in the system, probability of certain events, causal relationships and costs) and the needs of the stakeholders, reflecting their understanding and intuition about the problem. Once complete, the narrative model is then instantiated as a process map of the system and serves as a reference point for further discussion and refining the problem description.

2.1. Example: pediatric ultrasound screening for hip dysplasia

"Is implementing ultrasound screening for developmental dysplasia at infant health care centers (IHCs) in the Netherlands feasible and cost-effective?". Screening for developmental dysplasia of the hip (DDH) is performed in many countries [22]. Early detection helps avoid progressive pain and loss of function and ultimately total hip replacement later in life. Currently, in the Netherlands, screening is performed at the IHCs via physical exam. However, alternatively screening with hip ultrasound at the age of 3 months may significantly lower the number of false positives and be cost-effective [23,24]. The primary stakeholders in this question are the national child health service, health policy makers, the parents of the children, the children and clinicians (nurses and physicians in IHCs). Each year approximately 180,000 children are born in the Netherlands. Each IHC serves a fixed catchment area of children. IHCs are open only during normal office hours. Hip ultrasound presently is reserved for children referred by their IHC physician because of abnormal physical findings. The optimal time for a hip ultrasound for screening is in children 3 and 4 months of age. After 4 months, the likelihood of getting the child to undergo a hip ultrasound drops substantially. Before 3 months of age the sensitivity and specificity of hip ultrasound is poor. Several strategies to increase access to hip ultrasound screening have been implemented or are currently being explored. One alternative strategy being tested in two regions in the Netherlands comprises training local IHC nurses and physicians to conduct the ultrasound screening.
(see Fig. 1). In this strategy, one ultrasound machine (laptop) is available to each region and must circulate among the IHCs. A major difference between this and the baseline national screening program is that all children involved receive hip ultrasound screening as part of their routine 3-month consultation. What remains to be defined is the optimal configuration of these resources. For example, how often should the ultrasound machine be transported to each IHC and what level of skill (accuracy) do the current clinicians need to have for this strategy to be (cost-) effective?

3. Step 2: framing the question

Framing takes the conceptual model of the system and articulates the questions the model needs to address in a meaningful and answerable way. When framing a question using simulation models, one must keep foremost in mind two things. First, pay attention to what is important to the stakeholders. This is in part understood in the process of creating the project narrative. A model generates the kind of output for which it is designed. Second, define relevant outcome measures. The outcome measures largely define the most appropriate type of model to use and how the model should be structured. The model frame should also consider the available resources and tools in the hands of the decision makers, and the time frame in which the question must be answered. Finally, to avoid ambiguity it should be as specific as possible.

3.1. Example: framing the pediatric ultrasound question

The stakeholders needs and interests were to a large extent identified through creating the project narrative. The next step is choosing meaningful outcome measures for the problem. In this case, these are the number of eligible patients screened, time to screening, the screening false positive and false negative rate, short-term and long-term morbidity and cost. Next the problem needs to be constrained to the available resources and time. An analysis that assumes infinite resources will likely be meaningless to decision makers. So the question that needs to be answered here is whether there are there sufficient resources in the Netherlands: human (radiologists, radiographic technicians, nurses and physician man-hours), physical plants (hospitals and/or IHCs), and financially (budget) to perform this screening within the time-window desired for the current patient volume? If so, what is the optimal combination of these resources? What timeframe are the decision makers interested in, e.g., a single year’s budget or the lifetime health of the population being examined. The nature of this problem demands a modeling method that can help answer questions about cost, resource allocation and timing, i.e., a type of problem for which DES is well suited.

4. Step 3: building the DES model

4.1. M/M/1 – the paradigmatic model

The simplest, most constrained and most paradigmatic model for DES is the M/M/1 or single server model. This model represents a system with a single source of arriving entities and a single resource/server for which these entities compete. Both the arrival rate of entities and the time it takes to process these entities are Markovian, i.e., independent and without memory [25,26]. A very simple example is a bank teller with arriving customers. While most DES become rapidly more complex as the number of entities and resources increase and as these components are allowed memory, the M/M/1 provides useful lessons about the behavior of systems that can be applied to more complex models.

4.2. Model components

There are 3 main classes of objects in a DES; entities, resources and queues. Entities are self-contained objects, for example, patients, physicians, organs for transplant, medical records, etc. The entities are the moving parts of the DES model. An easy rule-of-thumb to determine which elements in a system are entities and which are resources in a DES model is to ask the question “who is doing the waiting (for a service)?”

Resources are facilities or entities that provide a service to a dynamic entity, for example, hospital beds, operating rooms, physicians, etc. Resource utilization is defined as the total time a resource is occupied divided by the total time it is scheduled to be available.

Finally, queues are waiting lists that an entity requesting the use of a resource enters if the resource requested is already occupied. An entity waits in the queue until the resource is available. Queues have their own logic and rules called a queue disciplines [26]. For example, a queue may follow a First In/First Out logic as in lines of people waiting for a bank teller, Last in/First out as in passengers getting on and off an airplane or Highest Value First as in emergency room triage. Using this information the optimal number of servers or fixed resources can be calculated. Queuing models can take many forms depending on the number of resources available and what assumptions are made about the inter-arrival rate and service rate [26].

Common inputs are the distribution of times between events, e.g., patient arrivals to a medical service such as ultrasound examination or an emergency department. The time interval for a transition in a discrete-event model is really the event (t0) to event (t1) time. When operationalizing the problem it is convenient to break it up into functional blocks which most DES software platforms readily allow. The first step is to map out each process described in the project narrative, then break these processes into smaller and smaller blocks until one comes to the lowest functional unit in the system. The advantage of operationalizing the system model into the smallest possible components is that it gives detailed control over simulation processes.

Most DES languages are object oriented and come with preset objects with which to help build models, for example, Create, Queue, Seize, Relay, Release (see below). The example that we present was built using Arena/SIMAN, though there are many other DES languages one might use. It is also possible to use generic programming languages such as C++ [27].
4.3. Model inputs

The main model inputs in most health care scenarios will be arriving individual entity objects, usually patients. Entities enter the system through instantiation, i.e., a new instance of the entity object is created. This is usually performed by class of objects dedicated to this task, usually a module named Create or Make. Creation of these entities may occur at either fixed intervals or stochastically, based on some underlying probability distribution. The interval function, \( f(t_{0}−t_{1}) \), in which these entities are created is usually chosen to fit the nature of the problem and can be derived from historic data or a hypothetical distribution.

Once a new instance of an entity is created it can be assigned attributes, specific to the entity created, for example, a new blank entity may be assigned age, gender and blood type. Attributes can also be used to record the entities’ history in the model, or characterize how the entity will respond to a variety of circumstances within the model and potential be modified in response to interactions in the system.

With DES it is not only possible to assign and create different populations with distinct characteristics, but it is also possible to let them compete for the same resources.

4.3.1. Example: model inputs

When comparing screening or therapeutic strategies, the modeler must be able to examine their effectiveness in different subject populations. In most DES models, the incoming entities, in our case newborns, are the subject population, as well as, the primary input. We can vary the characteristics of the subject population by varying the attributes of the entities as they arrive in the system. As entities enter into the system they are assigned different values for their attributes, such as health state, age, and gender. The value of these variables and attributes can be absolute relative to the model, e.g., time of entry into the system, or come from a distributions specific to the population (discrete [.49 male, .51 female], travel time = lognormal (10, 2)). In our example model these attributes define the risk profile of the children.

The scenarios we initially examine are, first, a general screening scenario where everyone in the population is included regardless of risk profile. Second we look at screening strategies targeting different risk strata. Finally, we also include the populations coming from different environments, e.g., urban, rural, affluent, poor, in the Netherlands. Each of these populations has different risk factors, such as, travel distance, attendance rate/propensity to keep or miss appointments, availability of resources, etc. These attributes can also be used to record or track the experience and the specific entity or a group of entities as it travels through the system. This makes tracking the accumulating costs and benefits to individuals and cohorts easy. The specific timing and characteristics of the entities running through the model can be created in the DES language itself or can be imported from a spreadsheet file.

4.4. Using resources

A resource in DES, typically involves the following sequence of operations: queue, seize (or get), delay and release. The queue is where the entities wait for a service. The seize operation is where the arriving entity obtains exclusive use of a unit of capacity of the resource. The delay operation is the time it takes for the serving process to run its course. The release operation frees the unit of capacity of the resource which the entity had seized making it available for the next entity to seize.

While running the model, the entities may use different resources at different times for different durations. It is possible that a single entity can seize more than one resource at the same time. For example, a patient entering an emergency room will ‘seize’ a bed. In addition, the patient can also seize the attention and services of one or more doctors and/or nurses for varying periods of time, since it is the combination of these resources that provides the relevant service. The patient can then release or retain control of these resources for varying lengths of time. The next patient must wait for the necessary resources to be released.

Just like entities, resources can be assigned individual characteristics. For example, the time it takes to complete a process or how much capacity it has. These can be absolute numbers coming from previous studies or distributions representing the uncertainty underlying certain numbers. Cost per unit time can also be assigned for purposes such as cost-effectiveness analyses.

4.4.1. Example: using resources

There are many instances in this system where entities compete for resources. For example the ultrasound machines, can only be at one IHC at any given time. This means that when it is in use at one IHC people at other IHCs will have to wait. In essence the capacity of the IHC to perform ultrasound screening is 1 when the machine is on site and 0 when it is elsewhere. The rooms in the IHC are also limited resources, not being dedicated solely for ultrasound screening purposes. Other clinical and non-clinical activities compete for the use of the same room. Another example is the screeners who perform the ultrasound examination. To increase this capacity means more training. Training also implies that this resources capacity will change over time assuming the retention rate of new screeners is greater than their retirement. Optimization in this problem means maximizing a conjoint resource capacity at the right time in the right place while minimizing cost. For national implementation it is also important to include in the model the specific population and geographic characteristics of the regions in which these IHCs are located. Different travel times and different numbers of IHCs therefore must be included. Furthermore, when children go to hospital for treatment they consume time, not only from their specialist during the consult, but also time that their parents spend instead of working (see Fig. 2).

4.5. Process pathways

The path that entities follow in a DES is not necessarily known a priori. It can be influenced by random local events,
memory of past events or changes in the system caused by other entities moving through the system. A DES model may also be used to model recursive and random events. Specifically, with DES it is possible to reenter the same path in the same run. The chance for each of the different pathways can be assigned either by stochastic distributions or fixed probability values.

4.5.1. Example: process pathways

In our example there are 4 different pathways (see Fig. 3). In our model children will reenter the same path if parents do not show up for the ultrasound screening, but will attend after a reminder. Note that the reminder itself can be a entity, for example, a letter.

4.5.1.1. The breakdown of the input process in detail. The inputs in our model are the children born in a given region, for example, a region may average 5 children born each day. The inter-arrival time of these children may be described using a random exponential function with a mean of 5/day.

4.5.1.2. The breakdown of the detection process in detail.

(1) Children arriving at the IHC. The screening process starts when children arrive at the IHC. The first part of this process is planning. Children need an appointment in their third month. Planning consists of matching an available building (room), screener and an ultrasound machine to the patient 3 months hence. Once an available time slot or appointment is set, children receive an invitation. The second part of the screening process is the attendance of the child at the screening location. If parents do not show, a new invitation must follow. Resources used are parents’ time (since we use a societal perspective), planning department, transportation costs, travel time, ultrasound machine, screener and room (overhead).

(2) Ultrasound-scan. During the ultrasound-scan consumables such as ultrasound-gel and towels are used. For each appointment 10 min is needed and for this period of time all the above-mentioned resources (room, screener and machine) are simultaneously in use.

(3) Screening result. The screening result can be DDH+, DDH−, or indeterminate. With DDH− no additional action is taken and the patient exits the system. With DDH+ the consult is delayed 5 min as the screener must give an explanation and write a referral letter. Resources released upon completion of the examination are: room, screener, ultrasound machine and parents. If the consult is indeterminate, another consult is made and the child is put back in the scheduling pathway.

(4) Treatment. Treatment is initiated when the screening result is confirmed by the specialist. If the specialists contradicts the initial ultrasound screening result, indicating a false positive at the IHC, an unnecessary consultation and trip to the hospital results. If the screening result is confirmed by the specialist, treatment is initiated. Resources that are used are the specialist, ultrasound-scanner in the hospital, parents’ and treatment costs.

4.5.1.3. The breakdown of the output process in detail. The output for this model is the cost-effectiveness of the policy, i.e., the total cost divided by the number of children detected and treated for DDH. With many simulation packages it is possible to read and write into spreadsheet files or other types of files, readily allowing calculations such as cost-effectiveness.

4.6. Validation

Verification and validation are important aspects of building a (simulation) model. Verification refers to whether the model is performing properly, i.e., does the computer model closely resemble the conceptual model.
Validation refers to whether the model is an accurate representation of the real system. This is checked by comparing the operational model with the real system [28] using current or historical data of a real system. Validation of simulation models of non-existent or proposed systems is less straightforward. One way of validating these models is to test parts of them against sub-systems in existing systems [27]. Another is to check the models behavior against expert consensus.

5. Step 4: statistical analysis/experimentation

A primary purpose for building a simulation model is to use it to experiment and compare alternative scenarios. This allows identifying the optimal scenario as judged by some output criteria. The statistical tests used to compare scenarios are determined by the types of outcome measures used. In DES the most common output types are observational and time-weighted. Observational outputs are equally weighted observations usually related to the entities history. Examples are counts, waiting times, flow times, and inter-arrival times. Time-weighted observations are usually related to the overall state of the system. One can think of time-weighted output as average behavior over a period of time, for example, average queue length, status of resources (busy, idle, etc.) and number of entities in the system over the run-time of the model. When comparing two strategies standard statistical tests, such as t-tests, may be applied.

The equivalent of an experimental sample, used for statistical analysis, in DES models is usually the value of the output measure for a single model run. For example, one can use the number of patients treated or the average queue length at the end of single model run as a sample. These values will vary from run to run as DES models are typically largely stochastic. The length of the run depends in large part on the whether or not the system being modeled is a terminating or non-terminating system. A terminating system is a system with a well-defined beginning and end, for example, an outpatient clinic. This opens in the morning and closes in the evening. A non-terminating system is a system that does not have a clearly defined beginning or end, for example, a 24/7 emergency room.

The appropriate number of runs, the sample size, is dependent on the variability of the observational outputs being examined. However, as in any situation where one uses simulation to compare strategies it is important not to arbitrarily increase your sample size to achieve a likewise arbitrary level of accuracy [25]. With terminating systems data loss will occur since some entities will not complete their run. Therefore, for testing scenarios enough runs must be completed to reduce the chance that distributions of values for a given element could adversely affect a given single run’s result, and therefore, a large number of runs will diminish the outlier effect and produce a more normally distributed set of output indicator data.

Also in simulation model analysis, data from multiple sources are integrated into a single explanatory and functional model. Because of the synthetic nature of simulation models care must be taken in choosing and understanding the quality of the data being used and making clear its source and assumptions. For example, it is usually preferable to use empirical data ahead of expert opinion to obtain probability distributions. Conversely, one of the intrinsic strengths of simulation models is the ability to interpolate missing data. However, the actual relevance of particular data set depends on the accurate modeling of the data measured. The variance and average values likely will be known for a given data element, based on their real-world data distribution. Thus a minimum number of runs will be needed to ensure representativeness. Data on processes that are uncommonly used in a given model will be subject to greater error. This will require either more actual data collected to reduce data dispersion or conversely implies the risk introducing a great deal of variance in the model.

Computer simulation and decision analysis problems are usually optimization problems rather than straightforward and standard hypothesis testing situations. When testing competing scenarios, each of which represent competing hypotheses as to which strategy is better, there are significant similarities to randomized-controlled trials. Determining confidence intervals for these problems, however, is somewhat different and depends on the unit of analysis (i.e., the total output from a single run of the model). The unit of analysis in turn depends on the length of the model run and this depends on the nature of the system: terminating or non-terminating.

Determining the number of model runs needed for statistically significant differences between systems and strategies is empiric. One of the most common methods is to start with the definition of a performance measure, for example, patient flow time or average queue length. Next, this is followed by a pilot study of the model. Just like a real-world pilot study, where a few patients are exposed to the intervention and their behavior is measured, for the DES a small number of model runs is performed and the output measured. This pilot, like a real-world pilot study before a randomized-controlled trial, allows the investigator to estimate the average value of the system output and its variance. This average value and variance can then be used to calculate the sample size (the number of model runs) for your experiment using standard techniques. The number of replications should be kept to the minimum needed to demonstrate differences between strategies, both minimizing computational overhead and the risk false accuracy.

5.1. Example: 2 competing strategies and statistical analysis

To demonstrate simply how DES can be used to examine the outcome(s) of different (competing) scenarios we present a simplified version of our model incorporating the variables travel time, consult time and probability of adherence with the scheduled visit. It also shows the results of a t-test for attendance rate for two competing implementation strategies. The first policy is aimed at reducing the travel time, at the expense of a longer consultation duration, for example, at their local IHC. The second strategy travel time is longer but the consult duration is shorter. For example, where all the parents have to travel to a central screening location. There is a longer travel time but on the other hand less waiting time and consultation time. The
outcome for comparing these strategies is the attendance rate. The model was run 1000 times to evaluate the average attendance rate for both policies. Since we assumed that for nationwide implementation an attendance rate of 85% would be required, a t-test is performed to see which of the strategies attains that goal, i.e., performs significantly above or below 85% attendance. We assumed that parents would not attend if total expected time surpasses 50 min.

The results (Table 1) indicate that a policy using a centralized screening center to reduce overhead will increase travel time and reduce parental attendance.

### Table 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel time (2×)</td>
<td>Normal (10, 5)</td>
<td>Normal (15, 3)</td>
</tr>
<tr>
<td>Wait time</td>
<td>Normal (3, 5)</td>
<td>Normal (3, 3)</td>
</tr>
<tr>
<td>Consult time</td>
<td>Triangular (10, 12, 15)</td>
<td>Triangular (9, 10, 12)</td>
</tr>
<tr>
<td>Number of runs</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Number of children</td>
<td>2500</td>
<td>2500</td>
</tr>
<tr>
<td>Mean attendance rate</td>
<td>95.4</td>
<td>97</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Minimum</td>
<td>94</td>
<td>62</td>
</tr>
<tr>
<td>Maximum</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>Confidence interval 95%</td>
<td>(95.3–95.4)</td>
<td>(79.1–79.8)</td>
</tr>
<tr>
<td>t-Test attendance rate &gt; 85%</td>
<td>7 = 0.0</td>
<td>7 = 1.0</td>
</tr>
</tbody>
</table>

### References


discrete event simulation. Pediatric Emergency Care 2007;23(January (1)):5–10.


