Introduction to Simulation in Healthcare

Thomas Roh
Senior Health Systems Analyst
Outline

➤ Introduction to Simulation

➤ Simulation Development Process

➤ PACU Simulation Model
  ➤ Empirical Inputs

➤ ED Simulation Model
  ➤ Theoretical Inputs
Simulation is a general name for various forecasting disciplines. In healthcare, Simulation centers are set up for education where care providers practice real patient care in an artificial environment. E.g. CPR on human dummies. Another type of simulation is creating a virtual environment through the use of computer software. Pilots can train using flight simulators that mirror real life events and conditions. Wind, rain, speed and its effect on the airplane’s flight can be theoretically calculated. The last type of simulation is mathematical. Data collected for time and other characteristics is used to give an approximation of a future outcome. A real life system is recreated using quantitative methods.
Operations research has a variety of tools each with their own advantages and disadvantages. Queuing theory is a high level model that cannot usually produce the complexity of the real life environment. It requires little data though, often only averages and variances are needed, and can provide good evidence for some simplistic rules. E.g. Splitting a queue into smaller queues increases the waiting times. Linear programming is another high level model that is not typically used in actual application because of its rigidity. Although it has worked well in the past for provider scheduling problems that are based on provider preference and institutional rules. Discrete event simulation is widely used in application because of the fine detail that can be put into the model. It requires large amounts of data to ensure accurate modeling. Stochastic modeling gives the closest approximations to real life, but is extremely complex undertaking and often requires questionable assumptions to be made. Massive amounts of data are needed to build some of these models.
Discrete Event Simulation

- Mathematical representation of a system that allows easy manipulation to study the effects
- Recreate history that closely mirrors past events
  - Assumes past events are a good predictor of future events
- Statistical forecasting tool
  - Fit distributions to account for randomness of demand, service times, etc.
- Graphical model
  - Flow chart of arrivals, processes, and decisions
  - 2-D and 3-D
  - Visual representations of how a system flows
When to Use Simulation

➢ Impossible or too expensive to test scenarios in real system
➢ Need to answer “what if” questions about the system
➢ Justification for improvements to be put into place
➢ Pinpoint problem areas, realize constraining resources
Averages are a standard metric that everyone uses for making decisions, but any time data is transformed, it loses information. Simulation uses variation to account for the lost information. If we simply assume that everyone arrives at identical, discrete time intervals, we are missing the inherent randomness in the real world. The first bullet shows that no patient should ever wait if everything is based on averages. The second bullet and table shows how although the averages are the same as the first bullet, the waiting time is not insignificant when accounting for the randomness of events.

**Why is Simulation Important?**

- **Service System**
  - 5 patients/hour
  - 12 minutes/patient
  - \( \frac{5 \text{ patients}}{\text{hour}} \div \frac{\text{patients}}{12 \text{ minutes}} = 1 \)

- **Includes variability**
  - Average Waiting Time
  - 5.2 minutes

<table>
<thead>
<tr>
<th></th>
<th>Arrival Time</th>
<th>Service Time</th>
<th>Departure Time</th>
<th>Time in System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Patient 2</td>
<td>4</td>
<td>16</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>Patient 3</td>
<td>15</td>
<td>16</td>
<td>44</td>
<td>29</td>
</tr>
<tr>
<td>Patient 4</td>
<td>50</td>
<td>12</td>
<td>62</td>
<td>12</td>
</tr>
<tr>
<td>Patient 5</td>
<td>55</td>
<td>4</td>
<td>66</td>
<td>11</td>
</tr>
</tbody>
</table>
An entity is the object that flows through the model. It can be a patient, worker, bed, device, etc. Most models typically have the patient/customer flowing through the model. It is typically created at the beginning of the simulation.
A resource in a simulation is something that an entity needs to pass through. The resource can service 1-infinity entities at a time. Resources have queues that if the resource is fully utilized can result in waiting times for the entities. Examples are beds, physicians, nurses, equipment, etc.
Queues are waiting places for entities. Queues can have characteristics that mirror real life policies. Examples: the waiting room has a limited number of seats, high acuity patients have priority over low acuity patients, several different types of patients wait in the same area.

**Components - Queue**

- Place where the entities form a line and wait for a resource or other delay

- Queues can have many aspects
  - Priority Rules
  - Limited Size
  - Shared/Parallel
Components - Attributes

- Characteristics of an entity

- Attributes are specific to the entity but entities can be assigned the same attribute value

- Time stamps, age, location, resources used, etc.
A process can be a simple delay or it can be a delay where one or more resources are seized and then released after the work is finished. A process can change an entity or combine entities together.

Components - Process

- An activity that is performed on an object
- Delays the entity in the system
- The time of the activity can be constant, randomly assigned from a distribution, or the output of an equation that possibly references the entity’s attributes
Simulation follows the scientific method as many other quantitative tools do. First, the specific problem needs to be identified and goals and metrics need to be set. The project can then move forward into data collection and early statistical analysis of the data. At this point, one starts conceiving how the model is going to be built, what further information might be needed, and maps out the flow of the system. Building the model is a phase that very entwined with the analysis. The analysis drives some of the things that you can and cannot do in building the model and sets the quantitative guidelines. The other part of model building is the art of reproducing actual processes, rules, behaviors, and policies in an artificial model. After the model is built, the simulation needs to be verified. Verification is checking the coding of the simulation and ensuring that the inputs are producing the outputs the way you want them. Validation is ensuring that the simulation model closely resembles the real life system. If either test fails, you have to go back and change the model. Once an agreed upon model is developed, the simulation is used to answer “what if” questions and output comparable metrics. The results are then presented and feedback is gathered. Usually, new questions arise so new scenarios are tested. Once an acceptable solution is found, implementation begins. After successful implementation, a new problem can be identified and we return to the model building phase.
Most people have done data mining without knowing they are doing it. It is simply looking for patterns, trends, and correlations. It is a precursor to finding causation. Data mining searches for anything that will go together even though they may not be related.

Tools:

- **Linear Regression** is a conservative estimate of the correlation between variables. It assumes that a relationship with a constant increase or decrease.

- **Partitioning groups predictors** based on an outcome. Continuous variables can be joined together e.g. People that weight < 130 lbs. Descriptive variables can be joined together e.g. Triple Bypass Surgery and Quadruple Bypass Surgery.

- **Non-linear Modeling** is similar to linear regression except that relationships can be much more complex. It is closer to real life but more difficult to prove relationships.
The next two graphs will show a brief graphical representation of data mining. The graph here represents a large group of patients (without any exclusions such as age, type of surgery, etc.) who stayed in the PACU. From this graph, one could make many assumptions that may or may not be true.

The PACU is overcrowded in the afternoon.

The patients in the PACU later in the afternoon had more complex surgeries.
Data mining was used to split the patients into two groups based on surgery times. This graph shows the difference in PACU LOS for surgery times over 200 min and below 200 min and shows a comparison between the case volumes. The lower PACU LOS at the beginning of the day is due to case mix. From here it’s easy to conclude, long procedures generally result in greater PACU stays.
The first step in data analysis is to make sure that the data you have is correct. After that the data needs to be “cleaned” into a usable format during which one learns the data limitations and the complexity of the problem. Performance measures are then developed so that the metrics will reflect the direct effect that changes have on the problem.
Discrete event simulation is mostly based on time data—how long a process takes and when a customer arrives. A set of times is considered a distribution. This distribution can be recreated in a simulation model.

Some characteristics that need to be considered before fitting a distribution:

If the data is in a finite form (e.g. all integers), then it is considered discrete. If data is continuous, then there is an infinite number of values e.g. time.

Data that is unbounded has no limitations. It ranges from negative infinity to infinity. Non-negative data is the most common. Data that is double bounded has minimum and maximum caps. When determining a fitting distribution, look at these characteristics to help make a decision.

Theoretical fitting uses random number inputs, probability, and a distribution equation to produce forecast number. Empirical distributions use the actual data in the real model.

Make sure that the data is random that is it each data point comes from the same distribution and data points do not have an effect on other data points in the set.
Here are some common distributions and their uses:

Arrival rates tend to follow the exponential family of distributions. The chance of an arrival usually has high variability and is random enough that a prior event does not have an effect on it. Arrival rates can be separated into specific time units such as months, weeks, days, hours since these units tend to differ from each other.

Service times traditionally follow a normal or lognormal distribution. Using a normal distribution can cause problems because it is boundless so there is always a chance of a generated service time being negative.

Transport times are typically exponential

The Johnson, Weibull, and Pearson distributions best describe a combination of service and transport times.

When data is scarce, the beta, triangular and uniform distributions are used as rough estimates. Often times these are used when the only information available is an expert’s opinion.
After inputting data into analytical software tool, certain statistical tests and graphs can be used to determine if the chosen theoretical distribution is a good fit to the data.

The first graph shows the empirical (real) data divided into segments and presented as a histogram. The curved line is a representation of the theoretical formula for the distribution. The bars and line should closely match as shown in this graph. Note: The size of the histogram intervals can have a large effect on the fit.

The second graph (top left) is a quantile-quantile (Q-Q) plot. The Q-Q plot should have the empirical points close to or on the 45 degree line.

The third graph (bottom right) is a cumulative distribution function comparison of empirical vs. theoretical. The points should closely match the line.

The fourth graph (bottom left) is a percentile-percentile (P-P) plot. The points should closely follow the 45 degree line.

The Q-Q plot accentuates problems at the tails of the fit. The P-P plot draws attention to problems in the middle of the plot.
Goodness of Fit Tests

- **Chi-Square Test**
  - Compute the $\chi^2$ statistic ($\sum_{j=1}^{k} \frac{(\text{Observed Value} - \text{Expected Value})^2}{\text{Expected Value}}$)
  - Biased by selected number of intervals
  - Can be used for discrete and continuous distributions

- **Kolmogorov-Smirnov Test**
  - $D_n = \max\{D^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F'(X_{(i)}) \right\}, D^- = \max_{1 \leq i \leq n} \left\{ F'(X_{(i)}) - \frac{i-1}{n} \right\} \}$
  - Largest vertical distance between the estimated functions and the right continuous stepwise function of $n$.
  - Only valid for continuous distributions
Kolmogorov-Smirnov Test
Creating the simulation to mirror the real life is not an exact science. Build the model as simple as possible because complexity decreases accuracy. Assess your data limitations and the limitations of artificially recreating a real life system. More likely than not, you will have to go back several times to data collection and analysis during the building of the model.

Building the Model

- Model conceptualization is an art not an exact science
  - Improves with experience

- Build the model around what you want to get from the model not from a detailed flow diagram
  - Keep it simple
  - Know your limitations
  - Go back to data collection and analysis if necessary
Verification

- Ensuring the model performs as it is intended to
- Debugging the simulation program
- Check inputs and outputs are properly recording
- Animation is very helpful

Verification is the first step in testing the completed model. Make sure that the code written is performing as intended. Animation is a good tool to make sure that variables are recording and rules are being followed.
The second step to assessing the accuracy of your model is validation. First, compare the data analysis from before to the outputs of the model. Use hypothesis testing and confidence intervals to test discrete outputs such as averages. Use X-Y plots to determine if whole distributions are similar. Second, set up meetings with people who work in the system and go over the testing results. Also, determine whether the model is a good representation of the real life system.
Scenario Testing

- Make sure output makes sense
- Can be paired with an optimization tool
- Trial & Error
- Use confidence intervals

Output should logically make sense when testing different scenarios. Small changes should have small effects. Some simulation software is packaged with an optimization tool. The optimization tool can test a variety of variable combinations and iterate through to the optimal solution. Sometimes, trial and error needs to be used with knowledge of the system helping to make educated guesses. Averages by themselves are not a good comparison. Use confidence intervals or SMORE plots to check for differences.
Implementation

- Work closely with stakeholders throughout project
- Wording and presentation is very important
- Understand most first reactions are defensive
- Be careful to not cater to leaderships expected results

Stakeholder involvement is key to any successful project. Remember that you are selling your ideas and work to the stakeholders. Make yourself visible to the department you are working for and get hands-on experience with the system. Simulation is not an easy concept to understand. Most people label it as a “black box.” Providers will not trust something that they don’t understand. Start with simple graphs and gradually guide people along. Do not be surprised if you are strongly challenged on your work, but do not shape results to the department’s expectations.
St Marys Hospital at the Mayo Clinic Rochester is undertaking a redesign of the current PACU areas. Currently, all patients after their procedure go to an “open-air” PACU unit. Nurses care for 1-4 patients at a time depending on the patient’s acuity. Nurses and supporting staff work as team units using visual cues to triage who needs help and assisting fellow coworkers during high workload times. Mostly, this happens during the initial intake of a patient. The future rooms are designed to enhance patient privacy and provide a place for families to stay in the room. Nurses care for 1-2 patients at a time. The rooms are built in twos and have four walls on each side with a door in between so that a nurse can move between the two rooms easily.
The model is a data-driven input model. The historical patients and their attributes are fed directly into the model. Four different times were used that assumed combined several processes and transportation times. The metrics that were assessed assumed that OR time was the most valuable. Therefore, the PACU needed to be staffed well enough to be able to receive patients from the ORs as soon as the patient’s procedure was completed. The first metric used was the bed usage sampled by half-hours and weekdays. Several different percentiles were assessed to aid decision making. The maximum number of overflow beds needed was used so that patient flow directors would know how many beds outside of the PACU might be needed. The percent time of overloaded capacity was recorded to give an idea of how long the PACU would be overloaded over the course of a day.
The patient is the entity that moves through the model. It is the input created at the beginning of the model and assigned all its attributes from an excel table. The patient then moves through the various activities. The resources seized at each of these activities can be either PED or Adult beds. The PED area is governed by different rules than the Adult area so the resources needed to be split. The PED beds are constrained in this model and the adult beds are not that is there is an unlimited amount of Adult beds. The rule changes include all pre/post op care is now done in the same room. Rooms are not saved during surgery and PED patient will use an Adult bed when the PED area is full.
The above figure shows the difference in patient flow for new vs. old.
The graph above shows Tuesday (busiest day) bed usage over the hours of the day. 60 beds would be required with the new room configurations.
This graph is a breakdown of patients requiring pre-operative care as opposed to post-operative care. Each type requires a different level of care and therefore a different amount of resources.
Given 54 adult beds are available, and 15 are available for PED patients, the PACU’s will be full 25% of the time on Tuesday and Thursday from 7:00-7:30. This results is not surprising since 1st case procedure start times are 7:45, 8:00, and 8:15.
This graph shows the number of patient bed diversions that will be needed when the PACU is overcrowded. The number of extra beds needed is really small.
This model can be used to address a variety of issues.

A 3% growth in patient cases. Since this model uses empirical inputs and surgery scheduling drives demand, one would have to determine the type and length of surgeries that will be added.

Reducing the length of stay of patients in for both pre-operative and post-operative care can be tested to see the savings.

Since the PACU is split into several areas, it would be worth determining which areas to open and close throughout the day.

Very similar to this would be to determine the staffing levels needed.
A very similar project is currently underway for the emergency department at St. Marys. Currently, the ED is separated into 6 areas that each deal with different types of patients. The new ED will create “universal” rooms that can accommodate patients with acuity 2-5. The 54 beds will be separated into 1 PED pod, 1 super pod (open all day), and 3 auxiliary pods. The auxiliary pods can be opened throughout the day in 8-10 hour shifts. It is assumed that one consultant led team will be required to service each pod. The purpose of the model is to match the physician’s schedules to the patient load throughout the day.
The inputs to this model are theoretical. The arrival rates were adjusted for day of the week and hour of the day.

Triage time data was not available so a rough estimate was used for triage time.

Bed LOS distributions were calculated from the bed arrival and departure time stamps. The Bed LOS were divided by acuity type. It is assumed that Bed LOS will not change in the future.

Abandonment rate based on a non-blind queue. The patient makes the decision to leave or wait based on the previous patients waiting time. If the patient decides the wait is too long for him/her, the patient immediately leaves.

Overall average length of stay was also used since timely care both provides a good patient experience and better quality of care.

Patients receiving a bed in under 60 minutes was another agreed upon measure. We assumed that time to bed is closely tied to time to provider.
The patient, once again, is the entity moving through the model. The resources are the PED Beds and the Universal Beds separated into their respective pods. The processes included are triage which seizes from a pool of two triage nurses. The patient characteristics that drive some of the policies include age and acuity. PED patients use PED area if it is not full. Adult patients use the adult pods. Low acuity patients can be taken to a staging area and cared for by a NP/PA during certain hours of the day.
Future State Assumptions

- 45 Universal Beds, 12 PED Beds
- Behavioral Rooms and Trauma Rooms not included
- PED Pod: Open 9AM-1AM
  > Two 8-hour shifts
- Staging Beds: 12PM-12AM
  > Patients spend majority of time in Sub-Waiting Area
- Optimal Balancing
- Bed LOS unchanged
- Patients LWBS at entrance to ED based on previous patient type’s waiting time
- Rapid Triage is 2-8 minutes
  > Two Triage RNs
- Teams work in 8-10 hour shifts
- 3% population increase per year
A primary concern from the physicians was the hours of the day when they were overwhelmed. The physicians had to take on so many patients that they felt it was affecting the timeliness and quality of care. This graph shows given a 3% increase/year the peak overload times with the current staffing model. Up to a quarter of patients will wait for more than 60 minutes during some hours of the day.
2015 Population with Current Staffing and Schedules (same for all DOW)

Sunday, Tues-Friday: Bed Usage with % Patients Waiting > 60 Min.

- Bed Capacity (Tuesday-Friday, Sunday)
- Mean Bed Usage (Tuesday-Friday, Sunday)
- % Patients Waiting > 60 Min. (Sunday, Tuesday-Friday)

Center for the SCIENCE of HEALTH CARE DELIVERY
By moving half a shift from Saturday to Monday and adjusting the start times of pod openings, the % Patients Waiting Time > 60 min curve was able to reduced and smoothed. This should result in less patient overload throughout the day, happier physicians and happier patients.
2015 with Current Staffing Revised Schedule (DOW Adjusted)

Sunday, Tuesday-Friday: Bed Usage with % Patients Waiting > 60 Min.
2015 with Current Staffing Revised Schedule (DOW Adjusted)

Saturday: Bed Usage with % Patients Waiting > 60 Min.

- Bed Capacity (Saturday)
- Mean Bed Usage (Saturday)
- % Patients Waiting Time > 60 Min. (Saturday)
A 3% increase each year is easy to apply across the board for a model with theoretical inputs.

How patients are assigned to each pod can be looked at.

The size of the waiting area needed can be computed based on the queue lengths calculated in the model.

As with the PACU model, the ED model can be used to determine the opening and closing of the pod areas and this can tie into the projected staff needed by time of day and location.
Contact Information

Thomas Roh
Senior Health Systems Analyst
Mayo Clinic, Rochester, MN
roh.thomas@mayo.edu
(507)293-0266
Questions?