Queuing Theory to Guide the Implementation of a Heart Failure Inpatient Registry Program

ADRIAN H. ZAI, MD, PhD, MPH, KIT M. FARR, MD, RICHARD W. GRANT, MD, MPH, ELIZABETH MORT, MD, MPH, TIMOTHY G. FERRIS, MD, MPH, HENRY C. CHUEH, MD, MS

Abstract
Objective: The authors previously implemented an electronic heart failure registry at a large academic hospital to identify heart failure patients and to connect these patients with appropriate discharge services. Despite significant improvements in patient identification and connection rates, time to connection remained high, with an average delay of 3.2 days from the time patients were admitted to the time connections were made. Our objective for this current study was to determine the most effective solution to minimize time to connection.

Design: We used a queuing theory model to simulate 3 different potential solutions to decrease the delay from patient identification to connection with discharge services.

Measurements: The measures included average rate at which patients were being connected to the post discharge heart failure services program, average number of patients in line, and average patient waiting time.

Results: Using queuing theory model simulations, we were able to estimate for our current system the minimum rate at which patients need to be connected (262 patients/mo), the ideal patient arrival rate (174 patients/mo) and the maximal patient arrival rate that could be achieved by adding 1 extra nurse (348 patients/mo).

Conclusions: Our modeling approach was instrumental in helping us characterize key process parameters and estimate the impact of adding staff on the time between identifying patients with heart failure and connecting them with appropriate discharge services.


Introduction
Coordinated Post Discharge Services Reduce Heart Failure Readmissions
Heart failure accounts for an estimated 1 million hospital discharges per year in the United States, with an estimated annual cost in 2008 of $34.8 billion. Hospital readmissions occur in 20–50% of patients with heart failure. These readmissions are one of the principal reasons for the high costs associated with heart failure. Hospital readmission rates have increased since introduction of the Medicare Prospective Payment System, a program that provides incentives for earlier hospital discharges. The increase in readmissions has been attributed to suboptimal assessment of readiness for discharge, fragmented discharge planning, a breakdown in communication and information transfer between hospital-based and community physicians, inadequate post discharge care and follow-up, or some combination of these processes. Multiple interventions have been developed in response to the readmission problem. For example, comprehensive discharge planning and post discharge heart failure disease management programs, including medication counseling and review along with increased communication and follow-up, have been shown to reduce heart failure readmission rates by up to 34%, all cause readmissions by up to 19%, and mortality rates by up to 25%.

Furthermore, multidisciplinary home-based intervention in the population with heart failure have also been shown to be sustained for periods of at least 18 months, resulting in both reduced hospital-based costs and mortality. One example of a successful home-based program is home tele-monitoring, which has been demonstrated to improve quality of care in patients with cardiac disease. However, its impact on heart failure admissions is unclear, with some studies...
showing a positive effect on readmission\textsuperscript{18,22–25} while others do not.\textsuperscript{21,26}

**Initial Implementation of an Electronic Heart Failure Registry Program**

To address increasing readmission rates, our institution established an inpatient heart failure management program called identify and connect to link newly admitted acute heart failure patients to appropriate post discharge services. This program consisted of 3 components: (1) identifying heart failure patients, (2) connecting those patients to post discharge services, and (3) creating quarterly reports as a source of feedback to assess our quality of care. To support this program, we designed and implemented a heart failure registry (HFR) that (1) identified inpatient with heart failure using a prediction algorithm developed using logistic regression, (2) aggregated all clinical and demographic data necessary to make decisions on post discharge services eligibility, and (3) provided real-time reporting for program activities. The variables used in our regression model included the admission diagnosis, the use of inpatient furosemide, angiotensin converting enzyme inhibitor or angiotensin reuptake blocker, admission to a medicine floor, and brain natriuretic peptide (NT-BNP) levels.

Implementing the HFR resulted in a marked increased in number of heart failure admissions identified per month (from 159 before the HFR [Jan–Dec 07] to 448 patients/mo after the HFR [Feb–Jul 08]) and in the number of identified patients who were successfully connected to post discharge heart failure services (from 88 to 209 patients/mo) (Fig 1). However, because of the increase in total number of patients identified, the corresponding proportion of patients successfully connected decreased from 55.2 to 46.6%. Similarly, while the average wait time improved from 14 days (before HFR) to 7 days (after HFR), this reduced wait time still exceeded the average length of stay of many patients, complicating the connection process because most patients needed to be contacted after discharge to arrange their referrals. Here, we present our use of queuing theory, which guided our decisions in redesigning the systems that connect patients with outpatient resources.

**Aligning Resources with Demand Using Queuing Theory**

Queuing theory is an approach to analyzing and modeling processes that involves waiting lines (Fig 2). Effectively applying queuing theory lets managers calculate the optimal supply of fixed resources necessary to meet a variable demand. Examples include the distribution of cars on highways (including traffic jams), data through computer networks, and phone calls through voice networks. Queuing theory is a product of mathematical research that grew largely out of the need to determine the optimum amount of telephone switching equipment required to serve a given area and population. Installing more than the optimum would require excessive capital investment, while installing less than the optimum would mean excessive delays in service.\textsuperscript{27} Now widely used in engineering and industry,\textsuperscript{28} queuing analysis occasionally has been applied to several hospital activities including cardiac care units,\textsuperscript{29} obstetric services,\textsuperscript{30} operating rooms,\textsuperscript{31,32} and emergency departments.\textsuperscript{33} To our knowledge, this method has never been applied to design and implement informatics-based patient registry programs.

**Quality Improvement Question**

Can queuing theory be used to guide the design of an informatics-based heart failure registry program by identifying the optimal amount of development and personnel resources?
Study Methods

Clinical Data from the Initial Heart Failure Registry Implementation

To generate primary clinical data for the queuing models, we compared differences in the period before and after implementing the initial HFR program. We collected data on (1) the arrival rate (λ) of heart failure inpatients needing connections, (2) the rate at which heart failure nurses connect patients to post discharge services (service rate (μ)), (3) the utilization rate (ρ) of the identify-and-serve system (we define the system as “stressed” if ρ > 0.8), (4) the probability for a server (one heart failure nurse, or S = 1) to be busy, (5) the average number of patients in the queue (Rq), and (6) the average waiting time (Twa) before being served.

Specifying the Queuing Theory Model

To assess the overall stress on the identify-and-serve workflow in the initial HFR implementation, we used a queueing theory model with random arrivals and exponential service time to determine the average length of the queue, the average waiting time before being connected, and the probability that the heart failure nurse would be busy.

Our heart failure nurse was treated as one server (S = 1) and a first come, first served queuing discipline was assumed. Such a system, in the queueing literature, is denoted as M/M/S—shorthand notation for systems involving Markovian interarrival times (M/M/S), which are modeled as a Poisson process, Markovian service times (M/M/S), and S servers (M/M/S). A computer simulation model of the identify-and-serve workflow was then constructed using spreadsheet software (Excel 2000; Microsoft Corporation, Redmond, WA, United States) and standard queueing formulas.34

It is crucial to select the proper queueing theory model to solve the problem in question. The main variables of a queueing model to consider when choosing a model are (1) the population size; (2) number of servers; (3) arrival patterns and service patterns; and (4) queue discipline.

Based on the characteristics of these variables, one can identify a specific model that best simulates the observed outcome.35 Examples of simple Markovian queueing models include single-server queues,36 multi server queues,37 queues with truncation,38 and queues with unlimited service.39 To build our queueing theory model, we made the following simplifying assumptions: (1) the arrival process follows a Poisson distribution with exponentially distributed random interarrival times, (2) the service time is an exponentially distributed random variable, and (3) the arrival process and the service process are independent of each other.

The most commonly used models assume that the patient arrival rate can be described by a Poisson distribution, and that the interarrival time, can be described by a negative exponential distribution.35 We confirmed the same behavior with our data by performing a one-sample Kolmogorov–Smirnov Goodness-of-Fit Test40 to the HF patient arrival rate from Feb to Jul of 2008 (a Poisson distribution was not rejected, p = 0.90, Z statistics is 0.571) using SPSS (SPSS software; Chicago, IL). Service time was empirically verified to follow an exponential distribution. Validity of the queueing model was assessed using a correlated inspection approach41 with agreement between observed rates and those predicted by the model assessed via linear regression, and paired t test (SPSS software; Chicago, IL).

Application of the Model to Guide Subsequent Registry Modifications

Before we developed the HFR, we simulated one scenario to answer the following question: what should the average service rate (μ) be if we are to retain the arrival rate (λ), set the utilization rate (ρ) to 80% and number of nurses (S) to 1 (Table 1—pre-HFR: Simulated Case)? Once the model was defined, we then simulated 3 scenarios to answer the following questions in the setting of the HFR: (1) what should the average service rate (μ) be in the setting of the new arrival rate (λ) while keeping the usage rate (ρ) at 80% and one nurse (S = 1) (Table 1—scenario 1)? (2) how many patients with heart failure can a nurse (S = 1) connect if we set the usage rate (ρ) at 80% (Table 2—scenario 2)? and (3) what would happen to the maximal patient arrival capacity, length of the queue, and the time it takes to be served if we hired another heart failure nurse (S = 2) (Table 2—scenario 3)?

Results

Impact on Workflow Efficiency of Implementing the Initial Heart Failure Registry Population Manager

The initial HFR application was deployed on Feb 1, 2008. An updated version with a revised user interface was released on Apr 1, 2008. We defined the predeployment period as Jan 1, 2007, to Dec 31, 2007; the initial post deployment period as Feb 1 to Mar 30, 2008; and the post revision period as Apr 1, 2008, to Jul 31, 2008. We did not count Jan 2008 because users had access to the HFR application for testing purposes even though they did not formally use it for operations.

Before deploying the HFR, the average number of patients receiving post discharge connections (arrivals) was 88/mo. The rate of connecting patients was 90/mo, resulting in a usage rate of 0.98 and an average queue length of 43 patients at any time (Table 1—predeployment of HFR). Patients had to wait an average 14 days before being processed for a connection.

A1. Pre-Heart Failure Registry Simulation Case

In the first simulation, we determined the rate at which patients should be connected to achieve an optimal usage rate of 0.8 if all other conditions remain the same (Table 1—pre-HFR: Simulated Case). We defined the optimal usage rate as 0.8, as the result of observing that the waiting time increased abruptly from hours to days or even weeks (Fig 3) when use of the existing heart failure nurse’s time exceeded 80–85% during any given time. Using a queueing theory model, we calculated that patients would need to be connected at a rate of 110/mo for the average waiting time to decrease to 1 d. Assuming that the number of heart failure patients identified remained constant and the probability that nurse usage would not exceed 0.8, our expectation was that patient waiting time would decrease significantly (to 1 d) if the HFR could achieve this connection rate.

Although we observed a decrease in waiting time (from 14 to 7 d) after deploying the HFR, the drop was not as significant as expected, and heart failure nurse usage rate
**Table 1** Applying Queuing Theory to Assess the Effectiveness of Changes Identifying and Connecting Patients with Heart Failure to Post Discharge Services

<table>
<thead>
<tr>
<th>Date</th>
<th>Comment</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$S$</th>
<th>$\rho$</th>
<th>$P_S(S)$</th>
<th>$R_I$</th>
<th>$T_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-deployment of HFR</td>
<td>Original workflow</td>
<td>88</td>
<td>90*</td>
<td>1</td>
<td>0.978*</td>
<td>0.98*</td>
<td>43*</td>
<td>14 d</td>
</tr>
<tr>
<td>Pre-HFR: Simulated Case</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the required rate of connecting patients ($\mu$) if our goal is to reach optimal utilization rate ($\rho = 0.8$) while keeping arrival rate $\lambda = 88$?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted answer is $\mu = 110$</td>
<td></td>
<td>88</td>
<td>110*</td>
<td>1</td>
<td>0.80</td>
<td>0.8</td>
<td>3.2*</td>
<td>1 d*</td>
</tr>
<tr>
<td>Intervention 1: Post-deployment of HFR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb–Mar 2008</td>
<td>With HFR</td>
<td>209</td>
<td>213*‡</td>
<td>1</td>
<td>0.981*</td>
<td>0.98*</td>
<td>51*</td>
<td>7 d</td>
</tr>
<tr>
<td>Post-HFR: Scenario 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the required rate of connecting patients ($\mu$) if our goal is to reach optimal utilization rate ($\rho = 0.8$) while keeping arrival rate $\lambda = 210$?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted answer is $\mu = 262$</td>
<td></td>
<td>210</td>
<td>262*</td>
<td>1</td>
<td>0.80</td>
<td>0.8*</td>
<td>3.2*</td>
<td>11 h*</td>
</tr>
<tr>
<td>Intervention 2: Post-user interface revision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April–July 2008</td>
<td>With HFR revised</td>
<td>209</td>
<td>218*‡</td>
<td>1</td>
<td>0.959*</td>
<td>0.96*</td>
<td>23*</td>
<td>3.2 d</td>
</tr>
</tbody>
</table>

HFR = heart failure readmission.
$\lambda$: average rate of patients identified as heart failure patients needing connections (per month).
$\mu$: rate of connecting patients (average rate of patients being connected per month).
$S$: number of nurses.
$\rho = \lambda/\mu$: average utilization rate of the heart failure postdischarge connection program.
$P_S(S) = \rho^S S!/(S-p)!$ where $p_0 = [1 + \mu/1! + \mu^2/2! + \ldots + \mu^{S-1}/(S-1)! + \rho^S/(1 + \rho/S)]^{-1}$; probability of a busy period.
$R_I = P_S(S) \rho/(S-p)$: average number of patients in line.
$T_w$: average patient waiting time.
*Numbers calculated are rounded for presentation clarity. Answers in **bold**.
†Stressed system.
‡Post HFR change in rate of connections.
Simulated cases: *italicized*.

remained high (98%). Further analyses revealed that this was the consequence of a marked increase (88–209) in the number of heart failure patients identified per month resulting from implementing the predictive logistic regression algorithm, a component of the HFR that lets one to identify in real time those patients with heart failure. The unanticipated additional heart failure patients identified by the HFR caused the identify-and-connect system to remain under significant stress and to perform less well than expected.

A2. Scenario (1) Increasing Service Rate

Assuming that the increase in the rate of identifying patients with heart failure (arrival rate $\lambda$) continued at approximately 210 patients per month, we determined that for all other variables to remain constant, the heart failure nurse would need to connect at least 262 patients each month (Table 1—Scenario 1).

After revising the user interface of the HFR, we measured the changes attained during the post revision period. We found only a modest increase in the rate that the heart failure nurse could connect patients (from 213 to 218 patients per month), falling short of the goal for optimal efficiency (262 patients per month). As a result, the waiting time only dropped from 7 to 3.2 days instead of to the predicted level of 11 hours. Moreover, the probability that the heart failure nurse would be busy remained high (96%).

The variability in the usage rate was due to a lack of consistency in rates of heart failure admissions (interarrival times) (Fig 4). Taking into consideration the existing variability of the usage rate, we found that the system was working beyond absolute capacity ($\rho > 1$) 42.2% of the time, and beyond ideal capacity ($\rho > 0.8$) 66.7% of the time.

**Table 2** Two Simulated Solutions Using the HFR to Decrease Time to Connection Using a Queuing Theory Model

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Questions</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$S$</th>
<th>$\rho$</th>
<th>$P_S(S)$</th>
<th>$R_I$</th>
<th>$T_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 2</td>
<td>What is the ideal HF patient arrival rate for optimal efficiency?</td>
<td>174*</td>
<td>218</td>
<td>1</td>
<td>0.8*</td>
<td>0.8*</td>
<td>3.2*</td>
<td>13 h*</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>What arrival rates can we accommodate if add 1 extra nurse?</td>
<td>348*</td>
<td>218</td>
<td>2†</td>
<td>0.8†</td>
<td>0.23*</td>
<td>0.15*</td>
<td>19 min*</td>
</tr>
</tbody>
</table>

HFR = heart failure readmission.
$\lambda$: average rate of patients identified as heart failure patients needing connections (per month).
$\mu$: rate of connecting patients (average rate of patients being connected per month).
$S$: number of nurses.
$\rho = \lambda/\mu$: average utilization rate of the heart failure postdischarge connection program.
$P_S(S) = \rho^S S!/(S-p)!$ where $p_0 = [1 + \mu/1! + \mu^2/2! + \ldots + \mu^{S-1}/(S-1)! + \rho^S/(1 + \rho/S)]^{-1}$; probability of a busy period.
$R_I = P_S(S) \rho/(S-p)$: average number of patients in line.
$T_w$: average patient waiting time.
*Numbers calculated are rounded for presentation clarity, answers in **bold**.
†Fixed variable.
Next Steps: Modeling Two Scenarios Using Queuing Theory to Assist Executive Decisions

The challenges posed by marked variability in heart failure admission rates made it unlikely that we could modify the HFR to further enhance connection rates or that doing so would be cost-effective.

To confirm this point, we calculated the optimal patient connection rate needed to maintain efficiency (\( \rho = 0.8 \)) if the arrival rate remained at an average of 210 patients/month (Table 1—Scenario 1). We found that the rate of connection (\( \mu \)), the average number of patients in line (\( R_0 \)), and the average patient waiting time (\( T_w \)) would be 262 patients per month, 3.2 patients, and 11 hours, respectively. Based on the development effort required to achieve a processing rate of 218 patients per month (Table 1—Intervention 2), we concluded that further modifying our information system to reach 262 patients per month was unrealistic. Instead, we used our model to simulate two noninformatics solutions for decision makers to choose from (Table 2—scenario 2 and 3) to maintain optimal system performance.

B1. Scenario (2) Changing Inclusion-Exclusion Criteria

Our current workflow connects heart failure patients to discharge services independent of disease severity, which raises the question of how much benefit there is to connecting patients who may be at low risk of readmission. The HFR will be able to rank patients by heart failure severity, which will let us determine whether the value of post discharge heart failure services varies with readmission risk. Finding a threshold score based on disease severity below which connection does not produce a cost-effective reduction in readmission risk, could be used to further filter the output of the predictive logistic regression algorithm, reducing the rate of arrival (A) and, in turn, the average time to connection.

With the potential of ranking heart failure patients by disease severity, we would like to answer the following question: what is the optimal number of incoming heart failure patients the nurse must process using the HFR to achieve an optimal usage rate (\( \rho \)) of 0.8 (Table 2—Scenario 2)?

In scenario 2, we have 1 nurse (\( S = 1 \)) using the HFR to connect heart failure patients with post discharge services (\( \mu = 218 \)). If we set the usage rate (\( \rho \)) to 0.8, we calculate the average arrival rate (A) to be 174 patients per month. In this scenario, the probability that the nurse is busy is 0.8, the average queue length (\( R_0 \)) is 3.2 patients at any given time, and the average waiting time (\( T_w \)) to receive connection services is 13 hours. Therefore, based on our model, the maximum number of incoming patients the current identify-and-connect system can accommodate is 174 patients per month.

B2. Scenario (3) Increasing Staff

In the current system, our heart failure nurse must dedicate all of her time to connecting patients to preserve the current level of efficiency (\( \mu = 218 \) patients/mo). The current connection process does not involve a face-to-face encounter with the patient because most patients are already discharged by the time the nurse can get to them.

To enhance the quality of our services, our goal is to have the heart failure nurse not only ensure that patients are connected while patients are in-house, but also, to meet with them to discuss their heart-failure–related post discharge plans. We would like a heart failure nurse to spend at least 50% of his or her time seeing patients. Therefore, the significant question is, how many nurses do we need to meet this goal and how busy will they be?

To simulate this scenario, we pose the following specific question to our model: what would the probability be that a nurse would be busy if we added 1 extra nurse (Table 2—Scenario 3)?
If the number of nurses (S) equals 2, the usage rate (ρ) is 0.8, and the average usage rate (μ) equals 218 patients per month, the probability for 1 nurse to be busy drops to 0.23, the average number of patients in line becomes 0.15, and the average waiting time is 19 minutes. Because our goal is to have our heart failure nurse spend 50% of his or her time seeing patients, we predict that adding 1 more nurse will be sufficient for both nurses to connect patients with outpatient services (probability of being busy for each nurse will be P[S] = 0.23) and to spend 50% of their time seeing patients.

Discussion

In complex stochastic dynamic systems such as in-patient heart failure programs, queuing theory offers a simple analytic approach for measuring such things as average waiting time, and average total process time. Applying this technique to our existing HFR system resulted in key quantitative insights that we can use to guide program modifications. One of the key findings was that our heart failure identify-and-connect system was working significantly beyond a reasonable capacity, driven primarily by unrecognized variability in demand.

Applying our queuing theory model led to our learning the following 4 lessons:

To Increase System Efficiency, We Must Minimize Variability of Patient Interarrival Times and Keep the Usage Rate Below 85%

Owing to the stochastic nature of queues, conventional wisdom dictates that a usage rate should not exceed 85%.42 Our findings are consistent with this observation as well as those from McManus et al.,43 in our hospital, usage rates greater than 85% result in rapidly increased waiting times (Fig 2). Therefore, we defined our optimal capacity (p) as 0.8. Minimizing such variability is a best means of attaining peak efficiency. In reality, however, this is a complicated task because it requires us to understand why heart failure patient arrival times are not completely random. This observation is well supported in the literature.44

There are two broad categories of variability that cause fluctuations in patient inter-arrival times: (1) a natural variability, which is based on unpredictable random patterns such as patients’ severity of disease, and their arrival patterns to the hospital; and (2) an artificial variability, which is based on predictable nonrandom patterns such as scheduled admissions. Whenever resources are limited, management of variability becomes critical to the efficiency and effectiveness of a complex system. Natural variability can usually not be eliminated but can be managed using tools and methods developed in the field of operations management.32,45–47 For artificial variability, the best solution is simply to remove the cause of the variability entirely.48 For example, in a study by McManus et al, the authors examined the variability in demand for intensive care unit services, and demonstrated that minimizing variability by controlling for scheduled surgical caseloads significantly reduced the variability of patient inter-arrival times, thereby increasing the efficiency of the intensive care unit.49 In our case with heart failure patients, there is no obvious artificial variability in inter-arrival time patterns that we could easily control to minimize the variability in patient inter-arrival time even though we detected higher admissions during weekdays than weekends.

Common Measures of Usage Are a Poor Proxy for Measuring the Stress on the Workflow of a System

Common measures of usage, such as daily census and number of nurses processing connections, used by themselves fail to capture flow-related stresses in the system; they mask the reality that patients frequently must wait for days before being connected, even if the nurses do not appear to be working at full capacity. A daily census, for example, does not provide you with patients’ inter-arrival variability patterns, nor does it provide you with rates at which patients are processed. Both of these dimensions are essential to understand the effectiveness of the system.

Informatics and Noninformatics Solutions Must Work Together to Achieve Optimal Efficiency

We were able to increase the efficiency of connecting heart failure patients significantly using an informatics solution. However, based on our queuing model, we realized that the efficiency we need to gain to achieve ideal capacity is unrealistic. Based on our model, we concluded that hiring an additional nurse is the appropriate next step. The Pareto’s Principle, or as it is more commonly known, the 80/20 rule, argues that 80% of the output comes from 20% of the input. We believe that the HFR is the 20% solution to achieve an 80% improvement. However, to eventually reach optimal efficiency, we must consider other solutions. Those solutions include noninformatics considerations such as policy changes, resource allocations, and others. The beauty of the queuing model approach is that it provides quantitative results to explicitly estimate how large a gain in efficiency is required to reach optimal usage, thus allowing administrators to make better-informed staffing decisions.

Workflow does not Linearly Correlate with Resource Use

As we learned in scenario 3, adding an extra nurse to a one-nurse system brings the usage rate from 0.98 to a predicted 0.23. This greater than-50% decrease may be higher than expected if our intuition was based on a linear relationship. The stochastic nature of patient flow can easily mislead decision makers. For example, it may appear intuitive that we can achieve optimal efficiency if the rate of patient arrivals equals the rate at which they are processed, or in queuing theory terms p = 1. However, when p = 1, unless arrivals and service are deterministic, and perfectly scheduled, no steady state exists, since randomness will prevent the queue from ever emptying out and allowing the servers to catch up. As a result, the queue will grow without bound. If one knows the average arrival rate and average service rate, the minimal number of parallel servers required to guarantee a steady-state solution can be calculated immediately by finding the smallest S such that \( \lambda/S\mu < 1 \). Preferably, we would like \( \lambda/S\mu = 0.8 \) to achieve ideal capacity. Therefore, if a real world workflow is properly modeled using queuing theory, hospital administrators may be able to accurately estimate the optimal amount of resource required to process the incoming patient flow.

Limitations

While queuing theory enjoys the advantage of being a quick analytic solution, it has several limitations when deriving
such analytic solutions. For example, mathematical models often assume an infinite number of patients; or there may be no bounds on interarrival or service times when it is obvious that these bounds exist in reality. In practice, however, careful assessment of model assumptions and bounds ensures that this queuing model approach provides valid results.

In the case of our heart failure identify-and-connect project, we made several assumptions: only patients destined to be connected were counted as part of the queue. We made this choice because the processing time for patients in the other categories (i.e., “excluded”, “no programs needed”, and “not connected”) is significantly less and is trivial compared with the time required to connect patients. We also assumed that the arrival rates of heart failure patients needing connections followed a Poisson distribution (coefficient of variation = 1), and that the time required to connect a patient is either constant or follows an exponential distribution. This may not be the case because there are usually more admissions during weekdays than during weekends. As demonstrated by Litvak, and associates, large interarrival variability is a major contributing factor to the inefficiency of a given workflow.44

In our case, we conducted a sensitivity analysis for April 2008 based on day-to-day variation in inter-arrival and service time distributions. We found that with 7.9 ± 3.3 heart failure patients arriving daily and an average processing rate of 8.13 patients per day, the usage rate is above the ideal capacity (ρ = 0.8) 57% of the time and above ρ = 0.99 36% of the time. This means that over half the patients have to wait at least 14 hours before being attended by the heart failure nurse, and over a third of the patients have to wait at least 14 days.

Although the M/m/1 model used in our heart failure model may be generalized to similar projects facing similar demand patterns, it is not necessarily applicable to projects that deviate from that pattern. Queuing theory provides other models (e.g., queues with rejections) that can accommodate different demand patterns.

Conclusions

We believe that the stochastic nature of patient flow may falsely lead decision makers to underestimate the resource required for projects with workflows similar to ours. Although building network simulation models may provide a more-precise representation of reality than those based on queuing theory, these models tend to be resource intensive and time-consuming. For information technologies projects involving workflow patterns similar to the one described in this paper, we believe that policy makers will benefit from the added data provided by queuing theory models. In addition to simulating multiple scenarios, we can use variables such as usage rate, queue length, and waiting time as process metrics to evaluate the overall stress level of a particular workflow. In today’s health care climate where we are expected to deliver maximum value using a minimum of resources, we believe that this methodology is another tool we can use to maximize our effectiveness and provide the best possible care to our patients.

References


