Priority Queuing Models for Hospital Intensive Care Units and Impacts to Severe Case Patients

Matthew S. Hagen, MSE[1,2,4], Jeffrey K Jopling, MD, MS[5], Timothy G Buchman, MD, PhD[5], Eva K. Lee, PhD[*,1,,2,3,4]

[1]Center for Operations Research in Medicine and HealthCare,[2] NSF I/UCRC Center for Health Organization Transformation,[3]School of Industrial and Systems Engineering, [4]College of Computing, Georgia Institute of Technology, Atlanta, GA; [5]Emory Center for Critical Care, Emory University School of Medicine, Atlanta, GA.

Abstract

This paper examines several different queuing models for intensive care units (ICU) and the effects on wait times, utilization, return rates, mortalities, and number of patients served. Five separate intensive care units at an urban hospital are analyzed and distributions are fitted for arrivals and service durations. A system-based simulation model is built to capture all possible cases of patient flow after ICU admission. These include mortalities and returns before and after hospital exits. Patients are grouped into 9 di fferent classes that are categorized by severity and length of stay (LOS).

Each queuing model varies by the policies that are permitted and by the order the patients are admitted. The first set of models does not prioritize patients, but examines the advantages of smoothing the operating schedule for elective surgeries. The second set analyzes the differences between prioritizing admissions by expected LOS or patient severity. The last set permits early ICU discharges and conservative and ag gressive bumping policies are contrasted. It was found that prioritizing patients by severity considerably reduced delays for critical cases, but also increased the average waiting time for all patients. Aggressive bumping significantly raised the return and mortality rates, but more conservative methods balance quality and efficiency with lowered wait times without serious consequences.

* Corresponding author: Eva Lee, eva.lee@gatech.edu

Introduction

The current climate of critical care has a heavy challenge to meet growing patient demands while hospital capacity continues to shrink at an alarming rate. According to American Hospital Association, the number of hospital beds has reduced by almost 25 percent in a period of 20 years.[1] Due to Certificate of Need (CON) regulations, an average occupancy level of 85 pe rcent was required before approval to increase capacity.[2] Since, many hospitals had average occupancy below these rates, there was an impression in the health care community that there was excess capacity. For nonprofit hospitals, average rates had reached as low as 66 percent.[15] Consequently, available beds have continued to decrease across states.

In April 2002, a Lewin Group survey reported 62 percent of U.S. hospitals reached or exceeded maximum operating levels. The percentages raised to 79 percent for urban hospitals and 82 pe rcent for level I trauma centers.[11] The Center for Disease Control reported the number of annual emergency department (ED) visits climbed by almost a q uarter for the decade ending in 2002. The number of EDs reduced by 15% for the same period. [29]

Setting hospital capacity by focusing on occupancy levels has led to serious circumstances. There have been access blocks and substantial increases in waiting times.[30] The relationship between waiting time and average occupancy is not linear. At a point, the average delay can start to rise exponentially relative to even small increases in utilization.[16] Wait time depends on the time between arrivals and begin of service. These measures have significant variability, and delays can be considerably different for identical utilization levels. It is not sufficient to only emphasize average occupancy levels when evaluating the process flow of a health care center.

Increasing average wait times for medical care has led to complications that are more significant than economic incentives. Poor patient flow has been found to be associated with elevated mortality rates, longer length-of-stay, and heightened readmission.[4,35] Sprivulis et al. linked ED overcrowding to a 30% relative increase in mortality.[41] Chalfin et al. identified delays to intensive care were correlated with longer lengths of stay and higher mortality.[3] During periods of stress, a decision to admit a patient may not be entirely clinically driven and nurses are prone to medical errors.[24]

Early discharges are more likely at high occupancy levels. The average length of stay can be reduced up to 16% for patients discharged from a busy intensive care unit (ICU).[20] However, the likelihood of returning increases substantially.[10,20,40] KC et al. found overall bounce-back probability was 14%, but

rose to 37.4% for early discharged patients.[20] Higher severity patients are associated with longer revisit stays raising their net total length of stay. These factors effectively reduce hospital's peak capacity, because the readmission loads add unexpected flow related stresses.[25] Readmitted patients have also been found to have higher mortality rates in addition to longer lengths of stays. Snow et al. identified mortality rates for returning patients were 26%, three times the general population for surgical intensive care units.[40] Readmissions from premature discharge can increase costs and lead to overall worsening of medical conditions for patients.[11]

It is essential to improve the process flow of health care centers with motivations that are not purely economic. The demand for intensive care is high. Green et al. determined 90% of ICUs in New York have insufficient capacity to provide proper medical care.[15] While economics tend to favor high occupancy,[13] the quality of care does not. This paper evaluates different priority methods (some from literature, and some we derive) to minimize waiting times for admission to intensive care units. An emphasis is placed on the severity of medical conditions rather than exclusively focusing on market factors. The goal is to maximize the number of patients served while maintaining good quality of care.

Related Work

In a perfect system, all patients would arrive at the same rate and all patients would have the same condition and require identical service time. This system would be 100% efficient as are many automated manufacturing plants.[25] This is not the case in the health care community. Patients arrive unexpectedly with an immense diversity of conditions. Therefore, it is necessary to optimally fit the distributions for patient arrival and service time. In most studies, the inter-arrival times are regarded as a negative exponential distribution.[36] The length of stay (LOS) can have different distributions for different patient types.[21] The fit distributions can vary from exponential, negative exponential, log-normal, or Weibull.[7,30,38,42] Kokangul et al. applied a Kolmogrov-Smirnov test on five years of admissions to a teaching hospital and found arrivals distributed as a Poisson process and LOS distributed as log-normal.[21]

Siddharthan et al. classified patients into emergency and non-emergency care.[39] After collecting data from an emergency department in Florida, patients were grouped as emergency care for major trauma, critical care, minor trauma, and non-critical care cases. Non-emergency care was classified only for primary care patients. 53.3 percent of patients were found to require emergency care and 46.7 percent were non-emergency. The average arrival rate, service rate and waiting time were calculated for both types. The study assumes arrivals follow a Poisson probability distribution and service rate follows an exponential distribution. Using a proper priority queue discipline,[19] it found the average waiting time to reduce by 10 percent for all patients. The queue gave highest priority to emergency care patients, because they had the larger average service time.

Chan et al. utilized a more sophisticated priority queue with 9 categories of patients.[5] Each category was classified by low, medium, or high LOS and by low, medium, or high severity. The severity of each patient was assessed using criteria from Escobar et al. where admission diagnoses and laboratory results were utilized.[12] All groups of patients were tested with three different priority models. The model assumed a patient must be discharged for new arrivals if intensive care units are at full capacity. This is due to the inherent urgency of intensive care. Each priority model enforced the discharge order for patients in intensive care. The three models were based on lowest nominal length-of-stay, smallest probability of readmission, and lowest readmission load. Readmission load is defined as return probability multiplied by average LOS for successive visits. The study results reported the readmission load model outperformed all other priority schemes by up to 10%.

Dobson et al. attempted to accurately estimate the expected number of patients transferred to accommodate more critical arrivals.[9] The study did not use a complex priority scheme compared to Chan et al. Instead, patients were simply discharged by lowest remaining length of stay. A Markov model was utilized to study the effects of ICU workload on patient bumping.

The difficulty of assigning priority to ICU admissions is to correctly identify the severity of incoming patients. Escobar et al. assessed the severity of each patient by assigning the probability of mortality based on sex, age, primary condition and chronic ailments.[12] 16,090 ICD admission diagnoses were grouped into 44 broad categories. Graham et al. used a simpler approach by classifying a diagnosis into "high", "medium", or "low" risk.[14]

Adding to the complications of accurately identifying patient severity, clinicians typically write diagnosis records in free-text format. There have been

successful attempts to use machine learning and natural language techniques to correctly associate notes with hierarchical codes, such as SNOMED-CT® and ICD-9.[6,8,32,33,34,37] However, these methods have been found to have considerably lower performance in data poor cases.[37] More successful results were attained when a large volume of clinical reports, laboratory results and follow-up reports were available.

Regarding strategies for analyzing ICU workflow, Chan et al. only prioritized patients by how they were bumped from the ICU rather than admitted. Discharges were enforced by attempting to minimize readmission load according to several factors, including return probability and LOS. Dobson et al. also prioritized patient transfers from the ICU.[9] They were ordered according to their remaining length of stay. Both of these studies used sophisticated priority schemes, but were not entirely realistic. Patients were automatically admitted when requesting ICU entry by bumping lower priority patients. However, in healthcare settings it is not uncommon for average wait times for an ICU to exceed 4 hours,[28] and bumping patients can cause significant medical complications.[10,20,40]

Methods and Computational Design

Data preparation

32,531 medical records were retrieved from a large urban hospital over a one year period from March 2010 to April 2011. Each record included the patient's id, registration number, diagnosis, and entrance and exit times of each reserved room during the entire hospital stay. Five separate intensive care units were analyzed for this study: Cardiovascular (CV) Surgery, Neurosurgery, Medical, Neuroscience, and Surgical. Since the distribution of LOS may vary among different patient types, [21] the Input Analyzer in Rockwell Arena® 13.5 was used to fit the LOS distribution for each ICU.

Of 5,465 hospital ICU visits, 813 contained a missing entry. 14.8 percent of records included the time a patient exits an ICU room without the time of entry. These offending records were temporarily removed to calculate the LOS distributions for each ICU unit (Table 1). The fitted distributions were then used to sample entrance times for the records with missing entries.

ICU	LOS distribution
CV Surgery	1 + LOGN(84.8, 115)
Neurosurgery	5 + LOGN(55.8, 75.7)
Medical	2 + LOGN(54.4, 53.9)
Neuroscience	4 + LOGN(56.6, 85.4)
Surgical	7 + LOGN(67.6, 101)

Table 1. Length of stay distribution

ICU	Emergency arrival distribution
CV Surgery	GAMM(9.37, 0.948)
Neurosurgery	EXPO(13.8)
Medical	GAMM(8.79, 0.937)
Neuroscience	WEIB(35.9, 1.06)
Surgical	WEIB(8.86, 0.984)

Table 2. Emergency arrival distribution

Arrival rates were calculated after all hospital ICU visits contained complete records for entry and exit. Full lists of entrance times were generated, and distributions were fitted from interarrival times for each ICU. Arrivals were separated by emergency and scheduled surgery admissions (Tables 2, 3). Other statistics were calculated to help identify the process flow of patients through the system. These included return rates after a patient leaves an ICU, after a patient leaves the hospital, and after a patient is forcibly bumped from an ICU. Mortality rates were determined for patients entering an ICU for initial and return visits (Table 4).

ICU	Scheduled arrival Distribution
CV Surgery	493 * BETA(0.554, 2.51)
Neurosurger	WEIB(23.2, 0.695)
Medical	2 + LOGN(2.98e+003, 3.97e+004)
Neuroscienc	WEIB(68.9, 0.883)
Surgical	GAMM(43.7, 0.715)

Table 3. Scheduled arrival distribution

ICU	P(R|r)	P(R|e)	P(M)	P(R|t)	P(M|R)
CV Surgery	0.068	0.039	0.077	0.118	0.226
Neurosurgery	0.034	0.069	0.172	0.111	0.288
Medical	0.073	0.113	0.160	0.056	0.213
Neuroscience	0.039	0.069	0.195	0.250	0.407
Surgical	0.061	0.059	0.075	0.257	0.159

Table 4. Probabilities for ICU returns and mortality. Return probability from room *P(R|r),* return probability from hospital exit *P(R|e),* mortality probability *P(M),* return probability after early discharge *P(R|t),* and mortality probability after return *P(M|R)*

Natural language processing of clinical diagnosis records

The medical records obtained were not comprehensive enough to conduct a full contextual analysis. In particular, the diagnoses from patient records received did not contain standardized codes, such as ICD-9. They were free text entries ranging only up to 54 characters at maximum. This limits text analysis for each record to a few words at most, but it is useful to test the applicability of natural language processing when the content is very minimal.

Due to the difficulty of uniquely matching a patient's diagnosis with minimal content and non-restricted entries, the goal is to instead classify the severity of a patient's condition based on these free text entries. Severity is then calculated by identifying key words shown to have high prevalence in cases of mortality.

Of 2,950 diagnoses, 486 resulted in mortality. The clinical terms used in mortality cases were treated with higher severity. A list of words was generated from all diagnosis records. Another list was produced only from the mortality records. NLTK, a natural language processing toolkit for Python, was used to tokenize the words in each list.[31] It was important to only include words in the English dictionary and remove any common stop words. Wordlist is a corpus included in NLTK that contains 234,943 unique English words, and the English Stopwords corpus contains 127 unique words. These corpora facilitate more significant words to be identified in diagnosis records, but many medical terms may be improperly excluded. It is possible that common words used by clinicians are not included in the standard English dictionary provided by the NLTK library.

SNOMED-CT® is a standardized reference that contains millions of medical concepts developed by the American Pathologists and the United Kingdom's National Health Service.[25] The July 2011 release contained 988,921 unique medical terms. We use this release to augment the list of English words provided by the NLTK corpus. SNOMED-CT was tokenized and stop words were removed using the NLTK library. SNOMED-CT was found to contain 94,581 unique words and when combined with the Wordlist corpus, the union created a joint corpus of 304,760 unique words. This added 69,817 medical words facilitating more content for analysis. With only utilizing the Wordlist corpus, 6,008 words were matched from diagnosis records. The joint SNOMED-CT Wordlist corpus matched 6,535 words increasing the data size by 8.7 percent.

In natural language processing, one of the challenges is to not treat words differently that have identical roots or map to the same stem. The words walking, walker, and walked all map to the stem "walk". Stemming is a process that reduces inflected or derived words to their appropriate root. In this study, the Lancaster stemmer provided by the NLTK toolkit was used. 683 unique words were found from the diagnosis records, and 231 unique words were found by diagnosis mortality records. With applying the Lancaster stemmer, the unique words were reduced to 635 and 222 respectively. The frequencies of each unique word were then calculated.

A severity score could then be calculated by utilizing the word frequency distribution for all diagnosis records and the distribution for mortality records. The TF-IDF score is a weight used in information retrieval. It measures the importance of a term in a document, but it is offset by the frequency the term appears in the entire corpus. The method in this study is not exactly identical to information retrieval, and there are many possible variants of the TF-IDF calculation.[27] The importance of the term is measured by the frequency it appears in all mortality records. It is offset by the frequency it appears in all diagnosis records. Therefore, a higher score will be given to a term that occurs often in mortality records but not often in all diagnosis records. Instead of summing the TF-IDF score for each term, the scores are averaged. This way more benign terms can reduce the severity of the diagnosis. The TF-IDF scores for each diagnosis record were calculated as

$$idf_t = \log \frac{N}{df_t}$$

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

$$Score(q,d) = ave_{t \in q}(tf - idf_{t,d})$$

where N is the number of words in the diagnosis records that were matched with the NLTK Wordlist corpus and the SNOMED-CT corpus, df_t is the term frequency in all diagnosis records, d is the set of words in mortality records, t is the set of words in the current diagnosis record, $tf_{t,d}$ is the term frequency in mortality records, and q is the set of words in the current diagnosis record that exist in d.

Patients were clustered into nine different groups similar to the study by Chan et al.[5] Each group has three possible levels for severity and three possible levels for LOS. LOS level is divided into three ranges by service hours (h). Groups are allocated by LOS < 25h, 25h ≤ LOS < 57h, and LOS ≥ 57h. This resulted in an equal amount of records for each LOS level. After the TF-IDF score was calculated for each diagnosis record, severity groups were clustered using the K-Means algorithm.[26] After K-Means clustering, the ranges for TF-IDF scores for each severity group were TF-IDF < 0.07, 0.07 < TF-IDF < 0.19, and TF-IDF > 0.21. The mortality rate for records in each severity group accurately reflected the average TF-IDF score. The lower severity groups both had mortality rates at roughly 15 percent. Seven percent of entries were classified with highest

severity and were found to have a mortality rate of 46 percent. LOS distributions were then fitted for each severity group (Table 5).

C_S	#	TF-IDF	$P(M_s)$	$P(M_s)$	Expression
1	1,42	0.026	0.143	0.852	1 + LOGN(72,87.4)
2	831	0.11	0.152	0.897	4 + LOGN(49.7, 60.3)
3	174	0.293	0.466	2.755	6 + LOGN(23.7, 19.4)

Table 5. Severity group results from K-Means clustering. Severity group C_S, average *TF-IDF* score for severity group, mortality rate for severity group $P(M_s)$, ratio between group mortality rate and average

After patients were successfully grouped into nine separate classes, multiple statistics were calculated for later use by priority models in the simulation model. These included average initial LOS, return rate and average return LOS (Table 6).

C_P	C_{LOS}	C_S	#	LOS_{Pinit}	$P(R_P)$	$P(R_P)$	LOS_{Pret}
1	1	1	1	19.468	0.075	0.950	41.971
2	1	2	1	19.606	0.074	0.928	28.543
3	1	3	4	19.720	0.065	0.822	44.027
4	2	1	1	39.491	0.076	0.960	54.448
5	2	2	9	38.179	0.041	0.520	50.014
6	2	3	2	34.405	0.091	1.147	44.280
7	3	1	2	159.712	0.124	1.562	134.017
8	3	2	9	149.481	0.041	0.520	182.770
9	3	3	6	110.284	0.167	2.103	124.621

Table 6. Patient group results after clustering. Patient group C_P, LOS group C_{LOS}, severity group C_S, average initial length of stay for patient group LOS_{Pinit}, return rate for patient group $P(R_P)$, ratio between group return rate and average return rate $P(RP)/\mu_R$, and average return length of stay for patient group LOS_{Pret}

Simulation Model

A simulation model was built using Rockwell Arena® 13.5 to aid in the development and evaluation of the process flow of five intensive care units. A separate submodel was created for each ICU: CV Surgery, Neurosurgery, Medical, Neuroscience and Surgical. Each submodel had both scheduled and emergency arrivals. Scheduled arrivals were direct transfers after an appointed operation or surgical procedure, and emergency arrivals were unexpected admissions. The inter-arrival distributions were fitted using the Arena Input Analyzer for both cases (Tables 2, 3).

Different numbers of beds were allocated and a separate queue was designated for each ICU. The full computermodel contains scheduled and emergency arrivals for all five ICUs. Further, each ICU is modeled in detail, including service, queues, clinical and patient workflow, and their inter-dependenices on patient care and resources.

After a patient departs an intensive care unit, they are transferred to an intermediate care room before dismissal. The patient may return to an ICU after transfer to an intermediate room, and they may also return after exiting the hospital. The distributions for LOS in intermediate rooms after ICU discharges were fitted with Input Analyzer. Distributions were also calculated for durations between patient hospital exits and subsequent ICU returns (Table 7).

Location	Expression
Intermediate Room before ICU return	LOGN(157, 294)
Intermediate Room before hospital exit	WEIB(76.4, 0.697)
Outside hospital before ICU return	67 + 8.82e+003 * BETA(0.467, 2.15)

Table 7. Intermediate room and hospital exit distributions

Estimated probabilities from hospital records were utilized in the simulation model. The return and mortality rates were separately calculated for each ICU (Table 4). The Return module in our computerized model captures all possibilities for returns and exits. It also includes mortality cases where patients do not survive their ICU stay.

The simulation model tests six different queuing methods and each is executed in Rockwell Arena® for a period of 90 days with ten replications. The results reported for each queuing model are averages over all replications.

ICU Resource Allocation

The goal of this system is to aggressively test the process flow of the hospital under heavy conditions. The given numbers of beds were approximated for each ICU according to an *M/M/s* queuing model. The model assumes there are s identical servers with unlimited waiting room capacity. Service duration follows an exponential distribution while arrivals occur at a constant rate according to a Poisson process. Given the number of servers s, average arrival rate λ, and average service time $1/\mu$, the mean waiting time in the queue W_q can be calculated under the *M/M/s* model:[18]

$$W_q = L_q / \lambda$$

$$L_q = \frac{\rho}{1-\rho} p_D$$

$$p_D = 1 - \sum_{n=0}^{s-1} p_n$$

$$\rho = \lambda / s\mu$$

$$p_n = \begin{cases} \frac{\lambda}{n!\mu^n} p_0 & (1 \le n \le s) \\ \frac{\lambda^n}{s^{n-s} s!\mu^n} p_0 & (n \ge s) \end{cases}$$

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{(\rho s)^n}{n!} + \frac{\rho^s s^{s+1}}{s!(s-\rho s)} \right]^{-1} \quad \rho < 1$$

where L_q is the mean number of patients in the queue, p_D is the probability that an arrival will experience a delay for service, ρ is the average utilization for the queuing system, and s is the number of servers.

In the 2001 US National Hospital Ambulatory Medical Care Survey (NHAMCS), the average waiting time for an ICU bed reported was approximately 4.1 hours.[28] In this study, the average arrival rate and service duration were determined for each intensive care unit. Using the *M/M/s* model, the average wait times were calculated with the given number of beds for each ICU (Table 8).

ICU	λ	λ_s	μ	s_r	s_e
CV Surgery	0.125	0.013	0.011	18	16
Neurosurgery	0.105	0.034	0.015	20	11
Medical	0.122	0.002	0.017	14	11
Neuroscience	0.042	0.014	0.015	7	6
Surgical	0.140	0.032	0.013	20	15

Table 8. M/M/s Queueing Model parameters for each ICU. Arrival rate (patients/hour) λ, Arrival rate from scheduled surgeries (patients/hour) λ_s, service rate (patients/hour) μ, number of beds in the hospitals s_r, number of beds in simulation model s_e

Parameters in the simulation model are determined empirically so as to match the hospital statistics for ICU admission delay to accurately evaluate the benefits for different test settings. Using the *M/M/s* model, performance measures were calculated for each ICU for different levels of bed availability. Since the M/M/s assumption of exponential service times can lead to underestimating actual congestion,[17] the number of beds selected by the simulation model were associated with mean waiting times between 1.8 – 3.2 hours (Table 9): CV Surgery (16), Neurosurgery (11), Medical (11), Neuroscience (6), Surgical (15)

CV Surgery		Neurosurgery		Medical		Neuroscien		Surgical	
S	W_q	s	W_q	s	W_q	s	W_q	s	W_q
11	705.54	8	52.162	8	60.122	3	267.6	11	413.78
12	51.580	9	14.948	9	15.611	4	23.19	12	42.430
13	18.190	10	5.684	10	5.834	5	5.606	13	15.265
14	7.972	11	2.333*	11	2.396*	6	1.531*	14	6.725
15	3.755	12	0.972	12	1.006	7	0.417*	15	3.170*
16	1.809*	13	0.400	13	0.419			16	1.525
17	0.871	14	0.160	14	0.171**			17	0.733
18	0.414**	15	0.062					18	0.348
		16	0.023					19	0.161
		17	0.008					20	0.073**
		18	0.003						
		19	0.001						
		20	0.000**						

Table 9. Estimated wait times for each ICU using M/M/s Queueing Model. Number of beds s, average wait time (hours) W_q.

* W_q for s used by simulation model
** W_q for s used by the hospital

Classification of severity group

After a patient arrives at the hospital in the simulation model, they are classified into one of nine different groups based on their severity score and LOS. The LOS is generated from the distribution for the requested ICU. There are prior values for the percentage of patients in each severity group. However, the LOS distributions are slightly different for each severity group (Table 5). For example, it is rare to find a patient with high severity and high LOS. It would not be entirely accurate to assign the severity group based only on prior probabilities. Therefore, a p osterior probability is calculated by multiplying the prior probability with the likelihood given a patient's LOS:

$$P(C_s|LOS) = \frac{P(C_s)P(LOS|C_s)}{P(LOS)}$$

$$P(LOS) = \sum_{s \epsilon S} P(C_s)P(LOS|C_s)$$

$$P(LOS|C_s) = P(LOS; \mu_s, \sigma_s)$$

$$P(LOS; \mu_s, \sigma_s) = \frac{1}{LOS\ \sigma_s \sqrt{2\pi}}\ e^{-\frac{(\ln LOS - \mu_s)^2}{2\sigma_s^2}}$$

where C_s is the severity group class, LOS is the sampled value for length of stay from the ICU distribution, $p(C_s|LOS)$ is the posterior probability of

belonging to C_s given the LOS, $p(C_s)$ is the prior probability of belonging to C_s, $p(LOS|C_s)$ is the likelihood of observing the LOS given C_s, μ_s and σ_s are parameters of the log-normal distribution for C_s.

The severity group is assigned to the admitted patient based on the calculated posterior probabilities for each class. Each group has a set of mortality rates determining whether the patient will die during their stay in the ICU (Table 5).

Managing Artificial Variability

There is substantial natural variability in hospital admissions through the emergency department, but there is also artificial variability. In this study, we found that 28.2% of entries were admitted to an ICU from elective surgeries. If adjusted for patient volume, scheduled surgical admissions can vary even more than through the Emergency Department (ED).[23] This can have reciprocal effects where high surgical volumes can delay operations and increase waiting times for an available room. Operations can be cancelled due to a shortage of ICU beds.

In this study, the distribution is calculated for interarrival times to each ICU from scheduled and unscheduled admissions. A Passive model is first tested that uses no priority scheme and factors natural and artificial variability of arrivals. Each model reports the total patients served, severe patients admitted, average waiting times, utilization rate, return rate, and mortalities.

The Smooth Model is similar to the Passive Model, except it uses an ideal surgery schedule where there is no artificial variability. This is to help determine the effects the surgery schedule has on the hospital process flow. The average time between arrivals is calculated for scheduled admissions for each ICU (Table 8). Instead of using the fitted distributions for scheduled admissions, patients arrive at times equidistant from each other for each ICU.

The Smooth Model is not realistic, because even operation times can vary in ideal cases where elective surgeries are scheduled at efficient times. It is only used for evaluation purposes. All subsequent models utilize fitted distributions for scheduled admissions, but test different priority methods for admitting and bumping patients.

Priority models

Typically, a queue admits entries on a first-come-first-serve (FCFS) basis. However, priority queues allow different classes to be treated differently. Without preemption, higher class items can jump ahead of others within the queue. However, service cannot be interrupted for any items in process. In a preemptive priority class, higher class items can discontinue other items currently in service.[16] In this study, both preemptive and non-preemptive models were tested to analyze the process flow of intensive care units.

Four different priority models were evaluated in our simulation model. Specifically, we derive and test models that both restrict and allow bumping while factoring the consequent mortality and return rates.

Greedy: The greedy method[39] gives patients with highest LOS the greatest priority. Using queuing theory, Siddhartan et al showed that admitting patients with larger LOS before others lowered the overall average wait time.[39] The Greedy model is non-preemptive where bumping of patients is not permitted in any case. Higher priority patients in the queue are not permitted to interrupt lower priority patients in service.

Hybrid: The hybrid method admits patients based on their severity and their LOS. A patient in the highest severity group will be admitted first, but patients in the lower severity groups will be ordered according to their average LOS. The Hybrid model is also a non-preemptive method. It factors admission not only on efficiency, but also on the severity of the patient's condition.

The next two priority models are both preemptive. They allow the service of lower priority patients to be interrupted if a higher priority patient is admitted.

Severity (Conservative) Bumping: The Conservative Bumping model is identical to the Hybrid model in the order patients are placed in the queue. However, a severe patient ($C_s = 3$) in the queue can bump a non-severe patient ($C_s < 3$) from service. Non-severe patients cannot bump any patients from service. Non-severe patients are bumped by lowest remaining length of stay plus the associated readmission load:

$$LOS_{tot} = LOS_{rem} + P(R_P) \times LOS_{Pret}$$

where LOS_{rem} is the remaining service time, $P(R_P)$ is the average return rate for the patient group, and LOS_{Pret} is the average service time for returns for the patient group (Table 6), LOS_{tot} is the estimated total service time. The readmission load is the product of

return probability times return LOS, which is calculated using a similar method to the study by Chan et al.[5]

Aggressive Bumping: Severe patients can still not be discharged from the ICU while in service. However, non-severe patients will be bumped when any type of patient requests admission to the ICU. Patients are discharged in the same order as the Conservative Bumping model. Aggressive Bumping is similar to the method used by Chan et al., except severe patients are restricted from ICU transfer before completion of service.

If a patient is bumped while in service, they will have a higher return rate as found in our hospital transfer records data (Table 4). Subsequently, the returned patients also have a higher mortality rate. All four different priority models are tested to determine the effects on waiting time, return rate and mortality.

Results

Table 10 reports the results for all six queuing models. Without enforcing any priorities for admission, the Passive Model reported higher average waiting time in the queue (4.5 hours) and fewer total patients served (1,024). The utilization rate was also 4% higher than any other model.

The Smooth Model also does not enforce priorities, but arrivals from elective surgeries occur at a constant rate. The hospital only schedules surgeries Monday through Friday and operating hours can vary significantly. The Smooth Model is an ideal case that removes all variation from scheduled surgery arrivals. It gave impressive results when compared to the Passive Model at 2.5 hours for average waiting time and 1,035 for total patients served. This raised the amount of patients as well as lowering delays. This showed reducing artificial variability is beneficial if it is possible to enforce a more regimented surgery schedule.

Priority queuing models were tested with artificial variability utilizing the fitted distributions for scheduled surgery arrivals. The Greedy model only prioritizes patients by their expected LOS. It was able to serve 1,043 patients at an average waiting time of 2.95 hours. This model could not capitalize on the benefits of uniform patient arrivals as with the Smooth Model, but it was able to report better performance measures than the Passive Model. The average wait time for the Greedy Model was 0.4 hours higher than the Smooth Model, most likely due to temporary bottlenecks from variation in arrivals.

The Greedy model focuses on efficiency rather than patient severity. The Hybrid Model prioritizes severe patients above all others. Non-severe patients are prioritized by expected LOS identical to the Greedy Model. The Hybrid Model served 1,033 patients at 3.6 hours average waiting time. These are weaker results, but the average waiting time for severe patients was only 1.4 hours compared to 3.8 hours in the Greedy Model. The Hybrid Model also had the lowest return rate at 16.4%.

The Conservative Bumping and Severity Bumping models reported results with substantial differences. Both preemptive queuing models prioritize patients by severity identical to the Hybrid Model. The Conservative model can only bump less severe ($C_s < 3$) patients from service when the most severe ($C_s = 3$) request ICU admission. The Aggressive Model bumps less severe patients from service for any patient requesting admission. The Conservative Bumping model served 1,038 patients and only bumped an average of 7.8 from service. The average waiting time was 0.8 hour for severe patients and 2.7 hours for all patients. The mortality rate was only raised by 0.4 percent compared to the Hybrid Model. The Aggressive Model served 1,051 patients bumping 93 patients with an average waiting time of 1.1 hours. The return rate increased by 1.0 percent and it r eported the highest mortality rate for any model at 8.8%. It is clear that bumping can prove to be beneficial but only in heavily restricted cases.

Model	Priority Order	# Patients	P(R)	P(M)	B	W_q	W_{qs}	Util
Passive		1,024	0.173	0.078	0	4.515	4.556	0.693
Smooth		1,035	0.169	0.075	0	2.560	2.610	0.670
Greedy	7,8,9,4,5,6,3,2,1	1,043	0.169	0.074	0	2.946	3.840	0.660
Hybrid	9,6,3,7,8,4,5,2,1	1,033	0.164	0.075	0	3.562	1.411	0.655
Severity Bumping	*9,6,3,7,8,4,5,2,1*	*1,038*	*0.165*	*0.079*	*7.9*	*2.768*	*0.847*	*0.652*
Aggressive Bumping	9,6,3,7,8,4,5,2,1	1,051	0.174	0.088	93.4	1.062	0.961	0.666

Table 10. Priority Queuing Model Results. Priority order for patient groups, total patients served, return rate *P(R),* mortality probability *P(M)* , number of bumped patients *B*, average waiting time in the queue for all patients W_q, average waiting time in the queue for severe patients *Wqs,* average utilization rate

Conclusion

Healthcare centers that focus on operating at highest efficiency may consequently sacrifice the quality of care. By evaluating several different priority methods, the ICU system-based simulation model helps identify the costs of prioritizing by severity rather than efficiency. Severe priority methods do raise overall waiting times and lower the amount of patients served, but added benefits reduce further medical complications. Shorter wait times for severe patients result in lower return and mortality rates. Severe priority methods can show substantial enhancements by conservatively allowing bumping policies. Permitting early discharges with severe priority models resulted in wait times close to the most efficient models. However, without firm restrictions, bumping can significantly raise the mortality and return rates.

There are several potential future research studies that can be conducted with appropriate types of data. Our approach is applicable to other hospital data streams, for example, ICD diagnosis codes, patient resource needs, and hospital utilization status. Specifically, it would be beneficial to accurately categorize the diagnosis for each patient using individual ICD diagnosis codes. This would help determine if a patient return was due to an early discharge or because of an entirely new condition. Further, in our earlier readmission work[43], hospital resource usage and utilization information were employed to help predict patient readmission characteristics and the impact on patient needs and quality of care.

The studied hospital has five distinctly specialized intensive care units. There may be events where the requested ICU is full and a patient is diverted to an ICU of a different specialty.[22] It would be advantageous to examine the implications regarding permitted diversions for associated conditions. An analysis could be conducted whether patients benefit from diversions to ICUs of different specialties rather than remaining in the queue for the desired location.

Patient admissions can also be evaluated more globally. If estimated wait times were available for each hospital, the costs can be considered for redirecting patients to another hospital. The studied hospital herein has a sister medical center at a location about six miles away. It would be interesting to review records for cases where patients were blocked access and directed to this alternative location. A future study will analyze these cases and determine if transfer times were lower than estimated wait times for direct admission. Even in circumstances where total wait time were reduced by diversion, complications can result from the additional transit time. Optimizing patient flow in healthcare settings is a challenging balance between managing efficiency and maintaining quality of care. Hospitals can become more proficient and resourceful in daily operations by continuing to build system models that attempt to identify and investigate all significant interdependent factors.

Acknowledgement

The study is partially supported by a grant from the National Science Foundation.

References

1. American Hospital Association. 2000. *Hospital Statistics 2000*. Chicago, Ill.: American Hospital Association.
2. Brecher, C., and Spiezio, S. Privatization and Public Hospitals: Choosing Wisely for New York City. New York: Twentieth Century Fund, 1995. Print.
3. Chalfin, D. B., S. Trzeciak, et al. (2007). "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit." Crit Care Med **35**(6): 1477-83.
4. Chan, A., G. Arendts, et al. (2008). "Causes of constraints to patient flow in emergency departments: a comparison between staff perceptions and findings from the Patient Flow Study." Emerg Med Australas **20**(3): 234-40.
5. Chan, Carri W, Nicholas Bambos, and Gabriel J Escobar. "Maximizing Throughput of Hospital Intensive Care Units with Patient Readmissions." *working paper* (2010): 1-41
6. Chien-Hsing Chen, and Chung-Chian Hsu. (2009). "Indexing ICD-9 codes for free-textual clinical diagnosis records by a new ensemble classifier." International Journal of Computational Intelligence in Bioinformatics and Systems Biology 1(2): 177-192.
7. Cochran, J. K., and Bharti, A. (2006). "A multi-stage stochastic methodology for whole hospital bed planning under peak loading." Int J Ind Syst Eng **1**(1/2): 8-36.
8. de Bruijn, L. M., A. Hasman, et al. (1997). "Automatic SNOMED classification--a corpus-based method." Comput Methods Programs Biomed **54**(1-2): 115-22.
9. Dobson, G., H.-H. Lee, et al. (2010). "A Model of ICU Bumping." Oper. Res. **58**(6): 1564-1576.
10. Durbin, C. G., Jr. and R. F. Kopel (1993). "A case-control study of patients readmitted to the intensive care unit." Crit Care Med **21**(10): 1547-53.
11. Emergency department overload: A growing crisis. (April 2002). The Lewin Group analysis of AHA ED and hospital capacity survey data.
12. Escobar, G. J., J. D. Greene, et al. (2008). "Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases." Med Care **46**(3): 232-9.

13. Goetghebeur, M. M., S. Forrest, et al. (2003). "Understanding the underlying drivers of inpatient cost growth: a literature review." Am J Manag Care **9 Spec No 1**: SP3-12.
14. Graham, P. L. and D. A. Cook (2004). "Prediction of risk of death using 30-day outcome: a practical end point for quality auditing in intensive care." Chest **125**(4): 1458-66.
15. Green, L. V. (2002). "How many hospital beds?" Inquiry **39**(4): 400-12.
16. Green, L. "Queueing analysis in healthcare." Patient flow reducing delay in healthcare delivery (2006): 281–307.
17. Green, L. V. 2010. Using Queueing Theory to Alleviate Emergency Department Overcrowding. Wiley Encyclopedia of Operations Research and Management Science.
18. Gross, D., J. F. Shortle, et al. (2011). Fundamentals of Queueing Theory, John Wiley & Sons.
19. Jaiswal, N. K. (1968). Priority queues, Academic Press.
20. Kc, Diwas Singh, and Christian Terwiesch. "An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit." Manufacturing & Service Operations Management (2011): 1-16.
21. Kokangul, A. (2008). "A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit." Comput Methods Programs Biomed **90**(1): 56-65.
22. Kolker, A. (2009). "Process modeling of ICU patient flow: effect of daily load leveling of elective surgeries on ICU diversion." J Med Syst 33(1): 27-40.
23. Litvak, E. and M. C. Long (2000). "Cost and quality under managed care: irreconcilable differences?" Am J Manag Care **6**(3): 305-12.
24. Litvak, E. 2005. Optimizing patient flow by managing its variability. In Front Office to Front Line: Essential Issues for Health Care Leaders. Vol. 5, edited by S. Berman. Oakbrook Terrace: Joint Commission Resources. Pp. 91-111
25. M. Q. Stearns, et al., "SNOMED clinical terms: overview of the development process and project status," Proc AMIA Symp, pp. 662–666, 2001.
26. MacQueen, J. B. (1967). Some Methods for Classification and Analysis of MultiVariate Observations. Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.
27. Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Ed. Christopher D Manning, Prabhakar Raghavan, & Hinrich Schütze. Cambridge University Press, 2008.
28. McCaig L, Burt C: National Hospital Ambulatory Medical Care Survey: 2001 Emergency Department Summary. Advance Data from Vital and Health Statistics, Centre for Disease Control and Prevention, National Centre for Health Statistics, 2003 Report No 335.
29. McCaig, L.F., Burt, C.W. "National Hospital Ambulatory Medical Care Survey: 2002 Emergency Department Summary." Web Page, No. 340, March 18, 2004. www.cdc.gov/nchs/data/ad/ad340.pdf. Accessed April 2004
30. McManus, M. L., M. C. Long, et al. (2004). "Queuing theory accurately models the need for critical care resources." Anesthesiology **100**(5): 1271-6.
31. NLTK: the Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1. Philadelphia, Pennsylvania, Association for Computational Linguistics.
32. Pakhomov, S. V., J. D. Buntrock, et al. (2006). "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." J Am Med Inform Assoc **13**(5): 516-25.
33. Patrick, J., Y. Wang, et al. (2007). "An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology." Proceedings of the Australasian Workshop on Health Knowledge Management and Discovery (HKMD). Ballarat, Australia. **68**: 219-226.
34. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A Shared Task Involving Multi-label Classification of Clinical Free Text.BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ; 2007.
35. Richardson, D. B. (2002). "The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay." Med J Aust **177**(9): 492-5.
36. Ridge, J.C., S. K. Jones, et al. (1998). "Capacity planning for intensive care units." Eur J Oper. Res **105**(2): 346-355.
37. Ruch, P., J. Gobeilla, et al. (2008). "From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding." AMIA Annu Symp Proc: 636-40.
38. Shahani, A. K., N. Korve, et al. (1994). "Towards an operational model for preventing and treatment of asthma attacks." J Oper res Soc 45(8): 916-26.
39. Siddharthan, K., W. J. Jones, et al. (1996). "A priority queuing model to reduce waiting times in emergency care." Int J Health Care Qual Assur **9**(5): 10-6.
40. Snow, N., K. T. Bergin, et al. (1985). "Readmission of patients to the surgical intensive care unit: patient profiles and possibilities for prevention." Crit Care Med **13**(11): 961-4.
41. Sprivulis, P. C., J. A. Da Silva, et al. (2006). "The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments." Med J Aust **184**(5): 208-12.
42. Tu, J. V., C. D. Mazer, et al. (1994). "A predictive index for length of stay in the intensive care unit following cardiac surgery." CMAJ **151**(2): 177-85.
43. Lee E.K., F Yuan, et al. (2012). "A clinical decision tool for predicting patient care characteristics: patients returning within 72 Hours in the emergency department." Proc AMIA Symp, pp. 495-504.