Models for Assessing the Impact of Resource Allocation in Hospitals

Natalia Yankovic

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
under the Executive Committee of the Graduate School of
Arts and Sciences
COLUMBIA UNIVERSITY
2009
UMI Number: 3373583

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI

UMI Microform 3373583
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346
ABSTRACT

Models for Assessing the Impact of Resource Allocation in Hospitals

Natalia Yankovic

This dissertation focuses on three issues that reflect some of the most important challenges facing both hospital administrators and healthcare policy makers. First, we present an empirical research study on the effects of ambulance diversion on patients’ safety. To determine whether increased diversion activity is associated with poor patient outcomes, we analyze myocardial infarction deaths as a function of emergency department (ED) diversion status within the five boroughs of New York City. Negative binomial regressions are used to demonstrate a statistically significant association between the level and extent of ambulance diversions and increasing myocardial infarction deaths. The second project is dedicated to identifying effective nursing levels for specific hospital units. We represent the nursing system as a variable finite-source queuing model. An approximating assumption results in a reliable, tractable, easily parameterized two-dimensional model that represents the crucial interaction between the nurse and bed systems. We use this model to show how unit size, nursing intensity, occupancy levels and unit length-of-stay each affect the impact of nursing levels on responsiveness to patients’ needs and thus how inflexible nurse-to-patient ratios can lead to either understaffing or overstaffing. The model can also be used to determine the relative impact of lack of inpatient beds and nursing levels on ED delays for a particular unit. Finally, we examine the problem of capacity and admission decisions for a stroke unit. Access to a stroke unit with dedicated beds and staff decreases the mortality rate and the need for institutional long-term care after stroke. In recent years hospital utilization has reached levels at which the rationing of critical care beds has become unavoidable, so there is a need to support capacity and admission decisions. We present a two-dimensional queuing model and an iterative approximation for representing the flow of stroke patients in a hospital. We show how these models can help assess the impact of capacity and admission decisions on the performance of the medical units involved in the recovery of stroke patients.
4 Optimizing Bed Allocation and Admission Rule of a Stroke Unit

4.1 Introduction .................................................. 50
4.2 Model description .............................................. 51
4.3 Alternative modeling of the two stage chain .................. 52
  4.3.1 No shared capacity ....................................... 52
  4.3.2 Single LOS distribution in neurology unit ............... 53
  4.3.3 Iterative approximation ................................... 55
  4.3.4 Performance of the models and approximations ........... 58
4.4 Optimizing the capacity ..................................... 60
  4.4.1 Impact of sharing capacity with neurology patients ...... 61
  4.4.2 Evaluation of alternative policies for reserving beds for stroke patients .............................. 63
4.5 Evaluating severity-dependent admission policies .......... 64
  4.5.1 Impact of overflowing stroke patients .................... 68
  4.5.2 Impact of shared capacity ................................ 69
4.6 Discussion .................................................. 70

List of Tables

2.1 Characteristics of the New York City Myocardial Infarction Mortality Data, 1999 - 2000. 7
2.2 Characteristics of independent variables. ........................... 9
2.3 Characteristics of dependent variables. .......................... 10
2.4 Association between diversion and AMI deaths .................. 14
3.1 Performance of the queuing model. ............................. 38
3.2 Staffing needed to achieve delay targets - Size effect. ........... 41
3.3 Staffing needed to achieve delay targets - Nursing intensity effect. 42
3.4 Staffing needed to achieve delay targets - ALOS effect. ........ 43
4.1 Minimum capacity to achieve performance target. ............... 62
4.2 Evaluating benefits from admission policy. ...................... 68
4.3 Benefits from admission policy, overflowing patients. ......... 70
4.4 Benefits from admission policy, shared capacity. ............... 71
List of Figures

3.1 Simulation of a hospital unit ................................................. 35
3.2 Queuing model - Base case performance .................................. 39
3.3 Combined impact of nursing intensity and ALOS effects .............. 44
3.4 Impact of nurse staffing levels on ED overcrowding .................. 45
3.5 Heuristic performance ....................................................... 47
4.1 Description of the two stage chain - Stroke care ...................... 52
4.2 Description of the approximation schemes ............................. 58
4.3 Performance of iterative approximation ................................. 59
4.4 Impact of reserving capacity at neurology unit ......................... 60
4.5 Impact of sharing the neurology unit with non-stroke patients .... 63
4.6 Reserving $B_r$ empty beds for stroke patients in neurology unit ... 65
4.7 Reserving capacity for stroke patients in neurology unit ............ 66
1 Introduction

Hospital systems and physicians have not, in general, used operational approaches and methodologies that have been common in many other service industries. Moreover, there is an historical lack of involvement of the OR community in healthcare problems which is likely due, in part, to the complexity of the healthcare system with its highly regulated environment, complex price/cost structure, and performance factors that are often extremely hard to quantify.

However, increasing pressure to cut costs and improve service quality, as well as the availability of more operational data, is creating greater interest in analytical approaches to addressing operational problems in healthcare settings. Hospital administrators, physicians and other healthcare professionals are increasingly willing to try new methodologies to help them make better decisions. Managing bed capacity and managing the work-force capacity are two good examples of operational problems where analytical models could be useful to understand the trade-offs between cost and service quality.

Managing bed capacity has become an important and controversial issue. Even though the number of community hospitals in the U.S. has leveled out recently after decades of decline, the number of beds per 1,000 persons continues to fall (American Hospital Association 2005). The closing and downsizing of hospitals may be explained by the extensive belief among politicians, policymakers and hospital administrators that a desirable target for a cost efficient hospital is to have an 85% occupancy. The 85% target is arbitrary and can lead to severe congestion in hospitals with smaller clinical units, since many beds are not fungible but patient-type specific.

The closing and downsizing of hospitals have also affected emergency departments (ED's). The Survey of Hospital Leaders (American Hospital Association 2008) shows that ED visits have been steadily increasing since 1997, while the number of EDs has been going down. Nearly half of all ED report capacity problems with 65% of urban hospitals and 73% of teaching hospitals reporting that the ED is "at or over" capacity. Moreover, the majority of urban and teaching hospitals experienced periods of ambulance diversion in the last twelve months. The reasons behind the episodes of diversions are not only the overcrowding of the ED, but also the lack of acute care beds for specific inpatient units, and most often, the lack of staffed critical care beds.

This dissertation focuses on three projects that deal with some of the issues presented above: ambulance
diversion, staffing of inpatient units and management of capacity and admission policies for specific acute care beds.

Section 2 is devoted to empirical research on the effects of ambulance diversion on patients' outcomes and safety. Ambulance diversions have been shown to increase out-of-hospital transport times. However, it is not known whether increasing amounts of diversion affect clinical outcomes. To determine whether increased diversion activity is associated with poor patient outcomes, we analyze myocardial infarction deaths as a function of emergency department diversion status. All adult patients dying of myocardial infarctions in New York City during the study period of January 2, 1999, to December 31, 2000, are included. Diversion status data from 58 New York hospitals was obtained from the New York City Fire Department. Negative binomial regressions are used to model the association of myocardial infarction deaths and diversion data within the five boroughs of New York City and the results indicate a statistically significant association.

The third section is dedicated to identifying effective nursing levels for specific hospital units. Nursing care is arguably the single biggest factor in both the cost of hospital care and patient satisfaction. Inadequate inpatient nursing levels have also been cited as a significant factor in medical errors and emergency room overcrowding. Yet, there is widespread dissatisfaction with the current methods of determining nurse staffing levels, including the most common one of using minimum nurse-to-patient ratios. In this paper, we represent the nursing system as a variable finite-source queuing model. We show that though the exact model requires a four-dimensional state space, an approximating assumption results in a reliable, tractable, easily parameterized two-dimensional model. We use this model to show how unit size, nursing intensity, occupancy levels and unit length-of-stay each affect the impact of nursing levels on performance and thus how inflexible nurse-to-patient ratios can lead to either understaffing or overstaffing. Our model represents the crucial interaction between the nurse and bed systems and therefore includes the nursing workload due to admissions, discharges and transfers, as well as the observed impact of nursing availability on bed occupancy levels. The model can also be used to determine whether lack of inpatients beds or nursing staff is the bottleneck responsible for ED delays for a particular unit.

Finally, in section 4, we examine the problem of capacity and admission decisions for a stroke unit. The benefits to certain stroke victims of starting treatment in a stroke unit are well established in the medical literature: access to a stroke unit with dedicated beds and staff decreases the mortality
rate and the need for institutional long-term care after stroke. In recent years hospital utilization has reached levels at which the rationing of critical care beds has become unavoidable, so there is a need to support capacity and admission decisions. We present a two-dimensional queuing model and an iterative approximation for representing the flow of stroke patients in a hospital. We show how these models can help assessing the impact of capacity and admission decisions in the performance of the medical units involved in the recovery of stroke patients.
2 Ambulance diversion and myocardial infarction mortality

2.1 Introduction

Emergency department crowding is a growing problem in the United States (Burt and Schappert 2004, McCaig and Ly 2002, Derlet et al. 2001). A frequently employed method of mitigating emergency department crowding is invoking diversion status, where the central dispatcher diverts incoming ambulances to other hospitals (Burt and Schappert 2004).

As emergency department crowding has worsened, the frequency of ambulance diversions has increased. The American Hospital Association reports that over 25% of all hospitals experienced periods of ambulance diversion in 2007. For urban and teaching hospitals, the numbers are 56 and 64 percent respectively and 1 in 8 experienced diversion more than 20% of the time (AHA 2008).

Delays in emergency care can have grave consequences for certain emergency patients, particularly those suffering an acute myocardial infarction (AMI). In these patients, the rapidity with which reperfusion therapy (including thrombolytic therapy and percutaneous coronary intervention like angioplasty and stent placement) is initiated has a significant impact on patient mortality (GUSTO 1993, Boersma et al. 1996, CMMS 2002).

Although EMS protocol in most cities (including New York) mandates that a hospital’s diversion status be overridden for a patient in extremis, such as during an acute myocardial infarction, there is evidence that this rule is not always followed, but even if the rule is overridden the whole EMS system may be affected by diversions leading to delays.

Ambulance diversion disrupts the timely access to medical care in several ways. First, it increases out-of-hospital transport times, delaying emergency medical care (Schneider et al. 2003, Redelmeier et al. 1994). Also, since an ambulance will have to travel longer to find an available emergency department (ED) there is a decrease in overall ambulance availability that may delay the response to new ambulance requests. Moreover, in periods of ambulance diversion EDs are generally more crowded so there may be a longer delay in getting access to care after the patient arrives to the ED. This is supported by research that shows that for patients with suspected myocardial infarctions, time to thrombolysis was
longer during periods of emergency department crowding (Schull et al. 2004).

There is evidence that ED overcrowding is becoming an increasing problem. Wait times to see an ED physician have increased 11.2 percent per year from 1997 to 2004. The median wait for patients diagnosed with AMI increased from 8 minutes in 1997 to 20 minutes in 2004 with higher numbers reported for urban areas and for teaching hospitals (Wilper et al. 2008).

The effects of ambulance diversion have been the subject of several studies. A recent review of the literature (Pham et al. 2006) identifies 11 studies on ambulance diversion and mortality (as adverse effect). Five of them were anecdotal or case reports that attributed ambulance diversion to patient deaths. However, statistical analyses have failed, to date, to find an association between ambulance diversion and mortality.

The aim of our study was to determine whether ambulance diversion is associated with poor clinical outcomes. We hypothesized that ambulance diversion has a more substantial impact on critically ill patients for whom time to treatment is of utmost importance, such as patients with acute myocardial infarction.

2.2 Source of data and setting

The study setting was the city of New York, which is comprised of five boroughs - Manhattan, the Bronx, Brooklyn, Queens, and Staten Island - and has a population exceeding eight million.

Our study was observational and retrospective in nature. We relied on three sources of data, with a study period from January 2nd 1999 to December 31st 2000. Our study was restricted to this time period because of the difficulty of obtaining ambulance diversion data which, in New York, is controlled by the Fire Department and is considered politically sensitive. During this period, the New York City Fire Department-operated emergency medical response system included 58 area hospitals, including three on the border of Queens and Long Island that were included in the Queens catchment area, as per New York Fire Department protocol.

The study protocol (number AAAA0354) was approved by the Institutional Review Board of Columbia Presbyterian Medical Center.
(a) New York City death certificates

The first data-set was provided by the New York City Department of Health and Mental Hygiene. All myocardial infarction deaths occurring within the study period were included in our analysis if the patient was over the age of 18. These data included age, sex, race, ethnicity, zip code of residence, and the borough in which the death occurred of all persons reported as dying of myocardial infarctions in New York City during the study period. There was no information on whether the diagnosis of myocardial infarction was confirmed by a post-mortem examination and we had no information about the time or the specific place of death (e.g., hospital, ambulance, home).

(b) Inpatients with AMI diagnosis

The second source of data was the Statewide Planning and Research Cooperative System (SPARCS) database, which collects patient level detail information on patient characteristics, diagnoses and treatments, services, discharge information, and charges for every hospital discharge, ambulatory surgery patient, and emergency department admission in New York State. The New York State Department of Health provided a de-identified database for all patients admitted and discharged during 1999 and 2000 in NYC’s hospitals with a primary or secondary diagnosis of acute myocardial infarction (NYSDH 2008). One of the SPARCS discharge codes is death, however there was no information on whether or not the death was caused by myocardial infarction.

(c) Ambulance diversion

The third source of data was prospectively collected by the New York City Fire Department over the same study period. These data included time, date, duration, and nature (one of five mutually exclusive categories of diversion: total, critical adult, psychiatric, obstetric, or pediatric) of ambulance diversions for 58 area hospitals operating under their central dispatch. To capture those ambulance diversions that might affect patients suffering a myocardial infarction, we included in our study episodes of critical adult diversion - diversions of patients who would likely be admitted to a critical care unit, as well as episodes of total diversion - diversion of patients, including those presenting with chest pain or other possible symptoms of AMI, who did not fall into the other four categories.

The data was summarized by borough and date. Table 2.1 presents some of the characteristics of the mortality data during 1999-2000. The geography of New York City, where most of the boroughs are
Table 2.1: Characteristics of the New York City Myocardial Infarction Mortality Data, 1999 - 2000.

<table>
<thead>
<tr>
<th>DEATHS with AMI diagnosis</th>
<th>NYC Department of health and Mental Hygiene</th>
<th>SPARCS (primary diagnosis AMI)</th>
<th>SPARCS (any AMI diagnosis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of MI deaths</td>
<td>4986</td>
<td>4757</td>
<td>1492</td>
</tr>
<tr>
<td>Mean Age</td>
<td>78.1 ± 12.3</td>
<td>77.9 ± 12.6</td>
<td>77.8 ± 11.3</td>
</tr>
<tr>
<td>Age &lt; 80 %</td>
<td>48.0%</td>
<td>48.7%</td>
<td>51.1%</td>
</tr>
<tr>
<td>Male %</td>
<td>45.0%</td>
<td>47.7%</td>
<td>45.0%</td>
</tr>
</tbody>
</table>

Mean MI deaths per day

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>2.98 ± 1.71</td>
<td>2.76 ± 1.67</td>
<td>1.04 ± 1.02</td>
<td>1.02 ± 1.00</td>
<td>2.02 ± 1.49</td>
<td>2.07 ± 1.50</td>
</tr>
<tr>
<td>Bronx</td>
<td>1.99 ± 1.47</td>
<td>1.71 ± 1.33</td>
<td>0.54 ± 0.71</td>
<td>0.64 ± 0.81</td>
<td>1.10 ± 1.02</td>
<td>1.15 ± 1.07</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>4.14 ± 2.12</td>
<td>4.01 ± 2.10</td>
<td>1.12 ± 1.10</td>
<td>1.20 ± 1.03</td>
<td>1.94 ± 1.37</td>
<td>2.19 ± 1.50</td>
</tr>
<tr>
<td>Queens</td>
<td>3.58 ± 2.06</td>
<td>3.49 ± 1.95</td>
<td>1.10 ± 1.14</td>
<td>1.05 ± 1.08</td>
<td>1.96 ± 1.56</td>
<td>1.99 ± 1.44</td>
</tr>
<tr>
<td>Staten Island</td>
<td>1.01 ± 1.02</td>
<td>1.02 ± 1.04</td>
<td>0.31 ± 0.56</td>
<td>0.27 ± 0.51</td>
<td>0.51 ± 0.72</td>
<td>0.53 ± 0.78</td>
</tr>
</tbody>
</table>

± standard deviation for actual AMI deaths per day


connected only by bridges and tunnels (the exception is Brooklyn and Queens, which are contiguous, an issue we address in the sensitivity analysis section), makes the grouping of the data into boroughs especially appropriate allowing for a natural experiment since it would be extremely unlikely for an ambulance to cross borough lines.

2.3 Methods of measurement

The idea was to link the summary mortality data (numbers of deaths per borough per day) to summary diversion data (amount of diversion time per borough per day) by borough and date to capture diversions that may have affected patients' timely access to care.
2.3.1 Variable construction

Independent variables

The main independent variable was the borough diversion rate per day. This variable was defined for each borough as the total of emergency department hours on diversion divided by the number of daily available emergency department hours (i.e., the number of emergency departments in a given borough multiplied by 24 hrs.).

We hypothesized that due to increased ambulance travel times, deaths due to AMI would be greater when several hospitals in the same borough were on diversion simultaneously. To assess the effect of multiple hospitals being on diversion simultaneously, we defined “gridlock” as the event that more than 25 percent of borough hospitals are on diversion at the same time. Twenty-five percent was chosen as the cut-off so that Staten Island, which had four hospitals during the study period, would be considered in gridlock only when more than one emergency department was on diversion. We measured gridlock in two ways: we used a dummy variable for the event that a borough experienced time in gridlock during a day, as well as a variable for the daily percentage of time the borough was in gridlock.

Table 2.2 (a) gives the distribution of average hospital diversion rate per day and Table 2.2 (b) the distribution of the gridlock for New York City during the study period.

Dependent variables

We summarized the number of deaths per borough per day of all persons reported as dying of myocardial infarctions in New York City. This variable, total deaths, includes people who died and never contacted the EMS, people who died waiting for an ambulance or during transport to the ED, patients who died in the ED and also inpatient deaths.

We assumed that ambulance diversion on a given day should not impact inpatients that were admitted on a prior date, yet those are included in the total deaths variable. In order to better assess the impact of ambulance diversion on those patients who are most likely to be affected by diversions, we constructed a variable for non-inpatient deaths.

For each borough, we used SPARCS data for each day to compute the number of inpatients with an
Table 2.2: Characteristics of independent variables.

(a) Average hospital diversion rate per day, 1999 - 2000.

<table>
<thead>
<tr>
<th>Borough</th>
<th>Average hospital diversion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Manhattan</td>
<td>0.0291</td>
</tr>
<tr>
<td>Bronx</td>
<td>0.0221</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>0.0294</td>
</tr>
<tr>
<td>Queens</td>
<td>0.0337</td>
</tr>
<tr>
<td>Staten Island</td>
<td>0.0191</td>
</tr>
<tr>
<td>Total</td>
<td>0.0267</td>
</tr>
</tbody>
</table>


Average hospital diversion rate = Cumulative daily borough emergency department diversion time in hours / Number of borough daily available emergency department hours

(b) Distribution of gridlock, 1999 - 2000.

<table>
<thead>
<tr>
<th>Gridlock</th>
<th>N days gridlock</th>
<th>% Daily time experiencing gridlock</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>38</td>
<td>12.82%</td>
<td>0.1421</td>
</tr>
<tr>
<td>Bronx</td>
<td>33</td>
<td>19.87%</td>
<td>0.2363</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>39</td>
<td>18.60%</td>
<td>0.2038</td>
</tr>
<tr>
<td>Queens</td>
<td>45</td>
<td>21.00%</td>
<td>0.1773</td>
</tr>
<tr>
<td>Staten Island</td>
<td>17</td>
<td>3.71%</td>
<td>0.0632</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>16.72%</td>
<td>0.1878</td>
</tr>
</tbody>
</table>


Gridlock = Greater than 25 percent of a borough's emergency departments simultaneously on diversion at any point during the day (Diversion = Total + Critical Adult)
AMI diagnosis who died that day but who were admitted prior to that day. Because SPARCS data does not include cause of death, we constructed two variables to account for inpatient deaths that may have been caused by myocardial infarction: the first including only deaths of patients with a primary diagnosis of AMI (inpatients deaths 1), and the second including deaths from patients with a primary or secondary AMI diagnosis (inpatient deaths 2).

The main dependent variables were constructed by subtracting inpatient deaths from the total deaths (on a per borough per day basis). For some borough-days, the SPARCS data reported a greater number of AMI inpatient deaths than the total number of deaths from AMI recorded by The New York City Department of Health and Mental Hygiene. This is likely due to the fact that we used diagnostic codes rather than cause of death for computing the inpatient death variables. In these cases, we truncated the data at zero.

We defined the variable non-inpatient deaths 1 as the truncated variable constructed by subtracting inpatient deaths with a primary AMI diagnosis, as calculated above, from the total AMI death count. We defined the variable non-inpatient deaths 2 as the truncated variable constructed by subtracting inpatient deaths with any AMI diagnosis from the total AMI death count. This procedure resulted in 5.3% of truncated borough-days for the definition non-inpatient deaths 1 (with an average truncation of 1.1658 deaths for truncated cases) and 14.9% of truncated borough-days for non-inpatient deaths 2 (with an average truncation of 1.4239 deaths for truncated cases). Table 2.3 summarizes some of the characteristics for the constructed dependent variables.

Table 2.3: Characteristics of dependent variables.

<table>
<thead>
<tr>
<th></th>
<th>Total deaths</th>
<th>Non-inpatients deaths 1</th>
<th>Non-inpatients deaths 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.6693</td>
<td>1.9942</td>
<td>1.4729</td>
</tr>
<tr>
<td>Variance</td>
<td>4.1112</td>
<td>2.6692</td>
<td>3.9770</td>
</tr>
<tr>
<td>% truncated cases</td>
<td>---</td>
<td>5.27%</td>
<td>14.89%</td>
</tr>
</tbody>
</table>

Non-inpatient deaths 1 = [Total deaths - Inpatient deaths 1]¹
Non-inpatient deaths 2 = [Total deaths - Inpatient deaths 2]²
2.3.2 Construction of the estimation

Because the dependent variable, myocardial infarction deaths per day per borough, is a count variable, negative binomial regressions were used to model the predictive effect of ambulance diversion. Negative binomial modeling is the appropriate functional form when the dependent variable follows a distribution in which the variance is larger than the mean, as is the case in our data (Cameron and Trivedi 1998). For the negative binomial regression $\lambda_i$ is respecified so that

$$\log \lambda_i = \beta' x_i + \epsilon$$ (2.1)

where $\exp(\epsilon)$ has a gamma distribution with mean 1.0 and variance $\alpha^2$. This additional parameter $\alpha$ is estimated from the dispersion rate such that

$$\frac{\text{Var}(y_i)}{E[y_i]} = 1 + \alpha E[y_i]$$ (2.2)

The resulting probability distribution is

$$\text{Prob}[Y = y_i|x_i] = \frac{\Gamma(\theta + y_i)\Gamma(1 - \lambda_i)^{y_i}}{\Gamma(1 + y_i)\Gamma(\theta)}, y_i = 0, 1, \ldots; \theta > 0$$ (2.3)

where $r_i = \theta/(\theta + \lambda_i)$ and $\theta = 1/\alpha$

The estimated coefficient from a negative binomial regression can be interpreted as a percentage change in the dependent variable given a unit change in the independent variable. In the case of categorical variables, the unit change in the independent variable represents the movement from absence to presence of the marker. The analyses were recalculated assuming a Poisson distribution, but there was no significant difference in standard errors or coefficients of the independent variables. All analyses were carried out using Stata version 9.0.

To account for the known seasonal effect in both the incidence of myocardial infarction mortality and ambulance diversions and potential weekly and yearly variation, we included day of week, month of year, and year categorical variables as independent controls (Sheth et al. 1999). Since socioeconomic status and hospital quality varies among boroughs, we addressed these potential confounders and, more
generally, inter-borough variability in death rates by including a borough categorical variable.

2.3.3 Sensitivity analyses

To assess the robustness of our results among different patient groups, we repeated the analysis using patient subgroups by gender and age. All observations with non-missing values for the subgroup variable were included.

To address the possibility that an external event caused either increased ambulance diversions or myocardial infarction fatalities or both, we employed two strategies: first, subgroup analysis by year, and second, including a week-in-sample categorical variable. These methods serve to control for the possibility that, due to some external event such as influenza season, one week has a greater number of myocardial infarction (or ambulance diversions) than others. This is particularly relevant in our sample period as the 2000-2001 flu season, which included November and December of 2000, was much milder than the flu seasons of 1998-1999 and 1999-2000 (CDCP 2008). We also included an additional categorical control for the 15 days in which a severe weather event occurred.

We tested our results first assuming that Brooklyn and Queens operate independently though as explained previously, there may be some hospitals in each borough served by ambulances in the other. Because of this possible interdependence, we also tested the results assuming Brooklyn and Queens act as one large borough.

Finally, we conducted a counterfactual test. We tested to see whether on a given day, the number of inpatient deaths among previously admitted patients, was affected by ambulance diversion levels on that day. In this case we used Poisson regressions since the mean and variance were of the same magnitude.

2.4 Results

2.4.1 Sample Demographics

A total of 9,743 adults died of myocardial infarctions in New York City over the time period between January 2, 1999, and December 31, 2000. Forty-six percent were men; 65 percent were white, 19 percent
were black, and 11 percent were Hispanic. The boroughs’ mortality levels mirrored their populations; Brooklyn accounted for the most myocardial deaths with 2975, and Staten Island, the least, at 741. The mean number of myocardial infarction deaths per borough-day was 2.67 (95 percent confidence interval, 2.60 to 2.74). Over the same period there were 3023 inpatient deaths with primary myocardial infarction diagnosis, and 5643 inpatient deaths with primary or secondary myocardial infarction diagnosis (31% and 58% of total deaths due to myocardial infarction, respectively). Thus, our definition of inpatient deaths is consistent with the American Heart Association’s (AHA 2008) estimate that only 28% of deaths from myocardial infarction nationwide occur among inpatients (see Table 2.1).

We found a seasonal effect in the incidence of myocardial infarction mortality and in the number of admissions with a myocardial infarction diagnosis. We also found that the hourly number of admissions and discharges follows a pattern that peaks in the afternoon, with only 5.8% of discharges occurring from 12:00 AM to 8:00 AM.

2.4.2 Ambulance Diversion

Ambulance diversion was a frequent occurrence in New York City in 1999 and 2000. On average, three hospitals per day citywide went on total or critical adult diversion status, with each diverting ambulance admissions for approximately five hours on average. Diversion was most frequent in the winter months and in Manhattan. The distribution of the variable borough diversion rate is shown in Table 2 (a). We also found a daily pattern with diversion episodes starting at midnight accounting for 26.2% of all cases, and contributing to 24.6% of the total diverted time. 44% of all diversions occurred from 12:00 AM to 8:00 AM and we found a second peak at 4:00 PM accounting for 11.7% of the total diversion episodes and 15.8% of the total diverted time.

2.4.3 Gridlock

Gridlock occurred on 172 borough-days during the two-year study period. With 3,645 borough-days during the study period this represent a 4.7% of the observations. In days on which gridlock occurred an average of 16.72% of the available emergency department time of the borough was spent with more than 25% of the hospitals on diversion at the same time (Table 2.2 (b)).
Table 2.4: Association between diversion and AMI deaths

(a) Association of myocardial infarction mortality with ambulance diversion.

<table>
<thead>
<tr>
<th></th>
<th>Total deaths</th>
<th></th>
<th>Non-inpatients deaths 1</th>
<th></th>
<th>Non-inpatients deaths 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average hospital diversion</td>
<td>0.743 (2.68)*</td>
<td>0.679 (1.98)**</td>
<td>0.927 (2.16)**</td>
<td>0.390 (2.60)*</td>
<td>0.425 (2.32)**</td>
<td>0.684 (3.05)*</td>
</tr>
<tr>
<td>Gridlock dummy</td>
<td>0.390 (2.60)*</td>
<td>0.425 (2.32)**</td>
<td>0.684 (3.05)*</td>
<td>0.142 (2.94)*</td>
<td>0.155 (2.47)**</td>
<td>0.157 (1.97)**</td>
</tr>
</tbody>
</table>

* Significant at 1%. ** Significant at 5%.

In these negative binomial regressions, the independent variables included dummy variables for day of the week, month, year and borough.

(b) Association of inpatient myocardial infarction mortality with ambulance diversion.

<table>
<thead>
<tr>
<th></th>
<th>Inpatient deaths 1</th>
<th></th>
<th>Inpatient deaths 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef. (t-stat)</td>
<td>Coef. (t-stat)</td>
<td>Coef. (t-stat)</td>
<td>Coef. (t-stat)</td>
</tr>
<tr>
<td>Average hospital diversion</td>
<td>1.011 (1.74)</td>
<td>0.514 (1.21)</td>
<td>0.323 (0.75)</td>
<td>0.013 (0.04)</td>
</tr>
<tr>
<td>Gridlock dummy</td>
<td>0.323 (0.75)</td>
<td>0.013 (0.04)</td>
<td>0.086 (0.86)</td>
<td>0.122 (1.58)</td>
</tr>
</tbody>
</table>

In these Poisson regressions independent variables included were dummy variables for day of the week, month, year and borough.

2.4.4 Association between diversion and AMI deaths

Multivariate regressions of borough diversion rate on myocardial infarction mortality counts, controlling for day of week, month of year, year, and borough, revealed a significant association between borough diversion rate and increased overall and non-inpatient mortality as shown in Table 2.4 (a). There were also significant associations of both episodes of gridlock and time in gridlock with myocardial infarction deaths per borough-day, supporting the hypothesis of the existence of network effects.

One extra hour of average borough diversion was associated with a 3.1% increase in overall AMI deaths and a 2.9 (3.8) percent increase in myocardial fatalities using the independent variable non-inpatient deaths 1 (non-inpatient deaths 2).
Changing the borough diversion rate from zero to one was associated with a 74.3% increase in overall AMI deaths and a 69.7 (92.7) percent increase in myocardial fatalities using the independent variable non-inpatient deaths (non-inpatient deaths 2) or about 0.38 additional deaths per borough-hour experiencing gridlock and 0.31 (0.23) additional deaths per borough-hour using non-inpatient deaths 1 (non-inpatient deaths 2). Incidence of gridlock was associated with a 14.2% increase in overall deaths and a 15.5 (15.7) percent increase in deaths from myocardial infarction using non-inpatient deaths 1 (non-inpatient deaths 2). In the specification using time in gridlock, we found an associated 39.0% increase in the incidence of deaths, for every extra borough-day in gridlock status, in the overall deaths and a 42.5 (68.4) increase in AMI fatalities using the independent variable non-inpatient deaths 1 (non-inpatient deaths 2).

2.4.5 Sensitivity analysis

Subgroup analyses demonstrated a significant association between high borough diversion rates and cardiac mortality in non-elderly patients (less than 80 years of age) for overall deaths, non-inpatient deaths 1 and non-inpatient deaths 2 in the three models specified. The association was also found in males but not in females and there were non-statistically significant trends toward the association in older patients (greater than or equal to 80 years of age) and women.

Including an additional categorical control for the 15 days in which a severe weather event occurred had no effect on our results. Repeating the regressions with a week-in-sample categorical variable also had no effect. In the individual year analyses, the daily time in gridlock was statistically significant for the year 2000 but not for 1999, while the dummy for existence of gridlock was statistically significant for 1999 but not 2000 in the three models under study. The magnitude of the estimated coefficient for 1999 and 2000 were similar and the change of the significance may be due to the use of smaller sample sizes. Neither the elimination of Brooklyn and Queens form our analysis nor combining them had any effect on our results.

Finally, in our counterfactual test we found no association of inpatient AMI deaths with ambulance diversion (Table 2.4 (b)).
2.5 Limitations

There are several limitations to our study. First, the mechanism by which an increased number of myocardial infarction deaths occur during periods of significant levels of hospital diversion is unclear, and we cannot test whether a greater level of ambulance diversions leads to longer delays in treatment given the limitations of our data. It is possible that on days with higher rates of myocardial infarctions (and hence AMI deaths), more emergency departments are overcrowded, and thus requests for ambulance diversion increase. We believe this to be unlikely, as emergency department visits with the diagnosis of myocardial infarction make up only 0.7 percent of total visits (Burt 1999) and unscheduled AMI admissions accounted for only 2.37 percent of all unscheduled admissions in New York City hospitals in the years under study (NYHD 2008). Also, and most importantly, evidence suggests that diversion is primarily due to the unavailability of inpatient beds for patients who are waiting to be admitted from the emergency department (Fatovich et al. 2005). Since new ED arrivals first have to wait to see a physician, have tests performed and receive a diagnosis, there is generally a delay of hours before a decision is made to admit the patient and hence request an inpatient bed. In addition, since diversions occur when the number of “boarded” patients waiting for beds is large and thus the delay for an inpatient bed is long, this suggests that diversions are the result of arrivals to the ED that occurred many hours prior to the diversion. Given this time lag and the fact that diversions generally occur at the beginning of the day, this makes it very unlikely patients experiencing AMI significantly affect same-day diversions.

Another significant limitation is that our mortality data lacked a variety of critical elements. Since we knew only the borough and not the exact location of death, we couldn’t determine whether the person actually experienced the effects of emergency department overcrowding on the day of his/her death. We may also have incurred measurement errors in linking our datasets. For example, because the analysis was conducted at the level of the borough, we may have linked a death in Southern Brooklyn to a diversion in Northern Brooklyn. In addition, the temporal linking is not ideal; the mortality data did not include time of death, and we may have linked patients to diversions that occurred after their demise.

Most generally, the cause of death was not contained in the SPARCS database and we did not know how cause of death was determined in the data of the Department of Health and Mental Hygiene. However, in a study of one set of patients who died outside of the hospital, forensic pathologists were
able to correctly predict ischemic heart disease as the cause of death prior to post-mortem examination, the gold standard, in 79.7 percent of cases. Our mortality data is likely to have a higher percentage of correct diagnoses as it included patients who died in an ambulance or in a hospital, for whom there would generally be more ante-mortem data (Nashelsky and Lawrence 2003).

2.6 Discussion

The objective of this study was to examine the covariance of ambulance diversion with patient mortality. As in previous studies, we used ambulance diversion episodes as our indicator of significant ED crowding, but also to capture the adverse effects of ambulance diversion itself, e.g. longer transport time and longer response time when calling 911.

We quantified diversion levels using two measures: borough diversion rate, and gridlock. Recent studies have demonstrated delays in treatment time during periods of significant simultaneous hospital diversion, and our gridlock variables were designed to model this effect (Schull et al. 2003a). By using predictor variables that were percentages of total borough emergency department time, we standardized diversion and gridlock times among boroughs with differing numbers of available hospitals. Finally, in contrast to previous studies, our primary outcome was clinical - daily mortality due to myocardial infarctions - within a subset of patients who have been shown to significantly benefit from rapid initiation of treatment.

Our results demonstrate a statistically significant association between ambulance diversion and increasing myocardial infarction deaths. These findings differ from previously published results, which showed no increase in transport-related deaths over a concomitant period of increasing ambulance diversion. However, unlike previous studies, our primary outcome measure was deaths from myocardial infarction, a smaller and, we believe, more sensitive subset of patients. In addition, our analysis included all deaths within a day occurring either in the emergency department or outside a hospital setting, rather than solely those that occurred during transport. This inclusion may capture deaths that occurred after the time of transport but which were nonetheless associated with the degree of emergency department crowding.

Our results, while robust in subgroup and other sensitivity analyses, must be interpreted with caution. Because of the observational, retrospective nature of our study, the relatively small time period over
which it occurred, and the limited datasets on which it was based, there are many potential confounders for which it is difficult to control. While we included categorical covariates to control for the daily, seasonal, yearly, and inter-borough variability in both death rates and ambulance diversion, we were unable to incorporate individual patient or hospital characteristics into our analysis.

Despite the limitations of the data, we believe our findings will be useful in informing hospital administrators and policy makers concerning decisions about hospital capacities. Hospitals have been under continuous pressure to downsize or even close in order to cut healthcare costs. Yet these decisions are generally made without consideration of their impact on ED overcrowding and ambulance diversions. This study demonstrates the potential danger of ignoring these consequences and highlights the need for a better understanding of the specific capacity needs of a hospital in providing timely access to critically ill patients.
3 Identifying good nursing levels: A queuing approach

3.1 Introduction

Maintaining appropriate nurse staffing levels is one of the biggest challenges facing hospitals. Nursing is the largest single component of hospital budgets, typically accounting for over 50% of all costs (Kazahaya 2005), making it an important area for study given the increasing cost of care and pressures from payers to keep prices down.

Furthermore, there is a growing realization of the important role nursing care plays in the quality of healthcare. Over the last fifteen years, evidence has been accumulating relating higher levels of nurse staffing to lower rates of adverse patient outcomes (Needleman et al. 2002) and a decrease in the likelihood of death (Aiken et al. 2002). It is now recognized by many, including the Institute of Medicine (IOM), and the International Council of Nurses (ICN) that there is a preponderance of evidence establishing the positive relationship between nursing care and quality patient outcomes (IOM 2004, ICN 2006). However, there is still no scientifically based methodology to help nurse managers and hospital administrators efficiently allocate scarce nurse resources to promote quality patient outcomes within their own setting.

Minimum nurse-to-patient ratios are one of the most commonly used methods to determine staffing adequacy. California is the first and only state to mandate a minimum nurse-to-patient ratio. The 1999 law AB 394 went into effect in 2004 and set minimum licensed nurse-to-patient ratios of 1 to 6 on general medical-surgical wards. Many other states are now considering similar proposals (Health Policy Tracking Service 2005). However, there have been many arguments made against the use of mandated staffing ratios (Lang et al. 2004, SHS 2005, Kane 2007). Opponents of mandated ratios observe that none of the studies of staffing and quality have identified an optimal ratio and that ratios are too inflexible to account for variation in nursing skills and the severity of patients' illnesses. This has been confirmed by a study showing that the mandated staffing ratios implemented in California did not result in the expected patient benefits (White 2006).

Another common method is to budget the number of nursing staff needed by calculating the total direct productive hours of care per patient day (HPPD) for the number of patients expected to require nursing care over a given time period. The usefulness of the HPPD concept has been questioned by
the American Nursing Association (ANA) because it is a simple quantification of the average patient without considering outlier patients (ANA 1999).

Many hospitals use Patient Classification Systems (PCS) or acuity systems to adjust nurse staffing based on the individual characteristics of the patients. Many PCSs are developed by the institution and not standardized (Seago 2002). California Title 22 (Title 22, Division 5, Ch 1, Section 70053.1, p. 761) requires California hospitals to have a PCS in place to predict their nursing staffing needs on a shift-by-shift basis and to staff accordingly. Hospitals must submit their PCSs to the state, but there is little guidance about what characterizes a valid PCS.

Arguments against the use of static measures and calls for development of patient-centered staffing policies based on careful analysis of multiple variables such as differing patient needs, fluctuations in care needs by day and time, expertise and education of the staff, and other setting characteristics have been proposed (see e.g. Lang et al. 2004). The ANA established guidelines on what should be incorporated in optimal systems that inform staffing decisions. The ANA recommends that staffing decisions be flexible, consider patient characteristics, and tailored to the needs of the patient by incorporating the intensity of nursing care.

Ideally, nurse staffing levels should be based on a quantification of the actual patient needs for nursing and the amount of time associated with these needs, so that nurse services are able to be provided in a timely fashion. This is substantiated by many of the adverse patient events that have been linked to inadequate nursing levels such as failure to rescue and cardiopulmonary resuscitation which are clearly time-sensitive (Kane et al. 2007). The major nursing activities in most clinical units will include admissions, discharges, transfers, administration of drugs, monitoring, preparation for procedures and responding to patient requests made through the use of call buttons. However, there are many activities that are specific to certain types of patients, for example, wound care and pain management for surgical patients. This is one of several factors that suggest the need for a flexible methodology for determining nurse staffing levels.

In this paper, we develop a queuing model to guide nurse-staffing decisions. Queuing models have been around for more than a century and are routinely used in many other service systems, particularly emergency systems such as police, ambulance and fire. Given the stochastic nature of patient demands and services and the need for a high level of responsiveness, queuing methodology seems very well-
suited for guiding nurse staffing. We believe that a major obstacle to their use in the past has been
a lack of electronic data regarding the demands for nursing care and the associated service times.
However, information technology systems to support both clinical and operational decision-making are
increasingly being used in healthcare and this type of data may already be captured or could easily be
routinely collected in the future.

Nurse staffing in hospitals is generally done independently for each clinical unit (sometimes referred to
as wards) which typically vary between 30 and 60 beds. A clinical unit may consist of general medi­
cal/surgical beds or correspond to one or more specific hospital services such as cardiology, infectious
diseases, orthopedics, oncology, neurology, pediatrics or obstetrics. There are also clinical units desig­
nated as “intensive care units” used for the most critically ill or injured patients which are generally
quite small and have very high nursing levels to assure almost constant patient monitoring. For each
nursing shift, which is generally either 8 or 12 hours long, there is a nurse manager and a dedicated
nursing staff.

From an analytical perspective, a clinical unit can be viewed as a finite source queuing system since
demands are generated from the inpatients in that unit. However, due to admissions, discharges and
transfers, the number of inpatients varies over a shift. Furthermore, each of these changes in inpatient
census triggers a demand for nursing care. So in order to identify appropriate nursing levels, it is
necessary to model a single hospital clinical unit as a queuing system with two sets of servers: nurses
and beds. Though patients are usually assigned to a specific nurse for each shift, it is common practice
for any available nurse to attend to a patient if the designated nurse is busy with other patients. So
we assume that both the nursing system and the bed system are multi-server.

The complex interaction effects between the nursing workload and the bed dynamics would seem to
necessitate a four-dimensional state space, even under Markovian assumptions for arrivals and service
times. However, in this paper we devise a reliable, approximate two-dimensional representation that
captures the essential characteristics of the hospital and nurse dynamics and yet is analytically tractable.
Our goal is to develop an easy-to-use model for use by hospital managers in evaluating the impact of
any given nursing level on patient delays. This will enable hospital administrators and nurse managers
to determine nurse staffing levels for various units and shifts that best fit their target service standards.

A major contribution of this paper is the identification of the major factors that affect nurse-to-patient
ratios needed to achieve high levels of patient service. Specifically, we show that in addition to unit size and level of nursing intensity, occupancy levels and average length-of-stay may play a significant role in affecting nurse-to-patient ratios that are consistent with timely service. We also demonstrate that the standard of desired responsiveness to patient needs is also a potentially important factor in determining desirable nurse staffing levels. Since hospitals have dozens of clinical units with very different bed and patient characteristics which also change across the day and week, these observations support the need for a queuing-based approach in order to tailor nurse staffing levels appropriately.

Another major contribution is in demonstrating and quantifying the impact of nursing levels on backlogs in the emergency department (ED). Though a lack of inpatient beds is generally cited as the primary reason for ED overcrowding, we show that even when the number of beds is sufficient, nurse unavailability can cause major backlogs in the ED.

The models that we have developed are part of a project being undertaken in conjunction with the Hospital for Special Surgery (HSS) in New York City and funded by AHRQ, a major national health quality agency, to determine the feasibility and usefulness of using a queuing model to help guide nurse staffing levels in hospitals. In addition to the development of the model, the project will focus on the identification of nursing tasks and the ability to obtain the required input data from both electronic and non-electronic sources. A specific goal of the project is to inform the development of hospital IT systems with regard to operational data that should be routinely collected to inform nurse staffing decisions and evaluate the impact of nurse staffing levels.

The rest of the paper is organized as follows: In section 2 we review the related literature and in section 3 we develop our basic model and derive the steady-state probabilities and major delay metrics. In section 4 we describe a simulation we developed that more accurately represents nursing dynamics and use it to demonstrate the reliability of the queuing model in identifying nursing levels that are consistent with good patient service. We use our model in section 5 to explore the clinical and patient factors that are most likely to affect nurse-staffing levels and highlight situations in which the California mandated ratios are likely to be inconsistent with good system performance. In section 6, we develop and test a single-dimensional heuristic that can be more easily implemented and which can be expanded to include two priority classes. Finally, in section 7 we summarize and discuss our major results and insights and identify some future research directions.
3.2 Literature review

Resource allocation in hospitals has been a subject of operations research studies for many years. Bed occupancy and patient flows within a hospital are examples of early applications of Markov and semi-Markov processes (Young 1965, Kao 1974, Hershey et al. 1981). See Green (2006) for an overview of the use of OR models for capacity planning in hospitals. Given the cost and quality impact of clinical personnel, much of the literature has focused on staffing and scheduling problems. Most of this work has used linear programming models to provide guidance on the scheduling of nurses and other hospital personnel (Miller et al. 1976; Kwak and Lee 1997; Jaumard et al. 1998). Most recently, the work of Wright et al. (2006) uses linear programming to analyze the impact of mandated nurse-to-patient ratios on workforce costs and the ability to provide nurses with "desirable" schedules. All of these have assumed as input a given number of nurses needed for each shift, based on either mandated ratios or one of the other methods mentioned previously.

Queuing models have been used to determine staffing requirements in many service systems in order to provide timely responses to customers' requests for service. Timely service has been found to be an important factor in customer satisfaction and is often used as a service quality measurement, for example in call centers, see e.g. Gans et al. (2003). Queuing models have also been used for staffing in emergency response systems such as police patrol and ambulance dispatch where timely response is correlated with safety (e.g. Green and Kolesar 2004). In healthcare, the Institute of Medicine 2001 report on healthcare quality includes "timeliness" as one of the key "aims for improvement" (IOM 2001). Therefore it makes sense to consider timely response as a key measure for evaluating in nursing performance and hence queuing methodology as an appropriate tool for guiding nurse staffing decisions.

However, the use of queuing models in guiding staffing decisions in healthcare facilities has been limited. One exception was described in a study by Green et al. (2006) which used a non-stationary queuing approach to allocate physicians in the emergency department of an urban hospital in order to reduce the fraction of patients who leave before being seen by a physician. de Véricourt and Jennings (2006) is the only previous paper to use a queuing model in the context of nurse staffing. They use a standard finite source multi-server queue to demonstrate that fixed nurse-to-patient ratios as embodied in the 1999 California legislation cannot achieve consistent performance across different unit sizes. Because their model assumes a fixed number of inpatients, it doesn't include the work generated by new arrivals, departures, and transfers of patients. In contrast, the models developed in this paper incorporate the
potentially large fluctuations in the census of a clinical unit during a shift which significantly affect the
demand for nursing care (Lapierre et al. 1999; Volpatti et al. 2000).

3.3 Model description

We model a specific clinical unit of a hospital for a given shift as a queuing system in order to identify
the number of nurses needed to provide timely care to the patients on the unit. The demands for
nursing care in a clinical unit at any given time are primarily generated from the current inpatients
(e.g. call button requests, medication and monitoring). However, patients move in and out of the unit
due to admissions, discharges and transfers, changing the census level of the unit and hence the size of
the population generating nursing demands. Unlike other service systems where the number of servers
can be adjusted frequently over the day to accommodate changing demand rates, nurses typically
work 8 or 12 hour shifts which generally start simultaneously for all nurses in the unit. Therefore,
the nursing level remains the same over long periods of time during which the number of patients in
the unit can vary substantially. In addition, these changes in the inpatient census each require nurse
involvement. For example, a patient admission requires a nurse to move the patient into the bed, set
up any necessary monitoring and/or intravenous lines, take vital sign readings, etc. Similarly, once a
physician gives a discharge order for a patient, the nurse must be available to provide post-discharge
instructions, prescriptions, fill out forms, etc. before the patient can physically leave. So in order to
capture both the volume and variability of the workload associated with patient demands for nursing
in a clinical unit, it is necessary to consider both the inpatients in that unit as well as the patient
movements in and out of the unit.

We let $B$ be the number of beds in the unit and $N$ be the number of nurses where $B > N$. We assume
that patients arrive to the unit according to a homogeneous Poisson process. This assumption is very
reasonable for units for which most arrivals are unscheduled, such as medical units and obstetrics units
(Young 1965). Even in units where most patients are scheduled, such as surgical units, the exact number
and timing of patient arrivals into the unit (which typically come from the recovery room) have been
found to be very random due to variability, additions, and cancellations in the surgical schedule (Litvak
and Long 2000). We assume that patient lengths of stay (LOS) in the clinical unit are exponentially
distributed. This assumption is supported by empirical data from a large urban hospital for which the
coefficient of variation of LOS was close to one in most clinical units (Green and Nguyen 2001). At
any given time $t$, we assume that there are a fixed number of inpatients in the unit, each of which has an exponential time between requests for nursing care which are independent of each other and that the amount of time a nurse spends on each patient demand is exponentially distributed. Though the assumption of exponentially distributed service times is necessary for analytical tractability, it is also well-supported by the only study reported in the literature on the use of nurses' time in a medical-surgical unit (Lundgren and Segesten 2001). Initially, we assume that requests are performed on a first-come, first-served basis. We will relax this assumption to allow for priority classes in our heuristic model described in section 6. Finally, we assume that all nurses are equally trained and can perform all requests. This assumption is valid for many hospitals and is also consistent with other nurse staffing methodologies, e.g. nurse to patient ratios. We will discuss the implications of this assumption for more complex situations in the last section of the paper.

So our model consists of two queuing systems - the first one related to the *beds* in the unit and the second related to the requests for *nurses*. These queues are related since the number of occupied beds will determine the demand for nurses and nurse availability can influence admissions and discharges, and hence, bed occupancy.

In order to model the exact relationship between the changing level of inpatients and the workload for the nurses, additional dimensions are needed in our state space. In particular, to capture the possibility of a new admission to the unit being blocked when all nurses are busy (a situation often observed in hospitals, see e.g. Zollinger et al. 1999), it would appear to be necessary to keep track of the number of patients assigned to a bed but not yet physically admitted since it is possible to have a bed queue even when beds are not all occupied. Similarly, the number of patients who are ready for discharge needs to be tracked in order to know if a nursing job should result in a patient leaving the system or returning to the inpatient pool.

Even if we were to assume that nurse availability doesn't affect the admissions or discharge process, the state space of the system would need to keep track of whether or not a discharged patient was with a nurse to be able to determine the number of busy nurses subsequent to the discharge. Also, allowing a patient's length of stay to start and end independent of nurse availability could result in cases in which a patient leaves the system without ever actually generating a single nursing demand, including the admissions work. Therefore, the two dimensions can't be decoupled.
In order to obtain a tractable, two-dimensional state space which captures the essential interaction effects between beds and nurses, we model the system dynamics as follows: As observed in reality, we assume that a new patient arrival to the unit that occurs when all the nurses are busy but a bed is available, will have a bed assigned to him/her and generate a request for nursing care at that time (corresponding to the work associated with a new admission), even though the bed will not be occupied until a nurse is available. We further assume that the arrival rate for the nurse system includes the discharge requests, but that the timing of the physical departure of a patient from his/her bed does not necessarily coincide with the discharge work associated with it. In other words, patients may physically depart when their LOS is over regardless of whether or not they are currently being attended to by a nurse. That is, we assume that the work associated with discharge may occur in advance of the patient’s vacating the bed. This assumption is not unreasonable since some patients remain in their room after being officially discharged because, e.g., they are waiting for transportation or a relative to take them home. Based on our work at HSS shadowing nurses and conducting nurse focus groups, we have learned that patients in this situation often continue to generate demands for nursing, for example, medication administration. However, to capture the impact of nurse unavailability on patient flow and to facilitate our analysis, we will assume that discharges of patients who are not being served by a nurse are blocked whenever all nurses are busy. Note that this assumption prevents the hospital LOS for a patient who has been assigned but not yet admitted to a bed from beginning (and therefore also ending) before he/she occupies the bed. The system described above can be described as a bivariate Markov process with state space \((X_b, X_n)\) where \(X_b\) is the number of occupied beds plus the number of patients who are waiting for a bed and \(X_n\) is the number of inpatients being cared for by a nurse plus the number of patients waiting for a nurse.

We will use the following notation in our analyses:

- \(B\) Beds
- \(N\) Nurses
- \(\lambda_b\) Arrival rate for the bed system
- \(\mu_b\) Service rate for the bed system
- \(\lambda_n\) Demand rate for the nurse system
- \(\mu_n\) Service rate for the nurses
We first consider the situation where all patients arriving to the bed system join the queue, and there is no balking or blocking. In reality, many hospitals attempt to block new arrivals to a clinical unit when the number of patients waiting for admission to that unit exceeds some level. This can be done by placing patients in other units that have available beds and/or by going on ambulance diversion, i.e. instructing dispatchers to send ambulances to other hospitals. (Ambulance diversions can be restricted to certain categories of patients, e.g. those requiring critical care beds.) However, some hospitals have a policy of never going on ambulance diversion and in rural areas, this may not be an option. Also, in smaller hospitals, there may be no alternative clinical units to which patients can be admitted. In these cases, and also to study the impact of nurse staffing on ED congestion and bed admissions, the infinite waiting room model is more appropriate.

For any given state \((i, j)\), the set of possible successor states and the associated transition rates depend on the “macrostate” as follows:

(a) \((i > B, j < N)\). All beds are occupied, and all patients in the nurse system are being served. So arrivals to the bed system join the bed queue, changing the state space to \((i + 1, j)\) at rate \(\lambda_b\) since all beds are occupied and no nursing work is generated by this arrival at this time.

Departures from the bed system occur at a rate \(\mu_b \cdot B\), leading to one of two possible states:

- If the departing patient was not in the nurse system the transition is to \((i - 1, j + 1)\), because the vacated bed allows for the admission of a patient from the bed queue which generates a new demand for the nurse system. However, if the departing patient was in the nurse queue the resulting state is \((i - 1, j)\), since as before we generate a new nursing job associated with the admission of a patient from the bed queue, but the departure leaves both systems simultaneously.

Requests for nursing from inpatients occur with rate \(\lambda_n \cdot (B - j)\), changing the state to \((i, j + 1)\).

Completions of nursing jobs occur at rate \(\mu_n \cdot j\) changing the state to \((i, j - 1)\).

(b) \((i \geq B, j \geq N)\). All beds are occupied and all nurses are busy.

Arrivals to the bed system occur as in the previous case. Since all nurses are busy discharges from the bed system are blocked for those without a nurse. That is, we assume that when all nurses are busy, only the patients who are in service with a nurse will be physically discharged if their length-of-stay is over during their service, this occurs with rate \(\mu_b \cdot N\). If \(i > B\) a departure allows a patient from the bed queue to reserve an available bed and the transition will be to state \((i - 1, j)\). When \(i = B\) there is no queue for beds and the departures must be from those
currently in service, so the transition is to state \((i - 1, j - 1)\).

Requests for nursing from inpatients occur with rate \(\lambda_n \cdot (B - j)\), changing the state to \((i, j + 1)\). Completion of nursing jobs occur at rate \(\mu_n \cdot N\) changing the state to \((i, j - 1)\).

(c) \((i < B, j < N)\). Both beds and nurses are available.

New arrivals are immediately admitted and generate a demand for nursing so that transitions to state \((i + 1, j + 1)\) occur with rate \(\lambda_b\). Departures from the bed system can be of the two types described in case (a), but in this case we don’t have patients waiting to be admitted. The transition will be to state \((i - 1, j)\), if the departing patient was not in the nurse system or to state \((i - 1, j - 1)\) otherwise. Requests for nursing from inpatients occurs with rate \(\lambda_n \cdot (i - j)\), changing the state to \((i, j + 1)\) and nursing job completions occur at rate \(\mu_n \cdot j\) changing the state to \((i, j - 1)\).

(d) \((i = B, j < N)\). All beds are occupied but there is no queue, and nurses are available.

Arrivals to the bed system occur as in case (a). Departures from the bed system occur as in case (c). Requests for nursing from inpatients occurs with a rate \(\lambda_n \cdot (B - j)\), changing the state to \((i, j + 1)\) and completions of nursing service occur at rate \(\mu_n \cdot j\) changing the state to \((i, j - 1)\).

(e) \((i < B, j \geq N)\). Beds are available and all nurses are busy.

Arrivals to the bed system occur as in case (c), and discharges from the bed system are blocked for those without a nurse as in case (b), because all nurses are busy. In this situation we will have departures with rate \(\mu_b \cdot N\), and since there is no queue for the bed system and departures must be from those currently in service, the transition is to state \((i - 1, j - 1)\). Requests for nursing from inpatients occur with rate \(\lambda_n \cdot (i - j)\), changing the state to \((i, j + 1)\). Completions of nursing jobs occur at rate \(\mu_n \cdot N\) changing the state to \((i, j - 1)\).

The model described above corresponds to a Quasi Birth and Death Process (QBD) with generator \(Q\) which can be partitioned into blocks of size \((B+1) \times (B+1)\) and boundary conditions for all \(i < B\). This block structure allows us to check for stability of the system and to solve the steady state distribution using the ideas of matrix geometric analysis (Neuts 1981). Appendix 1 includes a complete description of the transition matrix and blocks for the case \(B = 3, N = 2\).

When \(Q\) is positive recurrent, we can solve for the steady-state probabilities \(p_{ij}\) which we can use to
derive performance measures for evaluating the impact of nursing levels on delays for both nurses and beds. These, in turn, can be used to help hospitals identify nursing levels that are consistent with any given target level of service performance.

We can compute the traffic coefficient for the QBD process, \( \gamma \), to determine if the system is positive recurrent or not (Neuts 1981). If we call \( A_0 \) the transition matrix from states \((i, j)\) to states \((i + 1, k)\) (for \( j, k = 0 : B \)), \( A_1 \) the transition matrix from states \((i, j)\) to states \((i, k)\) (for \( j, k = 0 : B \)), and \( A_2 \) the transition matrix from states \((i, j)\) to states \((i - 1, k)\) (for \( j, k = 0 : B \)) a necessary and sufficient condition for ensuring that the QBD process will be positive recurrent is given by:

\[
\gamma = x \cdot A_0 \cdot e_1 - x \cdot A_2 \cdot e_1 < 0
\]

Where \( x \) solves \( x \cdot (A_0 + A_1 + A_2) = 0 \) and \( x \cdot e_1 = 1 \), and in this case we will have stability for the bed system. Appendix 1 includes a complete description of the transition matrix and blocks for the case \( B = 3, N = 2 \).

With this model we can derive the steady-state probabilities and major delay metrics for both, the bed and nurse systems.

### 3.3.1 Bed system

Average bed occupancy rates are widely used and reported to evaluate hospital efficiency (Green 2002). The occupancy level of a clinical unit is simply the bed utilization rate, \( \rho_b \). In our model, this rate is a function not only of the arrival rate \( \lambda_b \), the service rate \( \mu_b \), and the number of beds available \( B \), but also the probability of discharges being blocked because all nurses are busy, \( P(\text{block}) \). We get

\[
P(\text{block}) = \sum_{i=N}^{B-1} \sum_{j=N}^{i} p_{ij} + \sum_{i=B}^{\infty} \sum_{j=N}^{B} p_{ij}
\]  

The average stay in the unit is given by \( \mu_b \cdot (1 - P(\text{block})) \). Using Little's law

\[
\rho_b = \frac{\lambda_b}{B \cdot \mu_b \cdot (1 - P(\text{block}))}
\]
To compute the probability that an arriving patient must wait for a bed, \( P_b(delay) \), or the distribution of the waiting time for a bed, we must consider two different sources of delay. First, an arrival may find all beds occupied \((i \geq B)\) and will have to wait until \( i - B + 1 \) patients are discharged, where the time for each discharge has an exponential distribution with rate \( B \cdot \mu_b \cdot (1 - P(block)) \). Second, a patient may arrive to the bed system at state \( i < B \) and still be delayed because all nurses are busy \((j \geq N)\), in this case he/she will have to wait until \( j - N + 1 \) patients are served, each of which has an exponential distribution with rate \( N \cdot \mu_n \). Since patients arrive to the bed system according to a Poisson process, we get

\[
P_b(delay) = \sum_{i=N}^{B-1} \sum_{j=N}^{i} p_{ij} + \sum_{i=B}^{\infty} \sum_{j=0}^{B} p_{ij} = P(block) + \sum_{i=B}^{\infty} \sum_{j=0}^{N-1} p_{ij} \tag{3.3}
\]

Other standard measures like the expected number of patients in queue or inside the unit can be easily computed. For instance, the expected number of patients waiting for a bed can be computed by:

\[
L_q = \sum_{i=B+1}^{\infty} \sum_{j=0}^{B} (i - B) \cdot p_{ij} \tag{3.4}
\]

It’s important to note that though in our model the bed system has Poisson arrivals and exponential service times, it is not equivalent to an \( M/M/c \) system where \( c = B \). This is because the times that patients spend in a bed are not independent of one another due to the blocking of discharges when all nurses are busy.

### 3.3.2 Nurse system

The nurse system is a finite source multi-server queue where the size of the source is the number of occupied beds and is therefore changing over time. This is further complicated by the fact that the rate at which the source size changes is dependent on the state of the nurse system.

In order to compute the average nurse utilization we need the average effective demand rate for nursing care, \( \lambda_{nf} \), which includes both the demands from current inpatients and the demands due to new admissions.
\[ \lambda_n^{\text{eff}} = \sum_{i=0}^{B-1} \sum_{j=0}^{i} (\lambda_b + (i - j) \cdot \lambda_n) \cdot p_{ij} + \sum_{j=0}^{B} (B - j) \cdot \lambda_n \cdot P_{Bj} \]

\[ + \sum_{i=B+1}^{\infty} \left[ \sum_{j=0}^{N-1} ((B - j) \cdot \lambda_n + B \cdot \mu_b) \cdot p_{ij} + \sum_{j=N}^{B} (B - j) \cdot \lambda_n \cdot p_{ij} \right] \]

In the last half of the above expression, which corresponds to the case when all beds are full, we have to account for nursing jobs due to the admissions of patients from the bed queue which occur when a current inpatient is discharged. Hence, the nurse utilization \( \rho_n \) will be:

\[ \rho_n = \frac{\lambda_n^{\text{eff}}}{N \cdot \mu_n} \]

Since the nurse system is a finite source system and therefore doesn’t have a stationary Poisson arrival process, PASTA does not apply. So in order to get patient delay measures we must use the steady state probabilities \( p_{ij} \) to derive the set of \( q_{ij} \), state probabilities given an arrival is about to occur (for proof see Gross and Harris (1985)).

Hence, the probabilities of an arrival finding the system at state \((i, j)\), for \( i = 0, \ldots \) and \( j = 0, \ldots \) \( \min(i, B) \) can be computed as:

\[ q_{ij} = \frac{\min(i, B) - j}{L_n - L_s} \cdot p_{ij} \]

where \( L_n \) is the expected number of patients in the nurse system \( (L_n = \sum_{i,j} j \cdot p_{ij}) \), and \( L_s \) is the expected number of inpatients \( (L_s = \sum_{i,j} \min(i, B) \cdot p_{ij}) \).

\[ P_n(\text{delay}) = \sum_{i=N}^{B-1} \sum_{j=N}^{i} q_{ij} + \sum_{i=B}^{\infty} \sum_{j=N}^{B} q_{ij} \]

We can use the probabilities \( q_{ij} \) to compute the distribution of the waiting time for a delayed inpatient. A patient arriving to the nurse system at state \((i, j)\) with \((j \geq N)\) will have to wait until \( (j - N + 1) \) patients finish their services \( j \geq N \) and his/her waiting time will follow an Erlang\((j - N + 1), N \cdot \mu_n\) distribution.
3.3.3 Finite capacity

As mentioned previously, many hospitals limit the number of patients waiting for admission into a clinical unit either by placing new arrivals "off-service" into a different unit or going on ambulance diversion. For use in those situations, we consider an alternative model with the same state space and transition dynamics as above, but assuming a finite waiting room, \( WR \geq 1 \), so that \( i \leq B + WR \).

For any given state \((i, j)\), the set of possible successor states and the associated transition rates depend on the "macrostate" as in the infinite model. The only difference is when the bed system reaches capacity, \( i = B + WR \), all arrivals to the bed system are "lost".

The steady state probabilities for this model, \( p_{ij} \), can be directly computed by solving the system of linear equations, or as in the previous case, we can take advantage of the special structure of the transition matrix. Akar and Sohraby (1997) present an unified approach for solving infinite and finite QBD processes using the ideas of the Matrix Geometric solutions without increasing complexity.

As in the infinite capacity model, the performance measures of interest can be computed from these steady state probabilities. The utilization of the bed system will, as before, be affected by the interaction with the nurse system due to the potential blocking of discharges. However, the average arrival rate is no longer \( \lambda_b \) because of the probability of finding the waiting room full and so the bed utilization is given by

\[
\rho_b = \frac{\lambda_b \cdot (1 - P(\text{full}))}{B \cdot \mu_b \cdot (1 - P(\text{block}))} \tag{3.9}
\]

where

\[
P(\text{full}) = \sum_{j=0}^{B} P_{(B+WR)j} \tag{3.10}
\]

As in the infinite case we can compute the effective arrival rate for the nurse system and its utilization rate.
\[ \lambda_n^f = \sum_{i=0}^{B-1} \sum_{j=0}^{i} (\lambda + (i - j) \cdot \lambda_n) \cdot p_{ij} + \sum_{j=0}^{B} (B - j) \cdot \lambda_n \cdot P_{Bj} \]
\[ + \sum_{i=B+1}^{B+WR} \sum_{j=0}^{N-1} ((B - j) \cdot \lambda_n + B \cdot \mu_b) \cdot p_{ij} + \sum_{j=N}^{B} (B - j) \cdot \lambda_n \cdot p_{ij} \]

Hence, the nurse utilization \( \rho_n \) will be:

\[ \rho_n = \frac{\lambda_n^f}{N \cdot \mu_n} \]  

(3.12)

For computing the probability of delay and the distribution of the waiting time for an inpatient demand for nursing care, we need to use the normalization factors associated with the finite source, using the same approach as for the infinite case, presented in equation (3.7).

### 3.4 Simulation of a hospital unit

As noted previously, our queuing model is an approximation of actual nursing dynamics and does not capture all of the elements of a clinical unit which may affect the performance associated with nursing care. To determine its reliability, we developed a simulation model that better conforms to the reality of patient, nurse and bed dynamics.

The major distinction of the simulation is in the modeling of the discharge process, which in the queuing model, is assumed to be blocked for patients not with a nurse when all nurses are busy. In addition, the simulation allows for the inclusion of a bed cleaning time when a patient leaves the unit.

This discrete event simulation shares most of the assumptions of our queuing model. We assume we have one unit of \( B \) beds and \( N \) nurses, which are fixed during the simulation in order to capture the dynamics occurring during a shift. Arrivals to the bed system follow a Poisson process with rate \( \lambda_b \), and an admission can only occur if both a bed and a nurse are available. Each patient on the unit generates nursing requests with an interarrival time that is exponentially distributed with mean \( 1/\lambda_n \) and the service times for requests are assumed to follow an exponential distribution with mean \( 1/\mu_n \). As in the queuing model, the patient stay in the unit begins at the moment the nurse starts the admission.
process, and it follows an exponential random variable with mean $1/\mu_b$.

Based on our observations and conversations with nurses at HSS and discussions with hospital personnel in other hospitals, we divide nurse patient interactions into three categories: initial service (admission), regular service, and final service (discharge). We assume that when there are several nursing jobs waiting for service, priority is given to regular requests (e.g. medication administration and call button requests), then discharge work, and finally new admissions. Each patient has a nurse assigned to him/her upon admission who, if available, responds to that patient’s needs. If the assigned nurse is busy at the time of the request, another nurse, if available, responds. However, only the assigned nurse can process the patient’s discharge.

As mentioned above, we assume that each time a patient physically leaves the unit, a bed cleaning time begins. This cleaning time is uniformly distributed between $[a_c, b_c]$.

The events that can trigger the clock to advance are the following: new arrivals to the bed system, requests for normal service, requests for final service, end of initial service, end of regular service, end of final service, and end of cleaning time.

The space state needed for tracking all the interactions is summarized in Figure 3.1. Patients in bed but not in need or waiting for final service generate regular nursing requests. If all nurses are busy the jobs join the nurse queue. A request for final service occurs when the remaining time in the unit of a patient hits zero. If this happens in the middle of a regular service, it is delayed until the current service finishes. If the assigned nurse is busy at the time of the discharge request, the request for final service joins the patient’s assigned nurse queue.

Initial and final services are assumed to have a different distribution than regular nursing requests. Based on data collected from nurse focus groups at HSS, we assume their durations follow uniform distributions ranging between $[a_i, b_i]$ and $[a_f, b_f]$, respectively.

### 3.5 Numerical results

In order to test the accuracy of our queuing model, we examined the nurse staffing levels needed to keep the probability of inpatient delay below a specified threshold over a range of parameter values for unit size, nursing intensity, average length of stay, and bed utilization.
To test the reliability of the queuing model in estimating the staffing level needed to assure a given delay target, we used as our “base case” the specific 42 bed HSS surgical unit that is the focus of our AHRQ research study. Since the majority of medical/surgical units in most mid-size and large hospitals range between 20 and 60 beds, we varied $B$ accordingly.

The estimation of parameters related to the bed system requires information about admissions, transfers and discharges, data that is readily available for most hospitals. Hospitals also generally record the average time patients spend in the hospital or length of stay (ALOS) as well as average occupancy levels which are often used as indicators of efficiency. Based on historical data from our HSS study unit, the ALOS is 4.3 days. Based on this and ALOS data from other hospitals (see, e.g. American Hospital Directory 2006), we varied ALOS from 3 to 8 days. The average bed occupancy level in HSS is 78% which is nearly identical to the average daily occupancy level in NYC of 79% (Final Report of the Commission on Health Care Facilities in the 21St Century 2006). We used a set of arrival rates to
the bed system $\lambda_b$ to roughly correspond to different bed occupancy levels. Since actual bed occupancy levels will depend upon nurse utilization due to the potential for blocking of discharges when all nurses are busy, we varied the *nominal bed utilization* as defined by

$$\hat{\rho}_b = \frac{\lambda_b}{B \cdot \mu_b}$$

which is the utilization rate of the bed system without considering the nurses' interaction. We tested for $\hat{\rho}_b \in \{0.65, 0.75, 0.85\}$.

It is far more challenging to get estimates regarding the nurse system. Inpatient demands for nursing care and the duration of nursing tasks are not generally tracked and documented by hospitals and HSS is no exception. Much of our project will focus on obtaining these data as well as identifying current and potential electronic sources for them. Since we don’t yet have these input values for HSS, we relied on the few published sources that we have found that report estimates for nursing tasks.

The duration of nursing tasks can vary from a couple of minutes to over 1 hour. Standards for all possible nursing interventions appear in the Nursing Interventions Classification (Dochterman and Bulechek 2004), based on the judgment of small groups of research team members that estimated the time needed for each intervention in their area of expertise. In the 2004 edition, 514 categories of interventions are listed including the required time and the minimum training required to perform each task. The estimated times are presented in broad range categories (< 15 minutes, 16 – 30 minutes, 31 – 45 minutes, over 1 hour) with the majority (52%) in the range less than 30 minutes. The only study we know that reports the actual duration of nursing tasks from direct observation is Lundgren and Segesten (2001). They studied the nurses’ use of time in a 22-bed medical-surgical ward, recording all nursing activities. The average service time found on the two sets of observations of 10 days each considered in that study was 15.3 minutes. Based on this, in our numerical experiments we used an expected time of 15 minutes, i.e. $\mu_n = 4$.

HSS, as many other hospitals, already collects some information regarding the frequency of demands for nursing care. In particular, admissions, discharges and transfers in and out of a unit are routinely recorded and can be easily obtained. They also have a separate database on the use of patient call button use which we will be accessing in our study. However, there are other nursing demands which are not currently captured such as medication and monitoring needs.
To estimate $\lambda$, the rate at which each inpatient generates a demand for nursing care, we again use the study of Lundgren and Segesten. From their data we computed an average demand rate of 0.38 requests/hr for each inpatient and so in our baseline numerical studies below, we use $\lambda = 0.4$. Since some hospital units, such as many surgical units, have more needy patients, we will also explore the impact of increasing this parameter value up to 0.5.

Since most hospitals have some finite capacity limit on the number of patients in the ED waiting for beds, we fixed the waiting room size to be 3. However, numerical tests on the size of the waiting room showed that our results were relatively insensitive to this parameter.

Perhaps the most important measure in determining nurse staffing levels is the patients' delay for nursing care. Although there are currently no existing standards of what constitutes safe delays for nursing care, the development of such standards will be part of our project with HSS. This may take the form, as in call centers, of a maximum fraction of patients who experience a delay of more than a given duration. As mentioned in the model description, the tail distribution for the waiting time conditional on the state of the system is Erlang, so we can easily compute these types of measures. For testing and illustrative purposes, we use probability of delay in our numerical examples.

The total number of combinations of the various parameter settings resulted in an experimental set of 216 cases. For each case, we used the queuing model to determine the minimum number of nurses needed to keep the probability of delay for an inpatient request to less than or equal to a specified value. Based on data and conversations with HSS personnel and the widely expressed desire of both patients and nurses to assure timely response to patient needs, we considered two levels of probability of delay targets: .05 and .10. We then used the simulation model to evaluate whether the nursing level suggested by the queuing model resulted in a probability of delay that met that target performance. To reflect the observation that for most managers the target is not a strict constraint, we counted as "unsuccessful", any case in which the simulation's estimate of probability of delay exceeded the target by more than 10%.

As indicated in Table 3.1, only 5 of the 216 cases - less than 2.5% - were unsuccessful. All of these cases corresponded to a short ALOS, i.e. 3 days, and 4 of the 5 had a nominal bed utilization level of 85%. We hypothesize that in these cases the queuing model slightly underestimates delays because of the frequency of admissions and discharges which in the simulation require more time than regular
Table 3.1: Performance of the queuing model.

<table>
<thead>
<tr>
<th></th>
<th>Unsuccessful cases</th>
<th>Simulation suggests 1 less nurse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of cases</td>
<td>Average probability of delay</td>
</tr>
<tr>
<td>P(delay &lt; 0.1)</td>
<td>1</td>
<td>13.6%</td>
</tr>
<tr>
<td>P(delay &lt; 0.05)</td>
<td>4</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

services. Also, as indicated, in another 8% of cases the simulation indicated that the target could be met with 1 less nurse than suggested by the queuing model. All of these latter cases were for fairly high value of ALOS and most were for very large units, i.e. 60 beds. Our hypothesis is that since large units have many patients, there is a higher number of patient discharges likely to be blocked when all nurses are busy. When ALOS is high, this can result in more patients occupying their beds for a considerable time after the actual length of stay is over, due to the queuing model’s discharge assumption, and hence considerably more nursing demands. Thus the queuing model is more likely to overestimate nursing needs in these cases.

Our results indicate that the queuing model will almost always suggest a staffing level that assures a high level of responsiveness. And given that there are very few nursing units that have on the order of 60 beds, our results indicate that for the vast majority of nursing units, the queuing model’s staffing estimates will be very reliable.

The simulation was programmed in Matlab 7.4 and the average time per run, using 20 beds, was 2.238 minutes using an IMac, with an Intel Core 2 Duo Processor 2.64 Ghz and 2 GB 667 Mhz RAM. This is in contrast to an average run time of 0.403 seconds for the queuing model.

3.6 Factors affecting nurse staffing ratios

Our model allows us to evaluate the impact on patient service of commonly used nurse to patient ratios, such as the minimum ratios mandated by the California legislation. In this section, we will identify the major factors that may cause the use of such ratios to under- or over-estimate the number of nurses needed to assure good performance. For this purpose, we will use some parameter values from our study hospital, HSS. As its name indicates, this is not a typical community hospital or academic
medical center, but rather a hospital which specializes in orthopedic surgery. Nonetheless, most of the key nursing tasks and practices, such as admission, discharges, medication administration, monitoring and call button use, are similar to those in other hospitals. Of course, since no one hospital unit is representative of the various conditions that exist across different hospitals and hospital units, we will vary the key parameters of our model to identify common characteristics that will likely affect the level of nurse staffing required for a high level of patient responsiveness.
3.6.1 Effect on inpatient delay

As mentioned previously, one key element in the quality of patient care is nurse responsiveness. In this subsection we use the probability of inpatient delay as the major performance metric to identify adequate levels of nurse staffing.

In Figure 3.2 (a) we show the probability of inpatient delay as a function of nurse staffing for our base case unit: \( B = 40 \) beds, nurse intensity \( \lambda_n = 0.4 \) requests per hour, average length of stay of 4.5 days, and waiting room \( WR = 4 \). We use two different levels of bed utilization \( \rho_b \in \{0.75, 0.85\} \), to capture the variation in occupancy levels that occur over the day and over the week.

To be consistent with the minimum nurse-to-patient ratio of 1 to 6 that was used in the California legislation, which mandates that the minimum ratio is required “at all times”, this unit would need to staff 7 nurses. (Recall that actual average bed utilization is higher than nominal bed utilization and so our use of nominal bed utilizations of .75 and .85 results in a substantial fraction of time that the unit is full.) In the case when nominal bed utilization is 0.85 this level of staffing will translate into a probability of inpatient delay of 4.37%. However, for the case with nominal bed utilization of 0.75 with 6 nurses (nurse-to-patient ratio of 1 to 7) the probability of inpatient delay is below 10%, which may be considered a reasonable target. Currently, HSS staffs 6 or 7 nurses in the unit under study, depending upon anticipated levels of admissions.

Studies investigating nurses’ use of time show that direct care accounts for only between 30 and 60% of nurses’ time (Jinks and Hope 2000; Lundgren and Segesten 2001). Other nurse activities, including indirect care or activities that occur away from the patient (preparing for nurse interventions, medications and therapies), rounds with the MD, report writing, communication with visitors and personal activities account for a significant fraction of nurses’ time. The numerical results for our model are consistent with those levels of nurse utilization as shown in Figure 3.2 (b).

We now look at some of the factors that are likely to affect the impact of a given level of staffing on delays. In the following analyses, we use the parameter values from our base case and vary only the factor under investigation.

Perhaps the most obvious factor that is likely to affect nursing levels consistent with timely response to patient needs is unit size, since bigger units will benefit from statistical economies of scale. The effect
of unit size is illustrated in Table 3.2, where we show the number of nurses and corresponding nurse to patient ratios needed to achieve the targets for inpatient delay for two levels of nominal bed utilization. We can see that varying the number of beds \((B \in \{20, 40, 60\})\) results in very different nurse to patient ratios. For a target probability of inpatient delay of 10% the required ratio goes from 1:5 to 1:7, and for the 5% target the ratio varies even more going from 1:4 for the smallest unit to 1:7 for the largest.

Another factor likely to affect nurse staffing levels is nursing intensity, i.e. the rate at which patients generate nursing needs. This is illustrated in the results presented in Table 3.3 in which this parameter varies from 0.35 to 0.5. It is interesting to note that for both delay targets and both levels of bed utilization, an additional nurse is needed when the request rate for nursing care increases roughly 10% from the average level reported in the literature. This seems to strongly indicate that this an important factor to consider in determining nursing levels. More specifically, units with more needy patients, e.g. many surgical units, are likely to need higher levels of nurses than other units, and more than suggested by the California mandated minimum ratios if a high level of responsiveness is desired.

We also explore the impact of average length of hospital stay since shorter hospital stays translate into more admissions and discharge events. Our results are summarized in Table 3.4. Though for most cases, ALOS does not seem to affect nursing levels, we do see an effect when ALOS is very low, i.e. 2.5. Given the results of some of our simulation runs which indicated that our queuing model may occasionally underestimate delays and hence staffing needs when ALOS is very short, we believe that...
Table 3.3: Staffing needed to achieve delay targets - Nursing intensity effect.

(a) Nurse intensity effect

<table>
<thead>
<tr>
<th>Nurse intensity</th>
<th>Nursing demand</th>
<th>( \lambda_n ) (requests/hr.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(\text{delay} &lt; 0.1) )</td>
<td>0.35</td>
<td>0.40</td>
</tr>
<tr>
<td>6 (1 : 7)</td>
<td>6 (1 : 7)</td>
<td>7 (1 : 6)</td>
</tr>
<tr>
<td>( P(\text{delay} &lt; 0.05) )</td>
<td>7 (1 : 6)</td>
<td>7 (1 : 6)</td>
</tr>
</tbody>
</table>

\( B = 40, WR = 4, \bar{\rho}_n = 0.75, \text{ALOS} = 4.5 \)

(b) Nurse intensity effect

<table>
<thead>
<tr>
<th>Nurse intensity</th>
<th>Nursing demand</th>
<th>( \lambda_n ) (requests/hr.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(\text{delay} &lt; 0.1) )</td>
<td>0.35</td>
<td>0.40</td>
</tr>
<tr>
<td>6 (1 : 7)</td>
<td>6 (1 : 7)</td>
<td>7 (1 : 6)</td>
</tr>
<tr>
<td>( P(\text{delay} &lt; 0.05) )</td>
<td>7 (1 : 6)</td>
<td>7 (1 : 6)</td>
</tr>
</tbody>
</table>

\( B = 40, WR = 4, \bar{\rho}_n = 0.85, \text{ALOS} = 4.5 \)

this factor may have a more pervasive impact than indicated by this table.

To further illustrate this last point, and to demonstrate the potential effect of several factors that are likely to exist concurrently for a single unit, we examine nurse-staffing levels for a 30 bed unit with a low ALOS of 3 days and a relatively high rate of demand for nursing care of 0.45. In this case, in order to meet the 10% target the proposed 1:6 ratio won’t be sufficient if the nominal bed utilization is greater than 0.7 and 6 nurses will be required. The probability of inpatient delay for this case, staffing 6 or 7 nurses is shown in Figure 3.3.

3.6.2 Effect on emergency department congestion

The nurse-staffing model can also be useful for estimating the performance of the bed system, measured either by the expected delay or expected number of patients waiting for a bed. As we already mentioned, our model is not equivalent to an \( M/M/c \) queuing system because service times are not independent. The possibility of admission or discharge blocking due to nurse unavailability adds another source of variability that can have a significant adverse affect on system performance.

Figure 3.4 shows the expected number of patients waiting in the base case unit, assuming no blocking or balking is allowed. We contrast the results from our nurse staffing model using equation (3.4), with the results from a standard equivalent \( M/M/c \) queueing model with an average utilization equal to the
Table 3.4: Staffing needed to achieve delay targets - ALOS effect.

(a)

<table>
<thead>
<tr>
<th>Average length of stay effect</th>
<th>Average length of Stay</th>
<th>ALOS (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5</td>
<td>4.5</td>
</tr>
<tr>
<td>P(delay &lt; 0.1)</td>
<td>7 (1:6)</td>
<td>6 (1:7)</td>
</tr>
<tr>
<td>P(delay &lt; 0.05)</td>
<td>7 (1:6)</td>
<td>7 (1:6)</td>
</tr>
</tbody>
</table>

\( B = 40, WR = 4, \hat{\rho}_b = 0.75, \lambda_n = 0.4 \)

(b)

<table>
<thead>
<tr>
<th>Average length of stay effect</th>
<th>Average length of Stay</th>
<th>ALOS (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5</td>
<td>4.5</td>
</tr>
<tr>
<td>P(delay &lt; 0.1)</td>
<td>7 (1:6)</td>
<td>7 (1:6)</td>
</tr>
<tr>
<td>P(delay &lt; 0.05)</td>
<td>7 (1:6)</td>
<td>7 (1:6)</td>
</tr>
</tbody>
</table>

\( B = 40, WR = 4, \hat{\rho}_b = 0.85, \lambda_n = 0.4 \)

It is very interesting to note that even with nursing levels that result in reasonable inpatient delays, i.e. below the 10% target, the expected backlog of patients waiting for a bed can be significantly higher than that predicted from the \( M/M/c \) queue. For instance, the probability of inpatient delay when \( \rho_b = 0.9 \), in Figure 3 is 5.36% yet the expected number of patients waiting for a bed for this case is 6 which is 2.01 above the prediction of the \( M/M/c \) model. This discrepancy can make the difference between having the ED going on diversion or not.

3.7 Heuristic

Though the two-dimensional models discussed above are tractable and easy to solve numerically, it would be preferable to have a simpler model that focuses on the nursing system to make it easier for hospital managers to use on a regular basis. For this reason, we developed and tested a one-dimensional
Figure 3.3: Combined impact of nursing intensity and ALOS effects.

\[ B = 30, \ WR = 3, \ ALOS = 4.5, \ \lambda_n = 0.45 \]

modified finite source queuing system (M/M/N/B mod). In order to do this, we assume a fixed number of occupied beds \( \bar{B} = E[\text{occupied beds}] \) that generate nursing requests at a rate \( \lambda_n \) each, plus an outside source of requests that arrive at rate \( \lambda_b \), representing the workload associated with admissions.

\[
\lambda_i = \begin{cases} 
(\bar{B} - i) \cdot \lambda_n + \lambda_b & \text{if } i < \bar{B} \\
\lambda_b & \text{if } i \geq \bar{B}
\end{cases} \tag{3.14}
\]

\[
\mu_i = \min(i, N) \cdot \mu_n \tag{3.15}
\]

The question is how to compute the expected number of occupied beds. If we assume there is no blocking from all nurses being busy we have

\[
\bar{B} = \rho_b \cdot B \tag{3.16}
\]

which gives us a lower bound for the actual number when the interaction between nurses and discharges is considered. This is not a bad assumption if the nurse system is lightly loaded but becomes a terrible one when bed utilization approaches 1 as illustrated in the graphs in the previous sections. In those situations it appears better to use \( \bar{B} = B \) since in practice most of the time we will see a full unit.
Figure 3.4: Impact of nurse staffing levels on ED overcrowding.

Base unit. $B = 40$, $WR = 4$, $N = 7 \ (1:6)$, ALOS = 4.5, $\lambda_n = 0.4$

We propose a two-step heuristic, where we start with a lightly loaded system, assuming the fixed source $\overline{B}$ as in equation (3.16). Then, we adjust $\overline{B}$ using the probability of finding all nurses busy in this modified model.

The heuristic works as follow:

(a) Define $\overline{B}^1 = \rho_b \cdot B$

(b) Solve the steady state probabilities for the death-and-birth process defined by the transition rates of (3.14) and (3.15). We call these probabilities $p^{h1}_i$.

(c) Compute the probability of having all nurses busy, where departures would have been blocked in the coupled model.

$$P^{h1}(block) = \sum_{j=N}^{\infty} p^{h1}_j$$  \hspace{1cm} (3.17)

(d) Compute an approximation for the utilization of the bed system, using the logic of equation (3.2).

$$\rho^{h1}_b = \min \left\{ 1, \frac{\lambda_b}{\overline{B} \cdot \mu_b \cdot (1 - P^{h1}(block))} \right\}$$  \hspace{1cm} (3.18)

(e) Define $\overline{B}^2 = \rho^{h1}_b \cdot B$
(f) Solve the steady state probabilities using this new fixed source. We call these probabilities $p_i^{k^2}$.

The procedure proposed involves twice solving a one-dimensional queuing model with a transition matrix of size $(B + 1) \times (B + 1)$ and can be easily implemented.

The proposed heuristic allows the approximation of the key performance indicators needed to guide nurse staffing decisions. For instance, inpatient delay can be approximated by the probability of having all nurses busy, and from this modified model we can compute nurse utilization and even tail probabilities.

Though the heuristic only considers the workload for nurses, we can still use it to say something about the bed system. Bed utilization can be approximated by equation (3.18), and the degree to which it differs from the nominal bed utilization $\bar{p}_b$ provides important information on the actual bed congestion.

Figure 3.5 presents the approximated nurse utilization and probability of inpatient delay from the heuristic together with the results for the infinite and finite models.

We can see that for both performance indicators the heuristic does a very good job of approximating the two-dimensional model.

Since the heuristic is one-dimensional, it can be modified to incorporate a priority service discipline so that staffing levels can be determined on the basis of virtually immediate response to emergent patient needs.

### 3.8 Discussion

In this paper, we have presented a two-dimensional queuing model to help guide decisions on nurse staffing based on providing a timely response to patient needs. We have demonstrated the model's reliability in identifying good staffing levels across a broad range of parameters corresponding to actual hospital units of different types. Our model represents the crucial interaction between the nurse and bed systems and therefore includes the nursing workload due to admissions, discharges and transfers, as well as the observed impact of nursing availability on bed occupancy levels. From a methodological perspective, this queuing model is unique in its assumption of a finite population of customers which is a random variable dependent on the dynamics of another queuing system. From a practical perspective,
Figure 3.5: Heuristic performance.

![Graph showing performance of queuing models with varying nurse ratios]

Base unit. $B = 40$, $WR = 4$, $\hat{\rho}_n = 0.75$, $ALOS = 4.5$, $\lambda_n = 0.4$

This is the first attempt to develop a quantitative model that can be used by hospital managers to evaluate the direct impact of their nurse staffing decisions on both patient delays and delays for inpatient beds. The model is very flexible, and conforms to the guidelines issued by the ANA for an optimal nurse staffing methodology. Furthermore, the model uses input parameters which correspond to clinical characteristics that hospital managers are familiar with such as admission rates, ALOS, and nursing intensity. Given the speed with which it can be solved, the queuing model could easily be incorporated into a software package for nurse scheduling, an application which is being implemented by many hospitals.

Numerical results using our model demonstrate the problem in using rigid nurse-to-patient ratios across a broad range of hospital units. Specifically, we have shown that unit size, nursing intensity, bed utilization and average length of stay will affect nurse staffing levels that are consistent with timely response to patients. Our findings show that unit size and nursing intensity are particularly critical parameters affecting staffing levels. The implications of these findings are extremely important from both a patient safety and hospital cost perspective. Using a 1:6 ratio in smaller clinical units with patients who are relatively needy and have short lengths of stay, e.g. gynecology and vascular surgery, may result in understaffing which has been shown to lead to higher levels of medical errors. On the other hand, in larger units, e.g. 40 or more beds with less needy patients and moderate or long lengths
of stay, e.g. many general medicine units, a 1:6 ratio may lead to overstaffing. This is also likely to be true for many units for night shifts when bed utilization is lower, patients are generating fewer demands, and both admissions and discharges are far less prevalent. Weekend shifts are likely to fall into this category as well since they are generally characterized by lower occupancy levels and fewer admissions. Based on an average annual salary of $60,000, the use of even one unnecessary nurse on a 12-hour shift for one unit can result in a wasted expense of $300,000 per year.

Our results also point to the need for identifying appropriate standards of responsiveness to patient needs. As indicated by some of our results, the probability of delay target itself is often a determinant of required staffing levels, yet there is no literature on the impact of delay in nursing response on clinical outcomes or patient satisfaction. One of the goals of our larger research project is to solicit opinions from nursing professionals on such standards.

Our model can also be very useful for providing insights not directly related to nurse staffing. There has been a great deal of concern about long waits for inpatient beds by patients in the emergency room and, as mentioned above, hospitals often react to these delays by placing patients in alternative units or going on ambulance diversion. As we've demonstrated above, our model can be used to determine whether lack of inpatient beds or nursing staff is the bottleneck responsible for ED delays for a particular nursing unit.

Of course, no model is a perfect representation of reality and this one is no exception. One limitation of our model is the assumption of a homogeneous workforce. In reality, some hospitals use both registered nurses (RNs) as well as nurses with lower levels of training, e.g. licensed practical nurses (LPNs) who cannot perform all types of interventions. In those cases our model may underestimate actual nursing delays (and hence underestimate the number of required nurses) since some requests will have to wait until the appropriate provider becomes available. Another possible complication is that there may be a priority scheme for responding to nursing demands depending on the acuity or urgency of the service being requested. In our preliminary discussions with HSS personnel, we have found that though there are no formal priority rules, there are clearly cases in which certain patient needs are attended to ahead of others. Our project will try to identify these as well as determine the importance of incorporating priorities into a future version of the model. As mentioned previously, this could be done by modifying our one-dimensional heuristic model.
Our findings clearly point to the need for more accurate data on critical parameters such as demand rates for nursing care and times for various nursing tasks. This is the focus of the next phase of our work with HSS. We are examining the data that can be captured from both existing and potential capabilities of electronic patient care support systems in order to make recommendations to hospitals and software vendors on future uses of hospital IT systems to support queuing-based staffing decisions. At HSS, we are working to get specific unit information to calibrate the model parameters. Once this is done, we will use the queuing model to estimate the impact of different staffing levels on the performance of the system. The hospital will then modify their nurse staffing for the unit based on these estimates and we will collect various service and quality performance measures from both before and after the staffing change to determine the usefulness of the model.

We are hopeful that our model will result in more cost-effective staffing decisions and that as a result of this project, hospitals will start to develop IT capabilities to keep track of the nursing-related data needed to adopt this type of staffing methodology. Given the link between adequate nursing levels and patient safety as well as the imperative for cost efficiency, we believe that the adoption of queuing methodology can be an important innovation in hospital management.
4 Optimizing Bed Allocation and Admission Rule of a Stroke Unit

4.1 Introduction

Stroke is a condition with high incidence and mortality rate, leaving a large proportion of survivors with significant residual physical, cognitive, and psychological impairments. However, there is a considerable amount of scholarly work which shows that the timely and appropriate treatment of acute stroke patients improves both short-term and long-term survival and functionality of the patient. In many large hospitals, particularly academic medical centers, acute stroke patients are treated in a stroke unit, a small unit with specialized care and a higher level of nurse staffing, which is used to closely monitor unstable stroke patients. After being stabilized, these patients are often transferred to a general neurological unit until they are deemed ready for discharge either to home or to a rehabilitative facility. Stroke patients with lesser acuity levels may be admitted directly to the general neurology unit, which is also used for non-stroke patients presenting a variety of neurological conditions such as epilepsy or multiple sclerosis.

The benefits of treatment in a stroke unit are well established in the medical literature. Timely admission to a stroke unit of a patient with acute ischemic stroke decreases the mortality rate (Langhorne et al. 1993, Rudd et al. 2005) and the need for institutional long-term care after stroke (Indredavik et al. 1991). All patients with acute stroke should be admitted to a stroke unit immediately after the initial assessment and no later than 48 hours from onset (Candelise et al. 2007, NHS 2008). However, in recent years hospital utilization has reached levels at which the rationing of critical care beds, including stroke beds, has become unavoidable (Sinuff et al. 2004), jeopardizing the well-being of patients needing this level of care. Identifying more effective policies for using these critical hospital resources is therefore extremely important in minimizing adverse clinical outcomes.

The major objective of this research is to help identify an optimal admission rule to the stroke unit and how that decision may be related to the relative capacities of the two units involved in the recovery of stroke patients. We model the flow of stroke patients as a tandem queue with blocking. We assume that all patients are first treated in the stroke unit and have a length of stay (LOS) that is comprised of two phases. The first phase, stabilization, must be completed in the stroke unit. Patients are moved
to the neurology unit for the second phase, recovery, if and when there is a neurology bed available. If there is no bed available by the time the second phase is completed, we assume that the patient is discharged from the stroke unit.

This work is related to Shmueli et al. (2003) which studied the impact of alternative admission rules to an intensive care unit (ICU) on the expected incremental number of lives saved. That paper assumed that there was ample capacity in the step-down unit to which patients were transferred from the ICU. Here we address a similar problem but for the care chain of stroke patients. We will consider not only the performance of the system under different admission policies but also the impact of modifying the relative capacities of the two units involved.

Capacity planning for critical care units has usually been studied using simulation models (Ridge et al. 1998, Costa et al. 2003). Akcali et al. (2006) used a network flow approach to help allocate hospital beds to different units. The work presented here is most closely related to de Bruin et al. (2007) which explored the impact of alternative bed allocations over a two-stage care chain for emergency cardiac patients using a two-dimensional queueing model. As is described in the next section, the model used in that paper is a special case of our stroke patient flow model.

In the next section we describe our model which is simple enough to be solved numerically, but flexible enough to allow for the evaluation of alternative admissions policies and capacity decisions. To help develop and test our model to identify good capacity and admissions decisions, we will use data from the stroke service at New York Presbyterian Hospital.

### 4.2 Model description

At New York Presbyterian and most other hospitals, stroke patients generally first arrive to the emergency department (ED) where they are diagnosed and the severity level is assessed using the NIH Stroke Scale (NIH 2008). Acute stroke patients are admitted to the stroke unit if and when a “stroke bed” becomes available and are transferred to the neurology unit when they are stabilized if a neurology bed is available. If a neurology bed isn’t available, they remain in the stroke unit until either a neurological bed becomes available or they are ready for discharge. The neurology unit is also used for non-stroke neurological patients and there are no reserved beds for the stroke patients. However, we will also explore the case, which we’ve encountered in other hospitals, where there are a specific number of beds
in the neurological unit reserved for patients coming from the stroke unit. Patients needing a stroke bed when the stroke unit is full may wait in the ED or be "overflowed" to the neurology floor, particularly if their condition is not considered too severe. Non-stroke neurological patients may be placed in a different ward if the unit is full when they arrive. Figure 4.1 illustrates the general situation. In the hospital under study there are four beds in the stroke unit and 35 in the neurology unit.

Assuming that the distribution of the LOS for neurology patients is different than the LOS of stroke patients in the neurology unit, a model of the system described above requires at least three dimensions: the number of stroke patients in the stroke unit, the number of stroke patients in the neurology unit and the number of neurology patients.

Below, we analyze some special cases of this model that allow us to use only two dimensions. This will allow for more efficient analysis of the impact of capacity and admission decisions.

4.3 Alternative modeling of the two stage chain

4.3.1 No shared capacity

We can simplify the problem by assuming there is no shared capacity in the neurology unit, i.e. stroke patients have dedicated beds in the neurology unit. If we further assume that a new arrival to the
stroke unit is admitted if a bed is available in the stroke unit but lost if not we recover the model presented in de Bruin et al. (2007), where $x_1$ ($x_2$) are the number of patients in phase 1 (phase 2). The assumption of a loss system for patients needing the stroke unit is reasonable given the urgency of admitting these patients as soon as possible. The LOS of each phase is assumed to be exponential with rates $\mu_1$ and $\mu_2$ respectively. Phase 1 must be completed in the stroke unit while phase 2 can be completed anywhere in the chain. Let $B_s$ be the number of beds in the stroke unit and $B_r$ the number of dedicated stroke beds in the neurology unit. Arrivals to the system follow a Poisson process with rate $\lambda_s$.

This model allows us to explore the impact of the relative capacities on the probability of blocking arrivals, as done in de Bruin et al. (2007) but it may also be used as the starting model for evaluating admission rules as in Shmueli et al. (2003), considering that the benefits of being admitted to the stroke unit are not homogeneous across patients. We will explore this issue in section 4.5.

### 4.3.2 Single LOS distribution in neurology unit

In order to have a model that better reflects the organizational structure in many hospitals, we need to consider the case where the capacity of the neurology unit is shared by both stroke and non-stroke patients. As explained above this may lead to a very complex model. However, assuming a loss system for both units and that the LOS of non-stroke patients and the phase two recovery of stroke patients follow the same exponential distribution, we can reduce the system to a two-dimensional model. Since many hospitals place neurological (and other) patients “off-service” when the appropriate unit is full, the assumption of a loss system is not unreasonable. Also, assuming a single LOS distribution for all patients in the unit is consistent with the practice of tracking LOS by unit rather than by patient type. We will relax this latter assumption below.

The space state can be characterized by $(x_1, x_2)$, where $x_1$ is the number of patients in phase 1 and $x_2$ is the number of patients in the second phase plus the non-stroke patients. Arrivals follow a Poisson distribution with rate $\lambda_s$ for the stroke patients and rate $\lambda_n$ for the neurology patients. We further assume that the duration of each phase for stroke patients is exponential with rates $\mu_1$ and $\mu_2$ respectively and that neurology patients have a LOS exponentially distributed with rate $\mu_n = \mu_2$.

Again, we assume that phase 1 must be completed in the stroke unit while phase 2 can be completed
anywhere in the chain. Let \( B_s \) be the number of beds in the stroke unit and \( B_n \) the number of beds in the neurology unit.

Our model allows for the implementation of a "cut-off" priority system in the neurology unit (Schaack and Larson 1986) so that we do not allow the admission of non-stroke patients if there are fewer than \( B_r < B_n \) empty beds. This type of policy can be very useful in assuring that new stroke arrivals are not blocked because of downstream blockage due to unavailability of neurology beds.

This two-dimensional model can be easily solved using numerical methods. The transitions are presented below.

**Transition rates from state \((i, j)\)**

- \((i < B_s, j < B_n - B_r)\)
  Arrivals to the stroke units occur with rate \( \lambda_s \) leading the system to state \((i + 1, j)\). The end of the first phase of stroke care occurs with rate \( i \cdot \mu_1 \) leading to state \((i - 1, j + 1)\). Arrivals of non-stroke patients occur with rate \( \lambda_n \) leading to state \((i, j + 1)\). The end of the second phase of stroke care (or service for non-stroke patients) occurs with rate \( j \cdot \mu_2 \) leading to state \((i, j - 1)\).

- \((i < B_s, j \geq B_n - B_r, i + j < B_s + B_n)\)
  Arrivals of non-stroke patients are blocked. All the other transitions occur as described above.

- \((i = B_s, j < B_n - B_r)\)
  Stroke patients cannot enter the unit and are blocked. All other transitions occur as described above.

- \((i = B_s, j \geq B_n - B_r)\)
  Neither stroke patients nor non-stroke patients can enter the system. End of first phase and discharges occur as described above.

- \((i + j = B_s + B_n)\)
  Neither stroke patients nor non-stroke patients can enter the system. End of first phase and discharges occur as described above.

This model is a first approach for dealing with situations of two stages of recovery considering shared capacity in the second stage, considering the possibility of having reserved capacity for stroke patients.
Even though it is unlikely that the LOS of neurology patients will have the same distribution as the second phase of recovery, this is a reasonable model from a practical perspective since most hospital data only captures ALOS by unit without distinguishing patient type.

4.3.3 Iterative approximation

In order to deal with situations in which the service rate of non-stroke patients $\mu_n$ cannot be considered equal to $\mu_2$ we construct an approximation by considering the two units independently and linking them with an iterative procedure. The idea is as follows: first we model the stroke unit assuming an infinite neurology unit, and compute the probability of new stroke arrivals being blocked (which is a lower bound of the actual probability due to possible blocking from the neurology unit). Then, we model the neurology unit as a multi-server queue with two types of patients, the patients from the stroke unit with arrival rate modified by the blocking probability found from the stroke unit model, and the non-stroke patients.

The iterative procedure, as the two-dimensional model presented above, allows for the implementation of a "cut-off" priority, but it also allows for the implementation of a different type of type reserved capacity. We can specify a $B_r < B_n$ such that no neurological patients can be admitted if there are $B_n - B_r$ neurological patients in the unit. This is a policy that has been implemented in some hospitals to help reduce the blocking probability of stroke patients. Figure 4.2 illustrates the two versions of the iterative approximation.

Description of the iterative approximation with "cut-off" priority system

(a) Stroke unit: We compute the probability of blocking new arrivals using a modified Erlang loss formula to allow for a non-integer number of servers, $B$, which may result from solving for the expected queue length in the other part of our iterative procedure. If we have $B$ available beds the factorial in the numerator is approximated by the Gamma function, while the sum of the denominator is computed up to the floor of the available capacity ($\lfloor B \rfloor$) to which is added the remaining fraction corresponding to the next value of the series ($i = \lfloor B \rfloor$).

We start the iterative procedure setting the initial available capacity $B^0 = B_\lambda$. For the $k^{th}$ iteration, if the available stroke unit capacity is $B^k$ the probability of blocking stroke patients
will be given by:

$$P^k(\text{block}) = \frac{e^{\rho B^k}}{\Gamma(B^k+1)} \sum_{i=0}^{\lfloor B^k \rfloor} \frac{\rho^i}{i!} + (B^k - \lfloor B^k \rfloor) \cdot \frac{\rho^{\lfloor B^k \rfloor}}{\lfloor B^k \rfloor!}$$  \hspace{1cm} (4.1)

Where $\rho = \frac{\lambda_{s}}{\mu_{s}}$.

With this probability, the effective rate of patients needing the neurology unit, when the available capacity at the stroke unit is $B^k$ is given by $\lambda_{s} \cdot (1 - P^k(\text{block}))$.

(b) **Shared capacity in the neurology unit with “cut-off” priority system:** We model the neurology unit beds which can be used by both types of patients as a two-dimensional Markov system with space state $(x_{s}, x_{n})$, where $x_{s}$ is the number of patients coming from the stroke unit (with a service rate $\mu_{2}$) and $x_{n}$ the external neurology patients (with a service rate $\mu_{n}$).

We assume the incoming flows are Poisson processes with an arrival rate equal to $\lambda_{n}$ for non-stroke patients and $\lambda_{s} \cdot (1 - P^k(\text{block}))$ for the stroke patients. The number of available beds is $B_{n}$, but we assume there are $B_{s}$ additional beds that can be used only for stroke patients representing the capacity at the stroke unit (that may be used if needed).

We assume a loss system for non-stroke patients who will be blocked if there are fewer than $B_{r}$ empty beds $(x_{s} + x_{n} \geq B_{n} - B_{r})$. Using $\Pi_{ij}^{k}$ to denote the steady state probability of finding the system at state $(x_{s} = i, x_{n} = j)$ at the $k^{th}$-iteration we can compute $L_{q}^{k}$ as:

$$L_{q}^{k} = \sum_{i+j>B_{n}} ((i + j) - B_{n}) \cdot \Pi_{ij}^{k}$$  \hspace{1cm} (4.2)

The available stroke unit capacity at the next iteration will be given by $B^{k+1} = B_{s} - L_{q}^{k}$

**Description of the iterative approximation with reserved capacity**

(a) **Stroke unit:** Same as in the previous case. The effective rate of patients needing the neurology unit, when the available capacity at the stroke unit is $B^k$ is given by $\lambda_{s} \cdot (1 - P^k(\text{block}))$.

(b) **Reserved capacity:** In some situations, it may be of interest to know the effect of reserving beds in the neurology unit for patients arriving from the stroke unit, i.e. specifying a $B_{r} < B_{n}$ such that no neurological patients can be admitted if there are $B_{n} - B_{r}$ neurological patients in the unit. This is a policy that has been implemented in some hospitals to help reduce the blocking probability of stroke patients.
In the iterative approximation, we can consider the reserved capacity by modeling it as an intermediate system with \( B_r \) servers with a service rate of \( \mu_2 \). Arrivals are assumed to be Poisson with an arrival rate of \( \lambda_s \cdot (1 - P^k(block)) \). Using the Erlang loss formula the probability \( (p^k(B_r)) \) of blocking arrivals in the reserved bed system is computed.

(c) Shared capacity in the neurology unit: We model the neurology unit beds which can be used by both types of patients as a two-dimensional Markov system with space state \((x_s, x_n)\), where \( x_s \) is the number of patients coming from the stroke unit (with a service rate \( \mu_2 \)) and \( x_n \) the external neurology patients (with a service rate \( \mu_n \)).

We assume the incoming flows are Poisson processes with an arrival rate equal to \( \lambda_n \) for non-stroke patients and \( \lambda_s \cdot (1 - P^k(block)) \cdot p^k(B_r) \) for the stroke patients (the blocked flow from the reserved capacity).

The number of available beds is \( B_n - B_r \), but we assume there are \( B_s \) additional beds that can be used only for stroke patients representing the capacity at the stroke unit (that may be used if needed). We assume a loss system for non-stroke patients who will be blocked if \( (x_s + x_n \geq B_n - B_r) \). Using \( \Pi_{ij}^k \) to denote the steady state probability of finding the system at state \((x_s = i, x_n = j)\) at the \( k^{th} \)-iteration we can compute \( L_q^k \) as:

\[
L_q^k = \sum_{i+j>B_n} ((i+j) - B_n) \cdot \Pi_{ij}^k
\]  

(4.3)

The available stroke unit capacity at the next iteration will be given by \( B_{s+1} = B_s - L_q^k \)

Convergence of the iterative approximation

Each step of the iterative procedure involves finding the blocking probability for one or two queuing systems, in case we have reserved capacity, and the expected queue of a two dimensional queue.

Equation 4.1 is decreasing in \( B \). The criterion for convergence used is that the difference of this blocking probability resulting from two consecutive iterations \( |P^k(block) - P^{k+1}(block)| \) is less than \( 10^{-8} \). The iterative procedure converges quickly requiring fewer than 8 iterations in all 18,900 tested cases \((B_s \in \{2, \ldots, 9\}, B_n \in \{B_s, \ldots, B_s + 20\}, \lambda_s, \lambda_n \in \{0.5,1,1.5,2,2.5,3\}, \mu_1 = 0.5, \mu_2 = 0.2, \mu_n \in \{\mu_2, \frac{\mu_2}{2}, \frac{\mu_2}{3}, \frac{\mu_2}{4}\})\).
4.3.4 Performance of the models and approximations

We tested the performance of the iterative procedure approximation using as our major performance metric the fraction of stroke patients that are blocked from access to the stroke unit. We use the 2-dimensional model as a benchmark since that is an exact model for the case $f_x = H_2$.

Figure 4.3 (a) shows the performance of the iterative procedure and the two-dimensional model assuming an average LOS for neurology patients equal to that for the second phase recovery of stroke patients, for three different stroke unit capacities, varying the arrival rate of stroke patients from 0.6 up to 2.5 patients/day. The biggest difference in the computed probability of blocking stroke patients using the iterative procedure vs. the two-dimensional model in these cases is less than 0.007, indicating that the iterative approximation is very accurate.

Figure 4.3 (b) shows the impact of increasing the average LOS for neurology patients using the iterative procedure. Not surprisingly, we see that longer stays lead to higher probability of blocking stroke patients, especially for higher arrival rates of those patients.
Figure 4.3: Performance of iterative approximation.

(a) Iterative procedure vs. two-dimensional model ($\mu_2 = \mu_n$)

\[ \mu_1 = \frac{1}{2}, \quad \mu_2 = \mu_n = \frac{1}{5}, \quad B_n = 30, \quad \lambda_n = 5 \]

(b) Impact of longer stay neurology patients - iterative approximation

\[ \mu_1 = \frac{1}{2}, \quad \mu_2 = \frac{1}{5}, \quad B_n = 6, \quad B_n = 30, \quad \lambda_n = 5 \]
Finally, we check the performance of the iterative procedure identifying the impact of reserving beds at the neurology unit for stroke patients.

Figure 4.4 shows the probability of blocking stroke patients when reserving an increasing number of beds at the neurology unit \((B_s = 4, B_n = 20, \lambda_s = 1, \lambda_n = 3, \mu_1 = \frac{1}{2}, \mu_2 = \mu_n = \frac{1}{8})\) for the iterative procedure with “cut-off” priority system, the two-dimensional model and the iterative procedure with reserved capacity for patients coming from the stroke unit. We can see that the iterative procedure does an excellent job approximating the 2-dimensional model, identifying as two the number of reserved beds needed to achieve a \(P(block) < 0.1\). By using the iterative procedure with reserved capacity we require nine reserved to achieve a \(P(block) < 0.1\), in section 4.4.2 we will discuss in detail the two alternative policies for reserving capacity for stroke patients.

### 4.4 Optimizing the capacity

The proposed models and iterative procedure can help to quickly determine the minimum capacity needed in order to achieve a specified level of performance. First, we use equation 4.1 to find the minimum capacity needed at the stroke unit \(B_s^*\). The minimum capacity needed at the neurology unit...
can be determined without considering the flow of non-stroke patients since we are only concerned with the impact of the neurology unit capacity on stroke patients.

To identify the number of beds needed in the neurology unit, \( B_n^* \), we use a queuing model with arrival rate \( \lambda_s \), a service rate \( \mu_2 \), and a maximum queue of \( B_s \), to find the minimum number of beds that will result in an expected queue less than \( ([B_*] - B_*^*) \), defining \( B_*^* \).

Table 4.1 (a) shows the minimum capacities for the two units, considering \( \mu_1 = 0.5 \) and achieving a probability of blocking stroke patients less than 10%. The arrival rate for the stroke unit \( \lambda_s \) goes from one to five patients a day, and the average duration of the second phase of recovery goes from 5 to 9 days.

This procedure assumes the hospital manager wants to use the fewest possible stroke beds. However, this doesn’t necessarily represent the trade-off between neurology and stroke unit beds, i.e. the cost of having a neurology bed may not be zero. The procedure can be easily adapted to include any relative cost situation by searching the capacity combination minimizing the cost while keeping the blocking probability below the specified target.

Table 4.1 (b) shows the optimal capacities considering a stroke unit bed is twice as expensive as a neurology unit bed, reflecting the fact that the nurse-to-patient ratio is twice as high on the neurological unit.

### 4.4.1 Impact of sharing capacity with neurology patients

Having a neurology unit shared with non-stroke patients is a very common organizational structure for stroke care. Pooling the neurology beds should allow for a better use of the resource, and should result in a better performance for the stroke patients given that non-stroke patients are blocked if the neurology unit is full while stroke patients can still enter through the stroke unit. However, if the capacity of the neurology unit is not larger than the minimum number of beds needed to achieve the performance target without sharing them with non-stroke patients, this practice may be dangerous.

Figure 4.5 shows the probability of blocking stroke patients with a small neurology unit (the minimum capacity without considering non-stroke patients is \( B_s = 7, \ B_n = 13 \) for a \( P(block) < 0.1 \)), using the iterative procedure. We can see that the system is especially vulnerable when the LOS for neurology
Table 4.1: Minimum capacity to achieve performance target.

(a) Minimizing use of stroke beds

<table>
<thead>
<tr>
<th>((B_s, B_n))</th>
<th>(\text{LOS second phase})</th>
<th>1/(\mu_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_s)</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>(4, 8)</td>
<td>(4, 10)</td>
</tr>
<tr>
<td>2</td>
<td>(7, 11)</td>
<td>(7, 13)</td>
</tr>
<tr>
<td>3</td>
<td>(9, 16)</td>
<td>(9, 19)</td>
</tr>
<tr>
<td>4</td>
<td>(11, 22)</td>
<td>(11, 26)</td>
</tr>
<tr>
<td>5</td>
<td>(13, 27)</td>
<td>(13, 32)</td>
</tr>
</tbody>
</table>

\(P(\text{block}) < 0.1\) \(\mu_1 = \frac{1}{2}\)

(b) Capacities - Cost of stroke unit bed is twice as much as the cost of neurology unit bed

<table>
<thead>
<tr>
<th>((B_s, B_n))</th>
<th>(\text{LOS second phase})</th>
<th>1/(\mu_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_s)</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>(7, 3)</td>
<td>(7, 4)</td>
</tr>
<tr>
<td>2</td>
<td>(9, 8)</td>
<td>(9, 10)</td>
</tr>
<tr>
<td>3</td>
<td>(12, 12)</td>
<td>(12, 15)</td>
</tr>
<tr>
<td>4</td>
<td>(16, 14)</td>
<td>(16, 18)</td>
</tr>
<tr>
<td>5</td>
<td>(18, 19)</td>
<td>(18, 24)</td>
</tr>
</tbody>
</table>

\(P(\text{block}) < 0.1\) \(\mu_1 = \frac{1}{2}\)
patients is larger than the average duration of the second phase recovery of stroke patients ($\mu_2 > \mu_n$).
However, if we have ample capacity at the second unit, stroke patients are better off sharing this capacity than using only the minimum number of beds needed to meet the target.

### 4.4.2 Evaluation of alternative policies for reserving beds for stroke patients

In this section we study the impact of reserving a specific number of beds in the neurological unit for patients coming from the stroke unit. The models and procedures presented in this work allow for the analysis of two different types of reserved capacity.

First we consider the situation in which $B_r$ beds are reserved for patients coming from the stroke unit and non-stroke patients are blocked if there are fewer than $B_r$ available beds in the neurology unit (without considering the type of patients).

Using the two-dimensional model described in section 4.3.2 we can deal with situations for which it is reasonable to assume a single LOS distribution for all patients in the neurology unit. Figure 4.6 (a) shows the probability of blocking stroke patients for average LOS of five, nine and fourteen days ($B_s = 6, B_n = 20, \lambda_s = 1.5, \lambda_n = 5, \mu_1 = 0.5, \mu_n = \mu_2$). The number of reserved beds needed to
achieve a $P(\text{block}) < 0.1$ are 0, 2 and 7 for ALOS 5, 9 and 14 respectively.

In cases for which $\mu_n < (>)\mu_2$, we can compute an upper (lower) bound for $P(\text{block})$ using the two-dimensional model with two different service rates, first we assume $\mu_n^1 = \mu_2^1 = \mu_2$ and then we assume $\mu_n^2 = \mu_2^2 = \mu_n$, and we can identify the number of reserved beds needed to achieve a specified performance target as explained above. However, the gap between the computed probabilities of blocking stroke patients can be quite large, so we need to have a better approximation. We use the modified iterative procedure to compute $P(\text{block})$ with $\mu_2$ and $\mu_n$. Figure 4.6 (b) shows the $P(\text{block})$ of the upper and lower bound and the iterative procedure with "cut-off" priority system ($B_s = 6$, $B_n = 20$, $\lambda_s = 1.5$, $\lambda_n = 5$, $\mu_1 = \frac{1}{2}$, $\mu_2 = \frac{1}{5}$, $\mu_n = \frac{1}{9}$).

The second type of reserved capacity that can be analyzed is the type of policy we use in the iterative procedure with $B_r$ reserved beds, where the maximum number of non-stroke patients in the neurology unit is limited by $B_n - B_r$. The two versions of the iterative procedure allow us to compare the impact of using this type of reserved capacity vs. using a "cut-off" priority system. Figure 4.7 (a) shows the effect of implementing a "cut-off" priority system, while Figure 4.7 (b) shows the effect of reserving beds in the neurology for three different arrival rates of the non-stroke patients ($B_s = 6$, $B_n = 20$, $\lambda_s = 1.5$, $\mu_1 = \frac{1}{2}$, $\mu_2 = \frac{1}{5}$, $\mu_n = \frac{1}{14}$).

We can see that the effect of reserving capacity has a lesser impact on decreasing the probability of blocking stroke patients than the implementation of a "cut-off" policy using the same $B_r$. This is especially important if the neurology system is congested. Having one bed as the threshold for the "cut-off" policy reduces the probability of blocking stroke patients in 2.3% when $\lambda_n = 5$. However, we need 4 reserved beds in order to achieve the same impact if we use the other type of policy.

4.5 Evaluating severity-dependent admission policies

There is evidence that those who have the most severe strokes benefit the most (Jorgensen et al. 2000) from care in a stroke unit. In this section we analyze the impact of using an admission policy that assumes stroke patients are non-homogeneous in the sense that the expected benefits of getting access to the stroke unit differ by acuity level. We model the benefit of getting treatment at the stroke unit by an incremental survival probability. For simplicity, and given the lack of specific data, we assume this incremental survival probability is uniformly distributed from zero to $b_{max} = 0.5$. 
Figure 4.6: Reserving $B_r$ empty beds for stroke patients in neurology unit.

(a) Single LOS distribution in neurology unit

\[ B_s = 6 \quad B_n = 20 \quad \lambda_s = 1.5 \quad \lambda_n = 5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \mu_n \]

(b) Different LOS distribution in neurology unit

\[ B_s = 6 \quad B_n = 20 \quad \lambda_s = 1.5 \quad \lambda_n = 5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5} \quad \mu_n = \frac{1}{9} \]
Figure 4.7: Reserving capacity for stroke patients in neurology unit.

(a) Iterative approximation using “cut-off” priority system

\[ B_s = 6 \quad B_n = 20 \quad \lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5} \quad \mu_n = \frac{1}{14} \]

(b) Iterative approximation reserving beds

\[ B_s = 6 \quad B_n = 20 \quad \lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5} \quad \mu_n = \frac{1}{14} \]
The work of Shmueli et al. (2003) studies the impact of using a threshold based policy, i.e. admitting patients if there is available capacity in the unit and if their incremental benefit is greater than a specified value, for an ICU without any downstream blocking. In this section we study if their findings hold for the two stage chain of care, and the implications of having a shared second unit.

The expected revenue of the FCFS policy, $V_{FCFS}$ is given by the arrival rate to the unit, the probability of entering the unit and the expected benefit of treating a patient in the stroke unit.

$$V_{FCFS} = \lambda_s \cdot (1 - P(block))E(B)$$  \hspace{1cm} (4.4)

The benefit of the threshold policy $b$, $V_{FCFS-H(b)}$ is given by:

$$V_{FCFS-H(b)} = \lambda_s S_B(b) \cdot (1 - P(block))E(B/B > b)$$  \hspace{1cm} (4.5)

where $E(B)$ is the expected incremental probability of survival of a patient admitted to the stroke unit and $E(B/B > b)$ is the conditional expected incremental probability of survival of a patient, with an incremental survival probability greater than $b$, admitted to the stroke unit.

First, we study the effect of different relative capacities fixing the total number of beds. To isolate the effect of having a two-stage chain of care from sharing the capacity, for this analysis we do not consider non-stroke patients ($\lambda_n = 0$) and we compute the maximum $V_{FCFS-H(b)}$ using the two-dimensional model. The effect of shared capacity with non-stroke patients is described in section 4.5.2.

Table 4.2 (a) shows the results of implementing the threshold policy for a total number of beds equal to 9, $\lambda_s = 1.5$, $\mu_1 = \frac{1}{2}$ and $\mu_2 = \frac{1}{5}$, while Table 4.2 (b) for a total number of beds equal to 7.

From Table 4.2 (b) we can see that the benefit of using the threshold increases when the system has less capacity, with an increasing $b^*$. The expected annual savings increase by 10.8 statistical lives (representing an 11.3% increase from FCFS policy) when the total beds equals 9. When reducing the capacity to seven beds the benefit of using the threshold policy depends on the relative capacities ($B_s, B_n$). When $B_s$ is high the expected number of additional lives saved is 13.8 (representing an 17.6% increase from the FCFS policy), however this benefit decreases as the number of stroke beds decreases to an additional 10.2 expected annual lives saved (16.6% increase from FCFS policy).
Table 4.2: Evaluating benefits from admission policy.

(a) Total number of beds $B_s + B_n = 9$

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V^*_{FCFS-H}$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8, 1)</td>
<td>0.264</td>
<td>0.294</td>
<td>0.15</td>
</tr>
<tr>
<td>(7, 2)</td>
<td>0.264</td>
<td>0.294</td>
<td>0.155</td>
</tr>
<tr>
<td>(6, 3)</td>
<td>0.264</td>
<td>0.293</td>
<td>0.155</td>
</tr>
<tr>
<td>(5, 4)</td>
<td>0.262</td>
<td>0.292</td>
<td>0.16</td>
</tr>
<tr>
<td>(4, 5)</td>
<td>0.255</td>
<td>0.284</td>
<td>0.16</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5}$

(b) Total number of beds $B_s + B_n = 7$

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V^*_{FCFS-H}$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6, 1)</td>
<td>0.213</td>
<td>0.251</td>
<td>0.19</td>
</tr>
<tr>
<td>(5, 2)</td>
<td>0.213</td>
<td>0.251</td>
<td>0.19</td>
</tr>
<tr>
<td>(4, 3)</td>
<td>0.212</td>
<td>0.249</td>
<td>0.19</td>
</tr>
<tr>
<td>(3, 4)</td>
<td>0.202</td>
<td>0.237</td>
<td>0.19</td>
</tr>
<tr>
<td>(2, 5)</td>
<td>0.168</td>
<td>0.196</td>
<td>0.20</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5}$

4.5.1 Impact of overflowing stroke patients

The use of an admissions policy based on stroke severity generally results in admitting those patients who fall below the specified threshold directly to the neurology unit. However, this “overflow” of stroke patients induces extra congestion in the neurology unit and thus can reduce the benefit of using a threshold admission policy. To model the overflow of stroke patients we can use either the two-dimensional model or the iterative procedure, considering the diverted stroke patients as an external arrival rate of “non-stroke” patients to the neurology unit. As in the previous section we do not consider external non-stroke patients ($\lambda_n = 0$).

We compute the optimal threshold policy assuming stroke patients with an incremental benefit below
the target are overflowed to the neurology unit and admitted there if there is available capacity, or lost if the unit is full. Stroke patients with an incremental benefit above \( b^* \) are lost if the stroke unit is full. Use of FCFS policy is equivalent to setting \( b^* = 0 \) and then no stroke patients are overflowed and we recover the same expected benefit \( V_{FCFS} \).

Table 4.3 shows the expected benefits of using the optimal threshold policies for the same configurations as in Table 4.2. We can see that by allowing the overflow of patients, we reduce the benefit of using a threshold policy especially for configurations with a rather small stroke unit. In these cases the optimal threshold is smaller than when we do not consider the overflow. By decreasing the threshold we also decrease the number of patients who are candidate to be overflowed to the neurology unit, reducing the probability of blocking suitable stroke patients due to the unavailability of neurology beds. This dynamic explains why the optimal threshold is non-monotonic if a fraction of the stroke patients are overflowed to the neurology unit.

4.5.2 Impact of shared capacity

When we include having non-stroke patients in the second unit, the benefits of using a threshold policy for admitting stroke patients will depend as in the previous case on the level of congestion we have in the system. This case is very similar to overflowing the less severe stroke patients, but without having any control on the arrival rate to the neurology unit.

Table 4.4 shows the expected benefits of using a threshold policy when there is a shared neurology unit. The benefits of using a threshold policy will depend on the number of stroke beds as well as the additional beds at the neurology unit.

When there is ample capacity in both units (\( B_s = 8, B_n = 6 \)) by using the threshold policy we expect an additional 2.6 statistical lives saved, representing an 2.1% improvement with respect to the FCFS policy. The benefit increases up to 6.2 lives saved when the capacity of the stroke unit is reduced (\( B_s = 4, B_n = 10 \)), representing an 5.9% improvement with respect to the FCFS policy. In all cases, the optimal thresholds are lower than the ones computed in the previous sections, resulting in more patients being candidate for access to the stroke unit. With very limited capacity the external flow of non-stroke patients decreases the benefit of the FCFS policy as shown in Table 5 (b). However, by using a threshold policy we can improve by 11% representing up to 10.8 statistical lives saved in a year.
Table 4.3: Benefits from admission policy, overflowing patients.

(a) Total number of beds $B_s + B_n = 9$

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V^{*}_{FCFS-H}$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8, 1)</td>
<td>0.264</td>
<td>0.294</td>
<td>0.15</td>
</tr>
<tr>
<td>(7, 2)</td>
<td>0.264</td>
<td>0.293</td>
<td>0.15</td>
</tr>
<tr>
<td>(6, 3)</td>
<td>0.264</td>
<td>0.291</td>
<td>0.15</td>
</tr>
<tr>
<td>(5, 4)</td>
<td>0.262</td>
<td>0.286</td>
<td>0.145</td>
</tr>
<tr>
<td>(4, 5)</td>
<td>0.255</td>
<td>0.275</td>
<td>0.145</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5}$

(b) Total number of beds $B_s + B_n = 7$

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V^{*}_{FCFS-H}$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6, 1)</td>
<td>0.213</td>
<td>0.251</td>
<td>0.19</td>
</tr>
<tr>
<td>(5, 2)</td>
<td>0.213</td>
<td>0.249</td>
<td>0.185</td>
</tr>
<tr>
<td>(4, 3)</td>
<td>0.212</td>
<td>0.242</td>
<td>0.185</td>
</tr>
<tr>
<td>(3, 4)</td>
<td>0.202</td>
<td>0.224</td>
<td>0.185</td>
</tr>
<tr>
<td>(2, 5)</td>
<td>0.168</td>
<td>0.184</td>
<td>0.19</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5 \quad \mu_1 = \frac{1}{2} \quad \mu_2 = \frac{1}{5}$

4.6 Discussion

In this work we have presented a two-dimensional queuing model and an iterative procedure for the chain of care of stroke patients which explicitly includes the interaction between the two levels of care, including the blocking of the stroke unit if patients cannot be transferred to the second unit.

The proposed models can easily determine the minimum capacity hospitals should have at each level of the chain in order to achieve a specific service level. The optimizations show that the minimum number of stroke unit beds can be computed considering that unit in isolation only if there is ample capacity in the second unit. Since this is not usually the case, the blockage from the second unit and the cost of adding one extra stroke unit bed vs. adding one neurology unit bed needs to be considered and this
Table 4.4: Benefits from admission policy, shared capacity.

(a) Total number of beds $B_s + B_n = 14$.

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V_{FCFS-H}^*$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(8, 6)$</td>
<td>0.344</td>
<td>0.351</td>
<td>0.07</td>
</tr>
<tr>
<td>$(7, 7)$</td>
<td>0.339</td>
<td>0.347</td>
<td>0.075</td>
</tr>
<tr>
<td>$(6, 8)$</td>
<td>0.330</td>
<td>0.339</td>
<td>0.085</td>
</tr>
<tr>
<td>$(5, 9)$</td>
<td>0.313</td>
<td>0.326</td>
<td>0.09</td>
</tr>
<tr>
<td>$(4, 10)$</td>
<td>0.285</td>
<td>0.302</td>
<td>0.12</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5$  $\mu_1 = \frac{1}{2}$  $\mu_2 = \mu_n = \frac{1}{5}$  $\lambda_n = 1$

(b) Total number of beds $B_s + B_n = 9$

<table>
<thead>
<tr>
<th>$(B_s, B_n)$</th>
<th>$V_{FCFS}$</th>
<th>$V_{FCFS-H}^*$</th>
<th>$b^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(8, 1)$</td>
<td>0.264</td>
<td>0.294</td>
<td>0.15</td>
</tr>
<tr>
<td>$(7, 2)$</td>
<td>0.263</td>
<td>0.293</td>
<td>0.155</td>
</tr>
<tr>
<td>$(6, 3)$</td>
<td>0.261</td>
<td>0.290</td>
<td>0.155</td>
</tr>
<tr>
<td>$(5, 4)$</td>
<td>0.255</td>
<td>0.283</td>
<td>0.155</td>
</tr>
<tr>
<td>$(4, 5)$</td>
<td>0.241</td>
<td>0.268</td>
<td>0.165</td>
</tr>
</tbody>
</table>

$\lambda_s = 1.5$  $\mu_1 = \frac{1}{2}$  $\mu_2 = \mu_n = \frac{1}{5}$  $\lambda_n = 1$

The benefits of using an admission policy different from FCFS has also been studied. In concordance with previous studies we found that in the presence of really scarce resources, with probability of blocking arrivals of about 40%, using a threshold based admission policy can result in about 18% improvement vs. the FCFS policy. The more congested the stroke unit, the higher the optimal threshold for getting access to the stroke unit. Our methodology allows analyzing the impact of other commonly used policies, like overflowing less severe stroke patients directly to the neurology unit. We found that this practice decreases the benefits of using a threshold-based policy, especially in cases
where the stroke unit is small and there is congestion in the neurology unit. Also, we show that when sharing the second unit with non-stroke patients, using a threshold policy can improve the benefits of the stroke unit, especially when there is very little capacity overall, which is the current situation in many hospitals.

References

American Hospital Association. 2007 Survey of Hospital Leaders.


APPENDIX 1

Example $B = 3, N = 2$.

<table>
<thead>
<tr>
<th></th>
<th>00 10 11 20 21 22</th>
<th>30 31 32 33</th>
<th>40 41 42 43</th>
<th>50 51 52 53</th>
<th>60 61 62 63</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>$B_{00}$</td>
<td>$B_{01}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>$B_{10}$</td>
<td>$B_{11}$</td>
<td>$A_0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>$A_0$</td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where $B_{11} = A_1$, $B_{21} = A_2$. All matrices $A_i$ are of dimension $(B + 1) \times (B + 1)$.

$B_{00}$ is a matrix of dimension $\frac{B(B+1)}{2} \times \frac{B(B+1)}{2}$, $B_{01}$ is $\frac{B(B+1)}{2} \times (B+1)$ and $B_{10}$ is $(B+1) \times \frac{B(B+1)}{2}$.

$$A_0 = \begin{pmatrix} \lambda_b \\ \lambda_b \\ \lambda_b \end{pmatrix}$$

$$A_1 = \begin{pmatrix} -(\lambda_b + 3\lambda_n + 3\mu_b) & 3\lambda_n & -\lambda_b \\ \mu_n & -(\lambda_b + 2\lambda_n + \mu_n + 3\mu_b) & 2\lambda_n \\ 2\mu_n & -(\lambda_b + \lambda_n + 2\mu_n) & \lambda_n \end{pmatrix}$$
\[ A_2 = \begin{pmatrix} 0 & 3\mu_b & 0 & 0 \\ 0 & \mu_b & 2\mu_b & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \]

and

\[ A = A_1 + A_2 + A_3 = \begin{pmatrix} -3\lambda_n + 3\mu_b & 3\lambda_n + 3\mu_b \\ \mu_n & -2\lambda_n + \mu_n + 2\mu_b & 2\lambda_n + 2\mu_b \\ 2\mu_n & -(\lambda_n + 2\mu_n) & \lambda_n \end{pmatrix} \]

\[ B_{01} = \begin{pmatrix} 0 \\ 0 \\ \lambda_b \\ \lambda_b \\ \lambda_b \end{pmatrix} \]

\[ B_{10} = \begin{pmatrix} 0 & 0 & 3\mu_b & 0 & 0 \\ 0 & 0 & \mu_b & 2\mu_b & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \]

\[ B_{00} \] keeps the same structure than \( B_{10}, B_{11}, B_{01} \) and can be decomposed in blocks of increasing dimension.

\[ B_{00} = \begin{pmatrix} -\lambda_b & 0 & \lambda_b & 0 & 0 & 0 \\ \mu_b & -(\lambda_b + \mu + \lambda_n) & \lambda_n & 0 & \lambda_b & 0 \\ \mu_b & \mu_n & -(\lambda + \mu + \mu_n) & 0 & 0 & \lambda_b \\ 0 & 2\mu & 0 & -(\lambda_b + 2\mu_b + 2\lambda_n) & 2\lambda_n & 0 \\ 0 & \mu_b & \mu_b & \mu_n & -(\lambda_b + 2\mu_b + \mu_n + \lambda_n) & \lambda_n \\ 0 & 0 & 0 & 0 & 2\mu_n & -(\lambda_b + 2\mu_b) \end{pmatrix} \]

\[ B[R] = \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11} + RB_{21} \end{pmatrix} \]

Where \( R \) is the minimal solution to:

\[ R^2 A_2 + RA_1 + A_0 = 0 \]

With this we can check if the system is positive recurrent by computing the traffic intensity \( \gamma \).

\[ [x_0, x_1] \cdot B[R] = 0 \]

\[ x_0 e + x_1 (I - R)^{-1} e = 1 \]

\[ x_k = x_1 R^{k-1} \]
\[ \gamma = x \cdot A_0 \cdot e_1 - x \cdot A_2 \cdot e_1 \]

Where \( x \) solves \( x \cdot (A_0 + A_1 + A_2) = 0 \) and \( x \cdot e_1 = 1 \). The QBD process will be positive recurrent iff \( \gamma < 0 \), ensuring stability for the bed system.