UCLA
UCLA Electronic Theses and Dissertations

Title
Allocation of Resources to Elective Patients under Stochastic Emergency Patient Demand

Permalink
https://escholarship.org/uc/item/5cb701k4

Author
Blondal, Margret Asta

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California

Los Angeles

Allocation of Resources to Elective Patients under Stochastic Emergency Patient Demand

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Management

by

Margret Blondal

2012

© Copyright by
Margret Blondal
2012

Abstract of the Thesis

Allocation of Resources to Elective Patients under Stochastic Emergency Patient Demand

by

Margret Blondal

Master of Science in Management

University of California, Los Angeles, 2012

Professor Rakesh Sarin, Chair

A simulation model is developed for scheduling elective patients to a hospital facility based on resource availability. The approach takes into account two different types of patients, emergency in which admission occurs randomly, and elective in which demand is random, but patients are only admitted if they have been scheduled. The simulation aims to maximize the hospital's rewards from treating patients, in the presence of penalties for overbooking the facility. Few different heuristics for admission policies are compared and results analyzed. Simulation was performed on data from the Neurosurgery department from the UCLA Ronald Reagan Hospital. Even though the simulation is based on data from a neurosurgery department it can be used when admitting patients to other medical departments as well.

The thesis of Margret Blondal is approved.

Kumar Rajaram

Christiane Barz

Rakesh Sarin, Committee Chair

University of California, Los Angeles

2012

Table of Contents

1 Introduction **1**

2 Literature Review **5**

3 Simulation **7**

3.1 Patients' Arrivals and Resource Usage 8

3.2 Resource Usage and Availability 10

3.3 Hospital's Reward 11

3.4 Scheduling Elective Patients 12

3.5 Numerical Examples of Admission Process 18

3.6 Simulation Parameters 21

4 Patient Arrival Processes **23**

4.1 Estimating Arrivals of Emergency Patients 23

4.2 Censored Data 25

4.2.1 Kaplan-Meier estimator 26

4.2.2 EM algorithm 28

4.3 Estimating Arrivals of Elective Patients 30

5 Results **32**

5.1 Results from heuristic 1 - No Resources Reserved for Emergency Patients 32

5.2 Results from heuristic 2 - A proportion of Resources Reserved for Emergency Patients 33

5.3 Results from heuristic 3 - Reservations for Emergencies and Contribution Constraints on Electives 34
5.4 Comparing Results 36

6 Data Analysis 42
6.1 Hospital's Resources 42
6.2 Patient Arrival Processes 43
6.3 Categorization of Patients 44
6.4 Patient Grouping 46
6.4.1 Priority of Patients 47
6.4.2 Admission Diagnosis 48
6.4.3 Contribution from patients 51
6.4.4 Patient Types Defined 53
6.5 Constructing Resource Matrices 53
6.6 Probabilities of Resource Matrices 55
6.7 Expected Total Resource Usage and Contribution of Patient Types 55
6.8 Demand by patient types 56
6.9 Maximum demand a day per patient type 56
6.10 Penalties for Overbooking Resources 56
6.11 Identification of Censored Data 58

7 Remarks 60

List of Figures

3.1 *Flow chart for patient admissions in simulation. The items within the shaded block have different admission policies for each heuristic.* 17

5.1 *Confidence intervals for the hospital's reward from each heuristic.* 36

5.2 *The reward and penalty are shown out of the total contribution for each heuristic.* 37

5.3 *The reward and penalty are shown in proportion to the total contribution for each heuristic.* 38

5.4 *The proportion of elective patients admitted to the hospital out of all that requested admission.* 39

5.5 *Shows the mean number of days that each resource was overbooked over a period of one year.* 40

5.6 *Shows the proportion of days that each resource was overbooked out of the total count of days where resources were overbooked.* 41

List of Tables

5.1 *Results from simulating heuristic 1, where there are no reservations of resources for emergency patients.* 33

5.2 *Results from simulating heuristic 2, where the reserved proportion of resources for emergency patients is 20%.* 34

5.3 *Results from simulating heuristic 3, version A, where* $\gamma_1 = 2, \gamma_2 = 1$*, and* $\gamma_3 = 84$*. The shadow prices are* $V_1 = 183.878$*,* $V_2 = 56.5364$*, and* $V_3 = 0$. 35

5.4 *Results from simulating heuristic 3, version B, where* $\gamma_1 = 12, \gamma_2 = 9$*, and* $\gamma_3 = 84$*. The shadow prices are* $V_1 = V_2 = V_3 = 0$. 35

6.1 *p-values from comparing emergency arrival distribution between weekdays.* 45

6.2 *p-values from comparing elective arrival distribution between weekdays. There are no elective arrivals during Saturdays and Sundays and therefore those days are not included.* 45

6.3 *p-values from comparing underlying distributions of resource usage for emergency patients and elective patients using K-S test at significance* $\alpha = 0.01$. 48

6.4 *Results from K-S test where the null hypothesis is that the usage of each resource is normally distributed* 49

6.5 *p-values from comparing underlying distributions of resource usage for different admission diagnoses.* 50

6.6 *p-values from comparing underlying distributions of resource usage for admission diagnoses 996A and 996C.* 51

6.7 *Length of stay, ICU usage and OR slots used by patients from all patient entries for diagnoses X2.a, X2.b, and X2.c.* 52

6.8 *Penalties for all resources used in simulation for all heuristics.* . . 58

CHAPTER 1

Introduction

Hospitals often have limited availability for patients needing treatments, operations, and care. Therefore one of the challenges when scheduling patients is to better manage the utilization of hospitals resources, without additional operational cost. The problem considered in this paper is how to allocate resources to patients under random admission requests with the objective of maximizing revenue.

Some medical facilities do not always have the option of diverting patients to other hospitals or denying them treatment. These customers can be patients that need an emergency procedure, or patients that require a specialized procedure or treatment that is not offered in other hospitals available to the patient. This fact makes it hard for the hospital to manage its resource availability. In most cases there is no possibility to know in advance what the demand will be.

The main difference in how quickly different patient classes are treated in the healthcare system is based on patients' priority status. In the work done here there are two priority groups, emergency patients which must be admitted instantly, and elective patients that can be, if necessary, admitted at a later date. Since all emergency patients must be admitted at time of arrival, but the same does not hold for elective patients, we look at the arrival process of elective patients and emergency patients separately.

New admission requests for elective patients arrive on weekdays, and the inter-arrival time between requests is assumed to be exponentially distributed. Elective

patients are never accepted if the hospital does not have enough resources available at the time of admission. The hospital can fill up their resource usage quickly, and thus, unable to admit all elective patients that are requesting admission. That is, elective patient demand is greater than the number of patients that the hospital can admit. In general, medical facilities record all admissions to the hospital, but do not record the number of cases they reject admission requests from elective patients. The result from this is that the elective demand data is *censored*. In this case, where demand is, at least equal to, if not greater than the number of admitted patients, the data is said to be *right censored*. If the distribution of the elective demand would be estimated without using methods to take into account the censoring, the demand would be greatly underestimated. On the other hand, since all emergency patient must be admitted, the emergency demand is *uncensored*.

Emergency demand follows a stochastic distribution. The facility will always have to meet this demand immediately, i.e. the facility will have to provide a bed space for the patient even if all beds are occupied, as well as have a surgical operation when needed. It will result in additional cost for the hospital if the it has to admit an emergency patient when all or some of the resources are fully booked. The additional cost can occur because of measures taken to find a suitable bed for the patient, medical specialists working overtime, or even increased number of staff taking care of the patients, to name a few.

If the hospital does not have enough resources for emergency patients, it must use alternatives to accommodate the patients. This results in penalties and in most cases negative contribution from a patient to the hospital. Therefore, penalties are applied in cases when the hospital overbooks any of its resources.

The reward function consists of the contribution received from patients and the cost of overbooking resources. The contribution from patients is the revenue from treating them, minus the direct cost of the treatment.

The data used in this analysis was provided by the Neurosurgery department at the Ronald Reagan UCLA Medical Center. They have three main resources to distribute to patients, regular beds, ICU beds, and OR slots. There were a total of 781 patient admission entries, of which 372 were for elective patients and 409 emergency patients, for the first half of year 2011. There were various fields given for each patient admission. It included the date of admission, total length of stay in hospital (in days), number of days in ICU, date of surgery, and total OR slots. It also has the characteristics of each patient's condition, e.g. priority type, admission diagnosis, final diagnosis, surgery type, the DRG (diagnosis related group) classification and insurance coverage. These characteristics are important when it comes to figuring out how to categorize patients for admission purposes. Revenue and cost figures from treating patients were also included in the data. It should be mentioned that, due to confidentiality, those figures have been scaled so that they are unrecognizable, but still in correct proportions to each other.

In the case simulations are performed for there are two different types of beds patients can be admitted to, either the ICU beds, which are for patients that need a great deal of care, e.g. after surgery, or the regular beds, which are for patients that do not require as much care, e.g. patients in their last stages of inpatient recovery, or patients that have had minor surgeries. Since there is a great difference in the two neurosurgery bed units, it is important to distinguish between them in the simulation. There is a possibility that all ICU beds are taken when a patient arrives that needs a bed that requires more equipment than a regular bed. In cases like these the hospital has to make some arrangements to give the patient the care needed for his/her treatment.

The remainder of the paper is organized as follows; in section no. 2 I discuss previous literature related to the problem of scheduling patients under stochastic demand. Section no. 3 talks about the structure of the simulation model,

and discusses the different heuristics compared. The patient arrival process is discussed in detail in section no. 4 and methods for dealing with censored data are introduced. Also, it specifies the simulation parameters. The results from the simulation are in section no. 5. Section no. 6. has a detailed description of the data analysis, and shows how we reached some of the values used when simulating. A discussion about the work is in section no. 7.

CHAPTER 2

Literature Review

Some of the earliest work considering patient arrival processes were carried out by Bailey (1952) and Balintfy (1960). Bailey studied a facility where all patients were scheduled and the arrival process was deterministic, but the service time was assumed to be gamma-distributed. Balintfy modeled the inpatient arrival process to have a negative binomial distribution, as a result from a risk function measuring the risk of hospitalization. Singer (1962) proposes a Poisson distribution for patient arrival process, as do Fetter and Thompson (1966) and Gabrielson (1962). Swartsman (1970) shows that arrival rates of patients can be represented by a Poisson process, with time-varying arrival rates. Litvak (2007) uses Poisson distribution for emergency patient arrivals to an ICU unit. Here, it is assumed that the arrivals of both emergency and elective patients follow a Poisson distribution.

Related to this paper is work done by Gerchak et al (1996). They provide a dynamic program for scheduling surgeries of elective patients, when emergency demand is uncertain. This is furthered here, by taking into account bed scheduling as well, although the analysis uses simulation rather than dynamic programming.

Demonstration of some of the earlier work when it comes to patient admissions are Kolesar (1970) and Collart and Haurie (1976). Kolesar introduced a Markovian decision model for scheduling admissions at a hospital under the uncertainty of emergency demand. However, he did not solve it numerically due to the computing limitations at that time. Collart and Haurie provide a semi-Markov chain to search for an admission policy for outpatients and construct a simulation model

of hospital admission.

The work that has been done for resource capacity planning is mostly carried out by simulation or queuing models. Harper and Shahani (2002) proposed a simulation model for planning and managing hospital beds. They used three patient priority classes, emergency, elective, and inpatients; they also grouped them by departments within the hospital. Ridge, Jones, Nielsen and Shahani (1998) created a simulation model for intensive care units (ICU). Like in this paper they focused on two different patient classes, elective and emergency patients. Instead of looking at the most profitable way of utilizing the hospital's beds, they look at how many ICU beds to offer. They also give a demonstration of using queuing theory models when it comes to managing bed capacity in hospitals. Gorunescu, McClean and Millard (2002) present a queuing model and a way to optimize the use of hospital resources, particularly bed occupancy.

One of the first to estimate censored data was Tobin. Kaplan and Meier - nonparametric estimators for right censored observations, that can possibly be left truncated. McGill (1995) looks into censorship of demand in the airline industry for two or more classes of demand. Huh et al. (2011) apply the KM-estimator for inventory management, and Ryzin et al. (2000) in revenue management, but apart from that there have not been many applications of the KM-estimator in operations management, and none to estimate patient arrivals that I know of.

CHAPTER 3

Simulation

There were a total of three heuristics carried out for the decision process of admitting patients. Each one represents a different admission policy for elective patients. Every emergency patient that arrives to the hospital must be admitted in all heuristics, no matter the availability of resources at the time of the emergency patient's arrival. A short description of each heuristic follows:

Heuristic 1: The heuristic is intended to see what happens if there is no reservation of resources for emergency patients. As long as there are enough resources available for an elective patients that arrives, the patient is admitted.

Heuristic 2: Here, a specific percentage of all resources is reserved for emergency patients, the same percentage for every resource. Elective patients are admitted to the hospital as long as the hospital does not have to use any of the resources that are reserved for emergency patients.

Heuristic 3: The heuristic is developed using a method introduced in by Barz and Rajaram (2012), Elective Admissions under Multiple Resource Constraints. They formulate the patient admission control process as a Markov decision process, and use approximate dynamic programming to find an amount of resources that is reserved for emergency patients only, as well as the shadow price for each resource. The shadow price is used in combination with the resource usage of elective patients to determine if patients should be considered for

admission.

All heuristics use the assumption that if an elective patient cannot be admitted on the day requested the hospital can delay the admission for some time, depending on the condition of the patient. It is assumed that if more than one admission date is feasible, the hospital admits patients at the earliest date, and that patients always accept the admission date suggested by the hospital. Also, it is assumed that there are no patient cancellations or "no shows". It could be incorporated into the model fairly easily, but won't be done here because the data provided for the simulation did not include any information regarding cancellations or patients not showing up. If the hospital does not admit the patient, the demand is lost, that is, the patient does not appear in another day's demand.

Before going into more details about the heuristics for elective patient admission the simulation model itself is described.

3.1 Patients' Arrivals and Resource Usage

When a patient arrives to the hospital he/she is categorized by priority and admission diagnosis. Based on those characteristics, the patient is assigned a patient type $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of all possible patient types. When a patient shows up at the hospital, he/she belongs to a type i with probability of p_i. The probability is based on the hospital's previous history. It is described in detail how the patient types are defined and probabilities are found in chapter 6. The maximum number of days a patient of type i has to be treated for is $N_i \in \mathbb{N}_+$, but it is possible that the patient stays for less time than N_i days.

The hospital has a total of R resources to distribute to patients. Each resource is denoted as $r \in \{1, \ldots, R\}$, where $r \in \mathbb{N}_+$. Every patient that arrives to the hospital has a specific resource usage, which is defined as a random $R \times N_i$ matrix U_i. If we let a patient type i have a total number of $J_i \in \mathbb{N}_+$ possible resource

matrices, a resource matrix number j for patient type i is defined as

$$u^{j|i} = \begin{bmatrix} u^j_{1,1} & u^j_{1,2} & \dots & u^j_{1,N_i} \\ u^j_{r,1} & u^j_{r,2} & \dots & u^j_{r,N_i} \\ \vdots & \vdots & \ddots & \vdots \\ u^j_{R,1} & u^j_{R,2} & \dots & u^j_{R,N_i} \end{bmatrix}, \qquad u^j_{r,n} \in \mathbb{N}_0, \tag{3.1}$$

where $j \in \{1, \dots, J_i\}$, and $n \in \{1, \dots\ N_i\}$ stands for the number of the day of a patients stay. Then, each element of the matrix, $u^j_{r,n}$, is the amount of resources r the patient needs at time t. Let the probability that a patient of type i has resource matrix $u^{j|i}$ be $\Pr(U_i = u^{j|i}) = p_{j|i}$ then we can write the expected resource usage of patient type i at the time of admission as

$$\mathbb{E}[U_i] = \sum_{j \in \mathcal{J}_i} p_{j|i} u^{j|i}, \tag{3.2}$$

where $\mathbb{E}[U_i]$ will be a matrix of size $R \times N_i$.

The number of patients that arrive each day is assumed to be Poisson distributed. Then the arrivals in one time period do not affect the number of arrivals in another period, that is, they are independent. Also, we assume the number of emergency patients that arrive is independent of the emergency demand. Because the patient demand is assumed to be Poisson, the time between patient arrivals is exponentially distributed. We let the time between arrivals of elective patients be $\vec{Y}$ and time between arrivals of emergency patients be $\vec{X}$, where $\vec{X}$ and $\vec{Y}$ are exponentially distributed random variables.

Let k denote the number of the arriving patient, $k \in \mathbb{N}_+$. The arrival time of the kth patient is $\tau_k \in \mathbb{R}_+$. Then the patients arrive in order

$$\tau_1 < \tau_2 < \dots < \tau_k < \dots < \tau_\infty. \tag{3.3}$$

The time horizon of the simulation is $t = 0, \dots, T$ days, where $t = 0$ is the first day of simulation data collection and T is the last day of the simulation period.

Patient k arrives on day $t_k = \lfloor \tau_k \rfloor$, where $\lfloor * \rfloor$ maps the real number $*$ to the highest previous integer.

At the time of a patient's arrival, the hospital has to decide whether to admit the patient or not. If the hospital cannot admit a patient of type i on the day of arrival it can delay admission by at most M_i days. The delay is m_k days, where $m_k \in \{0, \dots \ M_{i|k}\}$, and $m_k = 0$ means that the patient is admitted on the day of arrival. Each patient k is assigned a binary variable $a_k \in \{0, 1\}$ that takes the value 1 if patient k is admitted on day $t_k + m_k$, and 0 otherwise.

Let $\vec{z_r}^k = (z^k_{r,0}, z^k_{r,1}, \dots, z^k_{r,T})$ be the vector for resource usage of resource r for patient k. The vector stores the number of resources the patient needs on each day over the whole simulation period. Using previously defined resource matrix $u^{j|i}(k)$ for patient k, from equation (3.1), patient k has a resource usage of

$$z^k_{r,(t_k+m_k+n-1)} = a_{k,m} \left[u^{j|i}(k)\right]_{r,n}, \qquad \forall n \in \{1, \dots, N_i\}, \tag{3.4}$$

of resource r at time $t_k + m_k + n - 1$, where n is the number of the day of patients stay, and $[*]_{r,n}$ stands for the element in row r and column n of matrix $*$. In the equation, a_k becomes 1 if patient k is admitted m days after arrival, and takes the value 0 otherwise. If there is no resource usage of r defined for patient k at time t then $z^k_{r,t}$ takes the value of 0.

3.2 Resource Usage and Availability

The hospital's resource usage per day is found by summing over the amount of resources all admitted patients need per day (from equation (3.4)), or by

$$\sum_{\forall k} z^k_{r,t}, \qquad t \in \{0, \dots, T\}. \tag{3.5}$$

Maximum capacity of resource r per day is $c_r \in \mathbb{N}_0, \forall r \in \{1, \dots, R\}$. Combining this with equation (3.5), the hospitals resource availability of resource r per

day t is

$$\left[c_r - \sum_{\forall k} z_{r,t}^k \right]^+, \tag{3.6}$$

where $[\cdot]^+$ stands for $\max\{0, \cdot\}$. From this it follows that the hospital has overbooked resource r by

$$\left[\sum_{\forall k} z_{r,t}^k - c_r \right]^+, \tag{3.7}$$

at time t.

3.3 Hospital's Reward

In the simulations, the reward function is defined as the contribution the hospital receives from elective patients less the penalties for overbooking the hospital's resources,

$$\begin{aligned} \text{Reward} = &\text{ Contribution from Elective Patients} \\ &- \text{Penalties for Overbooking Resources.} \end{aligned} \tag{3.8}$$

Every patient that is admitted for treatment brings some contribution to the hospital. The contribution from a patient is defined as the net revenue from the patient, less the direct cost that comes from treating him/her. It is directly related to the admission diagnosis of the patient and the resources the patient uses. Thus, contribution from a patient of type i with resource matrix $u^{j|i}$ can be written as $g_{j|i}$, and patient k brings contribution of $g_{j|i}(k)$ to the hospital. The total contribution that the hospital receives from treating elective patients is

$$\text{Contribution from Elective Patients} \quad = \quad \sum_{\forall k} a_k g_{j|i}(k) \mathbb{1}_{EL|k}, \tag{3.9}$$

where $\mathbb{1}_{EL|k} = \begin{cases} 1 & \text{if patient } k \text{ is an elective patient} \\ 0 & \text{otherwise} \end{cases}$.

It is important to note that the reward function does not include the contribution the hospital gains from treating emergency patients. Since the hospital must always accept all emergency patients, their contribution does not need to be included when exploring the different admission policies for elective patients.

Penalty for overbooking resource r is π_r for each unit used beyond the capacity of the hospital per day. Using the penalty and formula (3.7), the overbooking cost of resource r at time t becomes

$$\pi_r \left[\sum_{\forall k} z_{r,t}^k - c_r \right]^+ .$$

Summing this over all periods t and all resources r, we get

$$\text{Penalties for Overbooking Resources} \quad = \quad \sum_{t=1}^{T} \sum_{r=1}^{R} \pi_r \left[\sum_{\forall k} z_{r,t}^k - c_r \right]^+, \quad (3.10)$$

where equation (3.10) gives the hospital's total overbooking cost over the whole simulation period. Inserting the results for the contribution and the pentalties into equation (3.8), the total reward of the hospital over the period carried out in the simulation is written as

$$\text{Reward} \quad = \quad \sum_{\forall k} a_k g_{j|i}(k) \mathbb{1}_{EL|k} - \sum_{t=1}^{T} \sum_{r=1}^{R} \pi_r \left[\sum_{\forall k} z_{r,t}^k - c_r \right]^+ . \quad (3.11)$$

3.4 Scheduling Elective Patients

When an elective patient arrives to the hospital, the hospital only has information about the patient's priority and admission diagnosis. Based on this information, the patient falls under a specific patient type i. The hospital does not know the exact resource usage the patient will need for the treatment, but it can find an estimation of resource usage by using information from prior patients that belong to the same type, as equation (3.2) in chapter 3.1 shows.

To decide if a patient is to be admitted, and on which day the admission should take place, the expected resource requirement of the patient is compared to the

resource availability of the hospital for the days the patient needs to be treated. First, it is checked if it is feasible to admit the patient on the day of request. If not, it is checked whether the patient can be admitted on any other day within the limits of maximum delay of admission. If more than one admission date is feasible, the patient is admitted on the earliest date.

Each heuristic carried out in the simulation has a different admission policy when it comes to admit elective patients. It will now be described how the decision process in the simulation works for all three heuristics. It is important to remember that, in all heuristics, the hospital must accept all emergency patients that arrive to the hospital.

Heuristic 1. In the first simulated heuristic, elective patients are admitted until the hospital has reached full capacity of its resources. That means that there are no resources reserved for emergency patients. An elective patient that arrives at the hospital is admitted on the day of arrival if the resource usage of the hospital, plus the resource usage of the patient, does not go over the resource capacity of the hospital. When a patient arrives at time $\tau_{k'}$, it can be considered for admission if the following constraint is feasible

$$c_r - \left[\sum_{k=1}^{k'-1}(z_{r,t}^k) + z_{r,t}^{k'}\right] \geq 0, \qquad r \in \{1,\ldots,R\}, t_{k'} \leq t \leq T, \tag{3.12}$$

where $t_{k'} = \lfloor \tau_{k'} \rfloor$ is the day that request from patient k' arrives. However, at this time the exact resource usage of the patient is unknown, the resource usage is the random matrix U^i, and thus, the expected value of the patient resource usage must be used when deciding whether to admit the patient or not. Also, the exact resource usage of all patients k that might stay on days after the day of patient's k' request arrival, is not known. Thus we have to take the expected value of resource usage for those patients as well. Taking the expected value of

the left side of equation (3.12),

$$\mathbb{E}\left[c_r - \left(\sum_{k=1}^{k'-1}(Z_{r,t}^k) + Z_{r,t}^{k'}\right)\right]$$

the left side becomes

$$c_r - \left(\sum_{k=1}^{k'-1}\mathbb{E}\left[Z_{r,t}^k\right] + \mathbb{E}\left[Z_{r,t}^{k'}\right]\right).$$

Using this, the result is that the following constraints must hold:

$$c_r - \left(\sum_{k=1}^{k'-1} z_{r,t}^k + \mathbb{E}\left[Z_{r,t}^{k'}\right]\right) \geq 0, \qquad \forall r \in \{1,\ldots,R\}, t = t', \quad (3.13)$$

$$c_r - \left(\sum_{k=1}^{k'-1}\mathbb{E}\left[Z_{r,t}^k\right] + \mathbb{E}\left[Z_{r,t}^{k'}\right]\right) \geq 0, \qquad \forall r \in \{1,\ldots,R\}, t_{k'} < t \leq T. \quad (3.14)$$

If the hospital cannot admit the patient at the time the patient's request arrives, it has the option to delay the admission for at most $M_{i|k}$ days. Since the assumption is that the hospital admits elective patients at the earliest date possible and the patients always accept the time suggested by the hospital, the admission policy for each patient k can be described as finding the lowest m_k such that the following constraints hold:

$$c_r - \left(\sum_{k=1}^{k'-1} z_{r,t'+m_k+n-1}^k + \mathbb{E}\left[U_{r,n}^{i'}(k)\right]\right) \geq 0, \qquad \forall r \in \{1,\ldots,R\}, m_k = 0, n = 1, \quad (3.15)$$

$$c_r - \left(\sum_{k=1}^{k'-1}\mathbb{E}[Z_{r,t'+m_k+n-1}^k] + \mathbb{E}\left[U_{r,n}^{i'}(k)\right]\right) \geq 0, \qquad \forall r \in \{1,\ldots,R\}, m_k \geq 0, n \geq 1. \quad (3.16)$$

If the constraints hold for at least one $m_k \leq M_{i|k}$, the patient is admitted, and the variable $a_{k',m}$ is assigned the value 1 for the lowest $m = m_k$. If not, $a_{k',m} = 0, \forall m \leq M_{i|k}$.

Heuristic 2. This heuristic is similar to heuristic 1, in the way that all elective patients are admitted until a specific limit is reached. However, in this case,

the hospital reserves a percentage $(1-q)$ of all resources for emergency patients and only admits elective patients if there are enough resources available, out of $c_r q, \forall r$, available for the patient. Therefore, elective patients are accepted if the total expected usage of each resource is lower than q percent of the hospital's resource capacity after the patient has been accepted. Using the same approach as for heuristic 1, changing only the resource capacity, the arriving patient is admitted, and $a_{k,m}$ gets the value 1, if the following resources are feasible

$$\lfloor c_r q \rfloor - \left(\sum_{k=1}^{k'-1} z^k_{r,t'+m_k+n-1} + \mathbb{E}\left[U^{i'}_{r,n}(k) \right] \right) \geq 0, \qquad \forall r \in \{1,\ldots,R\}, m_k = 0, n = 1, \tag{3.17}$$

$$\lfloor c_r q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{r,t'+m_k+n-1}] + \mathbb{E}\left[U^{i'}_{r,n}(k) \right] \right) \geq 0, \quad \forall r \in \{1,\ldots,R\}, m_k \geq 0, n \geq 1. \tag{3.18}$$

Heuristic 3. This heuristic differs from the previous two by being price-directed. It comes from a working paper by Barz and Rajaram (2012), Elective Admissions under Multiple Resource Constraints. They model the admission process as a Markov decision process. Using approximate dynamic programming, they suggest how much should be reserved of each resource, γ_r. They also solve for the shadow prices, V_r, of each resource, and use it in combination with resource usage and contribution of each patient to see if the patient should be considered for admission. They define

$$\text{Capacity Adjusted Contribution} = \text{ Contribution} - \text{Price of Using Resources}. \tag{3.19}$$

Using the patient's type at admission, the hospital finds the expected contribution $\mathbb{E}[G_{i'}]$ of the patient. The price of using each resource for the patient is the shadow price of the resource, times the number of units of the resource the

patient needs.

The Capacity Adjusted Contribution must be greater than 0 for the hospital to consider the patient for admission. Since we don't know the exact resource usage of the patient or the contribution, the expected value must be used. Then equation (3.19) becomes

$$\sum_{j=1}^{J} \mathbb{E}[G_{i'}] - \sum_{r=1}^{R} V_r \sum_{n=1}^{N_i} \mathbb{E}[U_{r,n}^{i'}] \geq 0. \tag{3.20}$$

If a patient's expected contribution is high enough we move on to see if the hospital has enough availability of all resources, reserving a total of γ_r of each resource r for emergency patients. The constraints that must hold for the patient to be admitted are the following:

$$\begin{aligned} c_r - \gamma_r - \left(\sum_{k=1}^{k'-1} z_{r,t'+m_k+n-1}^{k} + \mathbb{E}\left[U_{r,n}^{i'}(k) \right] \right) \geq 0, \\ \forall r \in \{1, \ldots, R\}, m_k = 0, n = 1, \end{aligned} \tag{3.21}$$

$$\begin{aligned} c_r - \gamma_r - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{r,t'+m_k+n-1}^{k}] + \mathbb{E}\left[U_{r,n}^{i'}(k) \right] \right) \geq 0, \\ \forall r \in \{1, \ldots, R\}, m_k \geq 0, n \geq 1. \end{aligned} \tag{3.22}$$

The simulation here differs in two ways from the method in the paper by Barz and Rajaram. First, in their work elective patients arrive at the beginning of the day, and emergency arrivals only occur after all elective patients have arrived that day. The second difference is that here the expected resource usage is calculated at the time of the patients arrival and not updated after that. Barz and Rajaram use the patient's state at a given time to find the expected value of the future resource usage. Therefore, the expected value of the resource usage of a patient is updated every time that patient moves into another state.

The simulation process flow is shown in figure 3.1. It should be noted that since there is no contribution requirement for patients in heuristic 1 and 2, the answer to the node "Contribution Meets Requirements" will always be *yes*. Also, all three heuristics have different feasibility constraints, as described in the sections above.

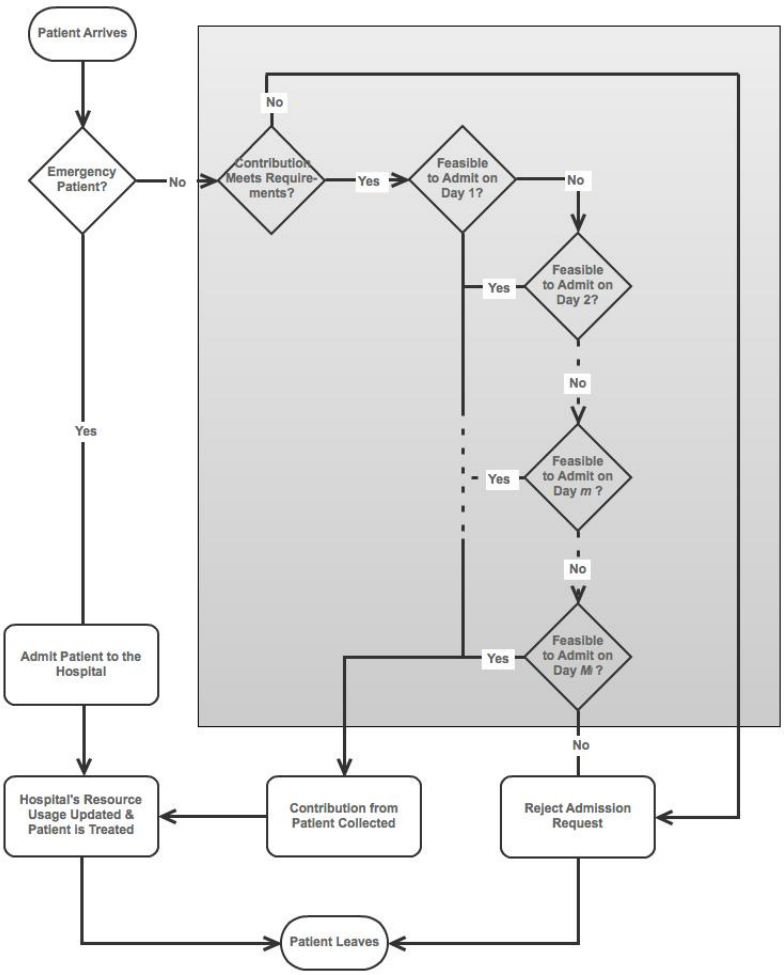

Figure 3.1: *Flow chart for patient admissions in simulation. The items within the shaded block have different admission policies for each heuristic.*

3.5 Numerical Examples of Admission Process

Suppose we have a hospital with $R = 2$ resources, where $c_1 = 8$ is the total bed capacity, and $c_2 = 18$ is the number of OR slots the hospital can assign for surgeries on each day. If $k' - 1$ number of patients have already arrived at the hospital at time t', the number of units (out of hospital's total resources) for resource r that has already been assigned to patients is written as

$$\min\left\{0, \sum_{k=1}^{k'-1} z_{r,t}^{k}\right\},$$

and let it take the value of 6 beds for $r = 1$ and 2 OR slots for $r = 2$ on day $t = t'$. For $t \in \{t' + 1, t' + 2, t' + 3, t' + 4\}$, let the expected values of the corresponding resource usage per day be $\{5, 4, 4, 2\}$ for $r = 1$, and {14,6,13,0} for $r = 2$.

Now suppose the hospital receives a scheduling request for patient k' at time $\tau_{k'}$, where $\lfloor \tau_{k'} \rfloor = \lfloor \tau_{k'-1} \rfloor = t'$, and that the request is for an elective patient. The patient belongs to a type i', with maximum length of stay $N_{i'} = 3$ days and maximum delay $M_{i'} = 2$ days. There are $J_{i'} = 2$ resource matrices $u^{j|i'}$ possible for the patient, with

$$u^{1|i'} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 0 & 0 \end{bmatrix}, \qquad u^{2|i'} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & 0 \end{bmatrix},$$

and corresponding contributions $g_{1|i'} = 300$, and $g_{2|i'} = 270$. Using equation (3.2), the expected resource matrix for a patient that belongs to type i' is

$$\mathbb{E}\left[U^{i'}\right] = \begin{bmatrix} 1 & 1 & 0.5 \\ 1 & 2 & 0 \end{bmatrix}.$$

A decision has to be made, whether to admit the patient or not. In the following paragraphs it is shown numerically how the decision process works for all heuristics.

Heuristic 1. We start by looking at the day of request arrival, where $m_{k'} = 0$. The constraints that must hold here for the patient to be admitted at the day of arrival, t', are

i) For $r = 1$

First day of stay: $c_1 - \left(\sum_{k=1}^{k'-1} z_{1,t'}^k + \mathbb{E}[U_{1,1}^{i'}]\right) = 8 - (6+1) \geq 0.$

Second day of stay: $c_1 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+1}^k] + \mathbb{E}[U_{1,2}^{i'}]\right) = 8 - (5+1) \geq 0.$

Third day of stay: $c_1 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+2}^k] + \mathbb{E}[U_{1,3}^{i'}]\right) = 8 - (4+0.5) \geq 0.$

ii) For $r = 2$

Fist day of stay: $c_2 - \left(\sum_{k=1}^{k'-1} z_{2,t'}^k + \mathbb{E}[U_{1,1}^{i'}]\right) = 18 - (2+1) \geq 0.$

Second day of stay: $c_2 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{2,t'+1}^k] + \mathbb{E}[U_{2,2}^{i'}]\right) = 18 - (14+2) \geq 0.$

Third day of stay: $c_2 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{2,t'+2}^k] + \mathbb{E}[U_{2,3}^{i'}]\right) = 18 - (6+0) \geq 0.$

All constraints hold for both resource types. Therefore, if the hospital would use heuristic 1 as admission policy, the patient would be admitted.

Heuristic 2. Let the proportion of total resources reserved for emergency patients be $(1-q) = 0.2$. That means that there are $\lfloor c_1 q \rfloor = \lfloor (8)(0.8) \rfloor = 6$ beds available for elective patient admission, and $\lfloor c_2 q \rfloor = \lfloor (18)(0.8) \rfloor = 14$ OR slots. Again, we start by looking at the day of request arrival, where $m_{k'} = 0$. The first constraint, for $r = 1$ and $t = t'$,

$$\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} z_{1,t'}^k + \mathbb{E}[U_{1,1}^{i'}]\right) = 6 - (6+1) \ngeq 0,$$

does not hold. There is no need to check the other constraints for time t' and thus we move on to see if it is possible to admit the patient at time $t = t' + 1$.

i) For $r = 1$

For first day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+1}^k] + \mathbb{E}[U_{1,1}^{i'}]\right) = 6 - (4+1) \geq 0.$

For second day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+2}^k] + \mathbb{E}[U_{1,2}^{i'}]\right) = 6 - (4+1) \geq 0.$

For third day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+3}^k] + \mathbb{E}[U_{1,3}^{i'}]\right) = 6 - (2+0.5) \geq 0.$

ii) For $r = 2$

First day of stay: $\lfloor c_2 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{2,t'+1}] + \mathbb{E}[U^{i'}_{1,1}] \right) = 14 - (14 + 1) \not\geq 0.$

Since we have reached a constraint that is broken we stop here, the patient cannot be admitted at time t = t'+1. But $M_{i'} = 2$, and we can try the last possible admission date, $t' + 2$.

i) For $r = 1$

First day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{1,t'+2}] + \mathbb{E}[U^{i'}_{1,1}] \right) = 6 - (5 + 1) \geq 0.$

Second day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{1,t'+3}] + \mathbb{E}[U^{i'}_{1,2}] \right) = 6 - (4 + 1) \geq 0.$

Third day of stay: $\lfloor c_1 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{1,t'+4}] + \mathbb{E}[U^{i'}_{1,3}] \right) = 6 - (4 + 0.5) \geq 0.$

ii) For $r = 2$

First day of stay: $\lfloor c_2 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{2,t'+2}] + \mathbb{E}[U^{i'}_{1,1}] \right) = 14 - (6 + 1) \geq 0.$

Second day of stay: $\lfloor c_2 q \rfloor - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z^k_{2,t'+3}] + \mathbb{E}[U^{i'}_{1,2}] \right) = 14 - (13 + 2) \not\geq 0.$

Again we reach a constraint that is broken. This is the last possible admission date, and thus, the patient is referred to another facility/declined admission.

Heuristic 3. For this heuristic, let $\gamma_1 = 3$ and $\gamma_2 = 2$, that is, the hospital reserves 3 beds and 2 slots of OR slots for emergency patients. Also, let the shadow prices are $V_1 = 50$ and $V_2 = 40$.

First, we check if the patient meets the capacity adjusted contribution restriction. If the patient meets the requirement, we move on to the next step, but if the patient does not meet the requirement he/she is denied admission and there is no need to go through the next steps. Also, let the contribution for patient of type i' be $\Pr(g_{1|i'} = 200) = \Pr(g_{2|i'} = 70) = 0.5$. Now, we can calculate the Capacity Adjusted Contribution:

$$\sum_{j=1}^{J} \mathbb{E}[G_{i'}] - \sum_{r=1}^{R} V_r \sum_{n=1}^{N_i} \mathbb{E}[U^{i'}_{r,n}] = 0.5(270) + 0.5(300) - (50(2.5) + 40(3))$$

$$= 285 - 270 \geq 0$$

The Capacity Adjusted Contribution is 15, it is greater than 0, and therefore the patient's contribution is enough to be considered for admission to the hospital. Now we must see if there are enough resources available to provide the patient with the care needed.

The first constraint we check,

$$c_1 - \gamma_1 - \left(\sum_{k=1}^{k'-1} z_{1,t'}^k + \mathbb{E}[U_{1,1}^{i'}] \right) = 8 - 3 - (6+1) \not\geq 0.$$

does not hold, so we move on and check if patient can be admitted on some other day. The first constraint is broken at time $t'+1$ as well, so the last possible day of admission,$t'+2$, is checked after that.

i) For $r = 1$

First day of stay: $c_1 - \gamma_1 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+2}^k] + \mathbb{E}[U_{1,2}^{i'}] \right) = 8 - 3 - (4+1) \geq 0.$

Second day of stay: $c_1 - \gamma_1 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+3}^k] + \mathbb{E}[U_{1,2}^{i'}] \right) = 8 - 3 - (4+1) \geq 0.$

Third day of stay: $c_1 - \gamma_1 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{1,t'+4}^k] + \mathbb{E}[U_{1,3}^{i'}] \right) = 8 - 3 - (2+0.5) \geq 0.$

ii) For $r = 2$

First day of stay: $c_2 - \gamma_2 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{2,t'+2}^k] + \mathbb{E}[U_{1,1}^{i'}] \right) = 18 - 2 - (6+1) \geq 0.$

Second day of stay: $c_2 - \gamma_2 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{2,t'+3}^k] + \mathbb{E}[U_{2,2}^{i'}] \right) = 18 - 2 - (13+2) \geq 0.$

Third day of stay: $c_2 - \gamma_2 - \left(\sum_{k=1}^{k'-1} \mathbb{E}[Z_{2,t'+4}^k] + \mathbb{E}[U_{2,3}^{i'}] \right) = 18 - 2 - (0+0) \geq 0.$

All constraints hold, and the result is that the patient is admitted at time $(t'+2)$, or two days after he/she requested admission.

3.6 Simulation Parameters

The simulation that was carried out here included three resources, $R = 3$. The resources are $r = 1$ for regular beds, $r = 2$ for intensive care unit beds (ICU), and $r = 3$ for OR slots.

When comparing the heuristics, the hospital's reward from a period of one year is used. Before collecting results from the simulations, all heuristics were run

for 10 year 'warm up' time. This was done to remove the initial transient states of the system, and begin the collection of results from a state that would be in line with the current heuristic simulated. After the warm up time, the run time was equivalent to 1000 years, or 1000 periods. Statistically, this was shown to be sufficient for the results to be significantly different from each other.

For the comparison of the heuristics to be as good as possible, the following model parameters were chosen to be the same in all heuristics:

- Arrival processes of both emergency patients and elective patients are the same in all heuristics.
- The inter-arrival time of emergency patients is exponentially distributed with parameter $\mu = 0.4403$ days. Section 4.1 shows how the parameter was found.
- The inter-arrival time of elective patients is exponentially distributed with $\mu = 0.1701$ days. Section 4.3 describes how this value was found.
- The regular bed unit has 23 beds.
- The ICU bed unit has 18 beds.
- Total OR capacity is 128 slots each day (4 ORs, 8 hours available per OR, and each slot is 15 minutes long).
- The penalty is 712 reward units for one overbooked regular bed per day. For one overbooked ICU bed per day it is 194 reward units. A penalty of 318 reward units was used for one overbooked OR slot per day .

CHAPTER 4

Patient Arrival Processes

There are two main priority categories of patients that arrive at the hospital, elective patients and emergency patients. Emergency patients must always be admitted instantly because their treatment cannot wait without serious consequences. However, the hospital does not have to admit all elective patients that request admission. When estimating the elective patient arrival distribution, this raises difficulties, since in practice, hospitals rarely record the number of patients that they decline.

When some of the observed data points in a data set do not provide the full information of the true state, it is called *censorship of data*. In cases like these, where the true data point is either equal or higher than the observation, it is said that the data is *right censored*. The following sections will discuss methods to find distributions of the patient arrival process, both for *censored* and *uncensored* data. All emergency arrivals to the hospital are recorded, and thus the emergency arrival data is *uncensored*. The same does not apply to the elective arrivals, the data for elective patient demand is *censored*.

4.1 Estimating Arrivals of Emergency Patients

The distribution of emergency patient arrivals can be estimated by using the maximum likelihood estimator (MLE) on the data provided by the hospital. The

formula for the likelihood of a distribution f is

$$L(\theta|\vec{v}) = \prod_{\forall i} f(v_i|\theta), \tag{4.1}$$

where $\vec{v}$ is the vector of observations and θ is the parameter of the underlying distribution f.

To find the maximum of the likelihood, the standard method is logarithmic differentiation. Define the logarithm of the likelihood function as $\mathcal{L}(\theta|\vec{v}) = \ln L(\theta|\vec{v})$, and then

$$\mathcal{L}(\theta|\vec{v}) = \ln L(\theta|\vec{v}) = \sum_{\forall i} \ln f(v_i|\theta). \tag{4.2}$$

The MLE estimator is the maximization of the differentiated log-likelihood function.

Patient arrivals have been modeled as a Poisson process before, Gabrielson (1962) and Fetter and Thomson (1966), and we use that assumption here. If arrivals are Poisson distributed, the inter-arrival times between patients follow an exponential distribution and we are able to derive the average inter-arrival rate of patients, using the MLE method. The likelihood function of a exponential distribution with a parameter λ, and the log-likelihood function are,

$$L(\theta|\vec{v}) = \lambda^n e^{(-\lambda n \bar{v})}, \qquad \text{and,} \tag{4.3}$$

$$\mathcal{L}(\theta|\vec{v}) = n \ln \lambda - \lambda n \bar{v}, \tag{4.4}$$

respectively, where $\bar{v}$ is the mean of the observations. By differentiating the log-likelihood function and then maximizing the result, the MLE estimator for an exponential function is found to be

$$\hat{\lambda} = 1/\bar{v}. \tag{4.5}$$

The estimator $\hat{\lambda}$ most likely reconstructs the true parameter, λ, the best. It must be kept in mind that it is only an estimator, and that the more data points it is

based on, the more likely it is to be close to the true parameter.

The data used here to estimate the arrivals of emergency patients had 318 data points and was derived from the first 20 weeks of 2011. The mean was $\bar{v} = 2.2714$, and thus, the estimation of the parameter is

$$\hat{\lambda} = 1/\bar{v} = 0.4403. \tag{4.6}$$

If we were to use the maximum likelihood estimator to estimate the parameter of the distribution for elective patients, it could result in a distribution that underestimates the number of arrivals of patients. That would happen in cases where some of the observations might show a lower value than what the elective patient demand actually is, that is, the data is *(right) censored*. Therefore, methods to deal with situations where this occurs will now be covered.

4.2 Censored Data

It is common that facilities and organizations do not have all information needed about the past when estimating future development. When basing forecasts on incomplete data, difficulties arise that can be hard to overcome. Examples of this are retailers that have stockouts of a certain product. Since the product is out of stock, they are not able to capture the real demand and this results in an observation that is lower than the actual demand. Here, the hospital wants to estimate the arrival process of both elective and emergency patients. In practice, hospitals' historical service to patients is in most cases recorded by admissions, not demand. Therefore, there is a possibility that some of the observations are right censored.

The goal here is to obtain an estimator for the distribution function of the patient demand. Two simple methods could be used; i) either to ignore that

some observations are censored and treat the censored data the same way as the uncensored one, or ii) to simply eliminate the censored data points and use only the uncensored data. Neither of these options are desirable; they do not give any information about the data beyond the censoring limit, and both could result in badly biased estimates when a large proportion of the data is censored. Either of these methods could be used in cases where a small proportion of the data is censored, but that is not the case here.

Numerous statistical methods have been introduced in recent years to better deal with censored data. The most common ones are the Kaplan-Meier method and the EM-Algorithm. Both will be covered in the following sections.

4.2.1 Kaplan-Meier estimator

The estimator was originally developed to estimate a survival function from life-time data. It is one of the most studied topics in statistics, and is commonly used in medical statistics. For example, it was used to measure the proportion of patients living for a certain period after a treatment. It is applicable when data sets are *right censored*, and can also be used for data that has been *left truncated*[1]. It has the advantage of being applicable to nonparametric probability distributions.

Let D_t be a random variable of patient admission demand in the time period t, that follows a probability distribution D. Different time periods have the same length, that is $|t_1| = |t_2|$, if t_1 and t_2 are two disjoint time periods, where $|*|$ is the length of the time period *. All time periods must be disjoint, the random variable D_t is identically distributed in all periods, and it is independent from other time periods.

Suppose there is admission data available for the period $t = \{1, \ldots, T\}$. Define

[1]With truncation, observations never appear outside some given range. Left truncation is when the observations are never lower than some specific value.

A_t as the number of admissions at time $t, \forall t$. Suppose the hospital does not record how many patients they refer elsewhere/decline admission, that is, the excess demand for admission is unknown. Then, when there is more demand than the hospital is able to admit, the observation is censored. Let the observation in each period be W_t. Then, each observation $W_t = \min\{A_t, D_t\}$. Since the hospital cannot admit more patients than the demand for admission is, the inequality $A_t \leq D_t$ must hold, and $W_t = \min\{A_t, D_t\} = A_t$.

To apply the Kaplan-Meier estimator on the sequence $\{A_t\}_{t=1}^{T}$, we define a binary indicator, or *censoring indicator*, δ_t to each observation W_t, where $\delta_t = 1$ if the observation is uncensored, and $\delta_t = 0$ if the observation is *censored*. If we have a censored observation, the only information available about the demand is that $D_t > A_t$. By applying appropriate indicators to the observations, we get a data set $\{(W_1, \delta_1), (W_2, \delta_2), \ldots, (W_T, \delta_T)\}$, where the observations are ordered in the sequence they take place. Now the observations are ordered from the smallest to the largest, by the value of W_t, if there are two or more observations that have the same value, the uncensored observations, $\delta_t = 1$ are put in front of the censored observations, but otherwise they are ordered arbitrarily. By letting V_1 be the observation with the smallest value of W_t (and is assigned the corresponding δ_1), V_2 be the observation with the smallest value of W_t that is greater than or equal to V_1, and so forward until V_T becomes the largest observation of W_t, we now have an ordered set $\{(V_1, \delta_1), (V_2, \delta_2), \ldots, (V_T, \delta_T)\}$. It holds $\forall t$ that $V_t \leq V_{t+1}$, and if $V_t = V_{t+1}$ then $\delta_t \geq \delta_{t+1}$.

The KM estimator can now be constructed with an iterative procedure:

1. Assign a weight of T^{-1} to all observations.

2. Find the censored observation with positive weight that has the smallest index. If there is no such that exists then terminate the procedure.

3. Allocate the weight of the censored observation found in step 2 equally between all observations that have an index that is higher than the index of the weight. This applies to both censored and uncensored observations.

4. Repeat step 2 and 3 until there are no more censored observations. If the highest indexed observation is censored, the procedure is not applied for that observation.

When the procedure has finished it has constructed an empirical cumulative distribution function (CDF) and the complimentary cumulative distribution function (CCDF) for the demand, denoted $F(d)$ and $\bar{F}(d)$ respectively. If the highest observation is uncensored, then $\bar{F}(d) = 0, \forall d \geq V_T$, that is, the probability that the demand is larger than V_T is 0. On the other hand, if the observation V_T is censored then the only thing we know about the tail of the distribution is the probability for $D > V_T$.

Kaplan and Meier (1958) give a formula for an estimate of the CCDF,

$$\bar{F}(d) = \prod_{t:V_t \leq d} \left[(T-t)/(T-t+1)\right]^{\delta_t}, \qquad \forall d \leq V_T. \tag{4.7}$$

If all the observations were uncensored, the KM estimator would correspond to the empirical distribution for the demand.

The advantages of the KM estimator are that it is computationally simple and statistically efficient. Also, it can be used for nonparametric distributions. However, it does not give any basis to estimate the distribution beyond the censoring point.

4.2.2 EM algorithm

The EM algorithm was introduced Dempster, Laird, and Rubin in 1977. There had been former investigations done based on a similar idea before that, but

Dempster et al were the first to introduce it in formal way. The algorithm uses the maximum likelihood estimator of a distribution in combination with the expected value to find an estimate that is more likely to represent the true parameter of the *censored distribution*. The algorithm can be used for both discrete and continuous distributions.

Suppose we have a vector $\vec{v}$ of observations with incomplete data. Let θ be the unknown parameter of the distribution of the data. Then the observations have a joint density function $f(\vec{v};\theta)$ The objective is to find an estimate for θ. The likelihood function for the observed data can be written as

$$L(\theta|\vec{v}) = \int f(\vec{v}|\theta)\,\mathrm{d}v \tag{4.8}$$

Finding the estimator for the distribution would be computationally "simple" in the case of uncensored data, like shown in chapter 4.1. However, it is hard to find the likelihood in the case of incomplete data. This is where the EM algorithm comes in.

We start by dividing the observed vector $\vec{v}$ into two different vectors, $\vec{y}$ and $\vec{z}$, having $\vec{y}$ the vector with all the complete data points, and $\vec{z}$ the vector with all data points that are censored. Now it is possible to write out the log-likelihood function as it would look like if all data were known,

$$\ln L(\theta|\vec{v}) = \mathcal{L}(\theta|\vec{y},\vec{z}). \tag{4.9}$$

Now suppose there is $\hat{\theta}$ available, which is a prior estimate to θ, and that we have the incomplete data set $\vec{v}$. It is possible to calculate the expected value of the log-likelihood function, conditional on $\hat{\theta}$ and $\vec{y}$:

$$Q(\theta,\hat{\theta}) = \mathrm{E}_{\vec{z}}[\mathcal{L}(\theta|\vec{y},\vec{z})|\vec{y},\hat{\theta}] \tag{4.10}$$

Given the known observations and and guessed parameters, Q can be taken as

the average possible value of the log-likelihood.

The EM-algorithm is as follows:

1. Define an initial value $\theta_1 \in \Theta$, and let $\hat{\theta} = \theta_1$.
2. **The E-step.** For the current $\hat{\theta}$, evaluate $Q(\theta, \hat{\theta}) = \mathrm{E}_{\vec{z}}[\mathcal{L}(\theta|\vec{y}, \vec{z})|\vec{y}, \hat{\theta}]$ for all $\theta \in \Theta$.
3. **The M-step.** Maximize $Q(\theta, \hat{\theta})$ over all possible θ. That is, solve $\theta_{k+1} = \arg\max_{\theta \in \Theta} \mathrm{E}_{\vec{z}}[\mathcal{L}(\theta|\vec{y}, \vec{z})|\vec{y}, \hat{\theta}]$.
4. Repeat step 2 and 3 until θ_k and theta θ_{k+1} are sufficiently close to each other to stop the algorithm. If the algorithm can stop, assign $\hat{\theta}_n = \theta_{k+1}$.

The evaluated parameter, $\hat{\theta}_n$, is used as the estimator for θ.

The algorithm is guaranteed to converge to some fixed value, but, there is no guarantee that the estimator converges to the maximum likelihood estimator. Another problem with the algorithm is that it can have a slow convergence, especially when there are a large number of parameters or the proportion of censored data is high.

4.3 Estimating Arrivals of Elective Patients

As mentioned before, the arrival data received from the hospital for elective patients is *censored*. Since this is the case, it would be advisable to use one of the methods mentioned before, either the KM-estimator or the EM-algorithm, to estimate the parameter for the distribution. Here, the EM-algorithm was chosen. Usually, when the EM-algorithm is carried out, there has to be some initial guess for the parameter, and it is common to use the KM estimator to do that. The

other reason is that we assume that patient arrivals are Poisson distributed, and thus, have an exponentially distributed inter-arrival time. As McLachlan and Krishnan (1997) show, the EM-algorithm estimator has the property of converging to

$$\hat{\mu} = \sum_{j=1}^{J} v_j / r, \tag{4.11}$$

where J is the total number of observations, and r is the number of uncensored observations.

The data used here to estimate the arrivals of elective patients was derived from the first 20 weeks of 2011. The mean of arrivals per day was $\bar{v} = 5.78$, with 50 days where the demand was censored, out of 100 days total. The estimation of the mean of the distribution is

$$\hat{\mu} = \sum_{j=1}^{J} v_j / r = 5.78. \tag{4.12}$$

The parameter of the distribution becomes

$$\hat{\lambda} = 1/\hat{\mu} = 0.1730. \tag{4.13}$$

CHAPTER 5

Results

5.1 Results from heuristic 1 - No Resources Reserved for Emergency Patients

The first heuristic carried out in the simulation does not save any resources for emergency patients, as described in section 3.4. The hospital accepts elective patients as long as there are enough units of all resources available for the patient. Since the emergency arrivals are constant throughout the period, and the hospital has to admit all emergency patients, the expected results from the simulation would be a considerable amount of overbooking. The output vector from the simulation is denoted $\vec{x} = (x_1, x_2, \ldots, x_N)$, where x_n is the nth run of the simulation. Each output can be regarded as a random variable, and thus it is possible to find a sample mean and a sample variance.

Total repetitions were $N = 1,000$. Since the repetitions are $1,000$ the central limit theorem is applicable, that is, the output can be said to follow a normal distribution. The results from the simulation were the following:

$$\text{Sample mean } = \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = -95,520.$$

$$\text{Sample variance } = s^2 = \frac{1}{N-1}\sum_{n=1}^{N} (\bar{x} - x_n)^2 = 51,899,078.$$

$$\text{Sample standard deviation } = \sqrt{s^2} = 7,204.$$

These results are also listed in table 5.1. Using the results, the 99% two sided

	Reward Units
Sample Mean	-95,520
Sample Variance	51,899,078
Sample Standard Deviation	7,204
Margin of Error	18,558

Table 5.1: *Results from simulating heuristic 1, where there are no reservations of resources for emergency patients.*

confidence interval for the rewards for the simulation is

$$\bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\bar{x}-x_n)^2}.$$

This turns out to be

$$-114,078 \leq \bar{x} \leq -76,963,$$

where the z-statistic is $z_{1-\alpha/2} = 2.576$ for $\alpha = 0.01$.

The result, that the hospital's reward from using heuristic 1 is low in the presence of penalties for overbooking, is not surprising. However, it is not necessarily true that the prediction would have been that the reward would turn out to be negative. It is clear that the hospital has to reserve some amount of resources for elective patients for the rewards to be positive.

5.2 Results from heuristic 2 - A proportion of Resources Reserved for Emergency Patients

It was decided to simulate this heuristic with a proportion of 20% of each resource reserved for emergency patients. Elective patients are not admitted if the hospital has reached 80% capacity of any of it's resources. Results are shown in table 5.2,

	Reward Units
Sample Mean	246,067
Sample Variance	19,674,106
Sample Standard Deviation	4,436
Margin of Error	11,426

Table 5.2: *Results from simulating heuristic 2, where the reserved proportion of resources for emergency patients is 20%.*

and a 99% two sided confidence interval for the hospital's reward is

$$234,641 \leq \bar{x} \leq 257,493,$$

where the z-statistic is $z_{1-\alpha/2} = 2.576$ for $\alpha = 0.01$.

This is a great improvement from previous heuristic. Not only is the reward the hospital receives positive, it improves drastically.

5.3 Results from heuristic 3 - Reservations for Emergencies and Contribution Constraints on Electives

Here, the hospital reserves a specific amount of each resource for emergency patients. Also, elective patients must have a positive capacity adjusted contribution.

Version A: The reservation of resources were, $\gamma_1 = 2$ for regular beds, $\gamma_2 = 1$ for ICU beds, and $\gamma_3 = 84$ for OR slots. The shadow prices were; $V_1 = 183.878$, $V_2 = 56.5364$, and $V_3 = 0$. Results from the simulation can be seen in table 5.3.

The 99% two sided confidence interval for the hospital's reward when using

	Reward Units
Sample Mean	347,239
Sample Variance	8,397,672
Sample Standard Deviation	2,898
Margin of Error	7,465

Table 5.3: *Results from simulating heuristic 3, version A, where* $\gamma_1 = 2, \gamma_2 = 1$, *and* $\gamma_3 = 84$. *The shadow prices are* $V_1 = 183.878$, $V_2 = 56.5364$, *and* $V_3 = 0$.

	Reward Units
Sample Mean	123,478
Sample Variance	6,014,634
Sample Standard Deviation	2,452
Margin of Error	6,318

Table 5.4: *Results from simulating heuristic 3, version B, where* $\gamma_1 = 12, \gamma_2 = 9$, *and* $\gamma_3 = 84$. *The shadow prices are* $V_1 = V_2 = V_3 = 0$.

heuristic 3, version A, is

$$339,774 \leq \bar{x} \leq 354,704$$

where the z-statistic is $z_{1-\alpha/2} = 2.576$ for $\alpha = 0.01$.

The results here improve both heuristic 1 and 2, if we look at the sample mean, and the confidence interval for the sample mean.

Version B:

The reservation of resources were, $\gamma_1 = 12$ for regular beds, $\gamma_2 = 9$ for ICU beds, and $\gamma_3 = 84$ for OR slots. The shadow prices were; $V_1 = V_2 = V_3 = 0$. Results from the simulation can be seen in table 5.4.

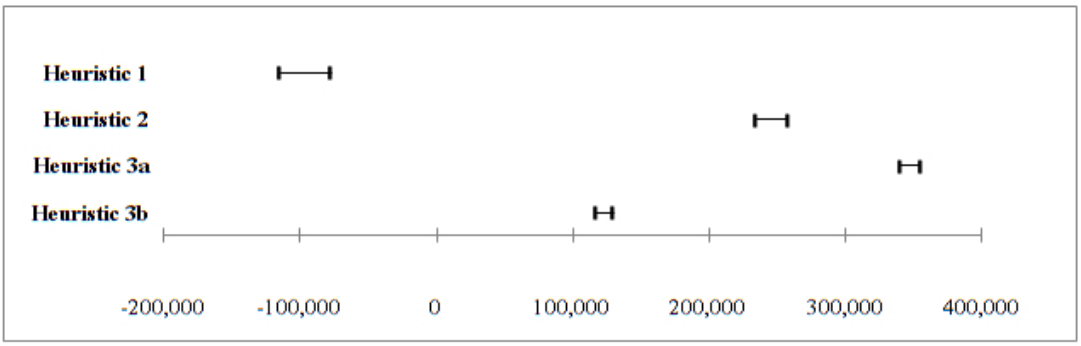

Figure 5.1: *Confidence intervals for the hospital's reward from each heuristic.*

The 99% two sided confidence interval for the hospital's reward when using heuristic 3, version A, is

$$117,160 \leq \bar{x} \leq 129,795$$

where the z-statistic is $z_{1-\alpha/2} = 2.576$ for $\alpha = 0.01$.

5.4 Comparing Results

From results above it can be seen that version A of heuristic 3 does best in the simulations. The second best heuristic is heuristic 2, followed by version B of heuristic 3. The heuristic that does the worst is heuristic 1, and it is the only heuristic where the hospital ends up with negative reward. The confidence intervals for the means of the rewards do not overlap for any of the heuristics. Therefore, there is a statistically significant difference between the rewards of all the heuristics. This is shown graphically in figure 5.1.

Heuristic 3.A results in a 41.12% higher reward than heuristic 2, and 181.22 % higher reward than heuristic 3.B (comparing the sample means). To figure out why this was happening it was decided to break the reward down by the hospital's mean total contribution and penalties from the simulation runs. This is demonstrated in figure 5.2 and 5.3. These show that the penalties from heuristic

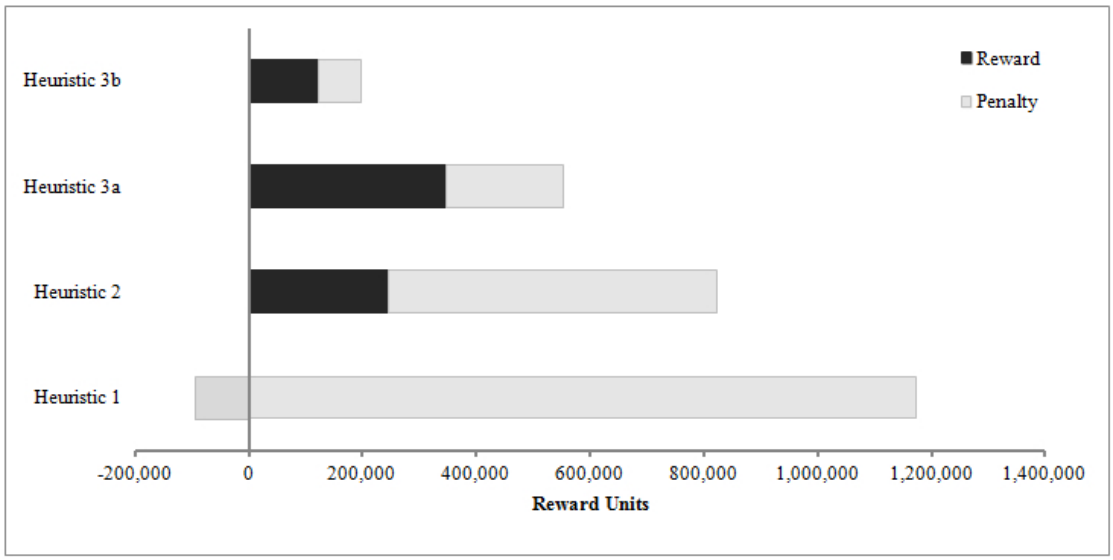

Figure 5.2: *The reward and penalty are shown out of the total contribution for each heuristic.*

3.A is eating up the least amount of the rewards, both when looking at the actual reward units and proportionally out of the total rewards. Even though heuristic 1 and 2 have considerably higher total contribution than heuristic 3.A, the reward ends up being lower because of too high penalty, caused by overbooking. The penalties are greater than the contribution in heuristic 1, resulting in a negative reward. Heuristic 3.B has a lower reward than both heuristic 2, and heuristic 3.A. In fact, the total contribution in heuristic 3.B is lower than the total rewards from both heuristic 2 and 3.A. The reason is probably that the admission policy is too protective, it reserves to much of the resources for emergency patients and thus does not allow for many admissions of elective patients.

Looking at the proportion of elective patients that are admitted, figure 5.4, it partially explains the difference in the contribution between heuristics. The first heuristic admits almost all patients since there is no reservation of resources, but then does not have room for all the emergency patients that arrive and must be admitted. An admission of 99% is really high, but it must be kept in mind that

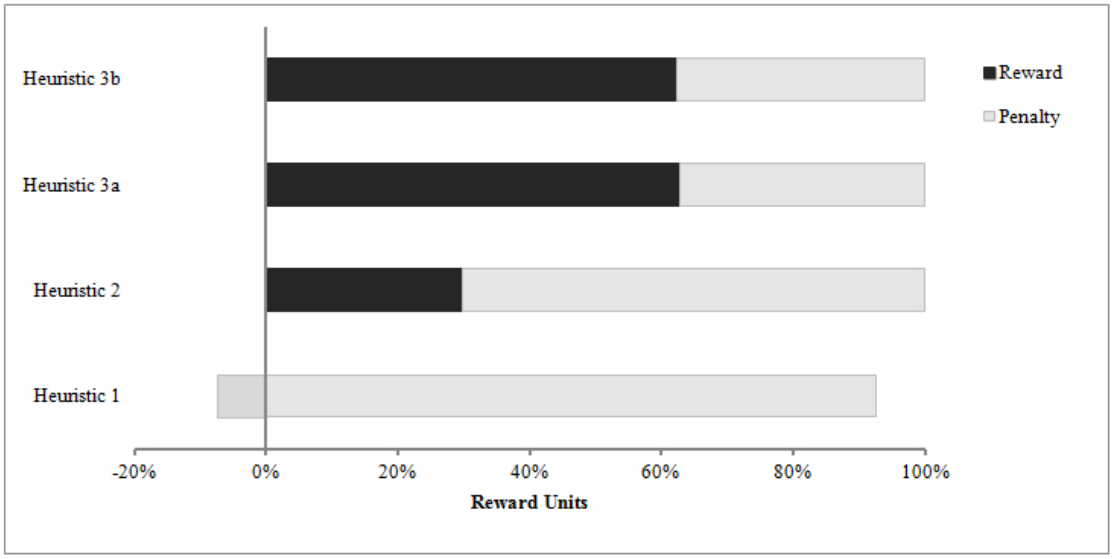

Figure 5.3: *The reward and penalty are shown in proportion to the total contribution for each heuristic.*

the hospital is able to delay admissions of elective patients, and thus fills up its resource capacity few days in advance.

The admission policy in heuristic 2 does not result in as great admission rate. However, the admissions are rather high, or just over 3/4 out of total requests. Therefore the penalties are not as great as in the first heuristic, but the total penalty is much higher than in heuristic 3.A. Heuristic 3.A was derived from running a program to find how much of each resource to reserve, based on the data from the hospital. Therefore, the reservations of resources are more specified. It is not necessarily true that the same amount of all resources should be reserved, one of them might have more emergency demand than the other and therefore different proportions of reserved resources might be a better option.

Heuristic 3.A accepts about 50% of elective admission requests. It is more selective of patients than heuristics 1 and 2 and makes sure not to fill up the hospital capacity with patients that give little or no contribution to the hospital. Since neither heuristic 1 or 2 have any requirements for patient's contribution,

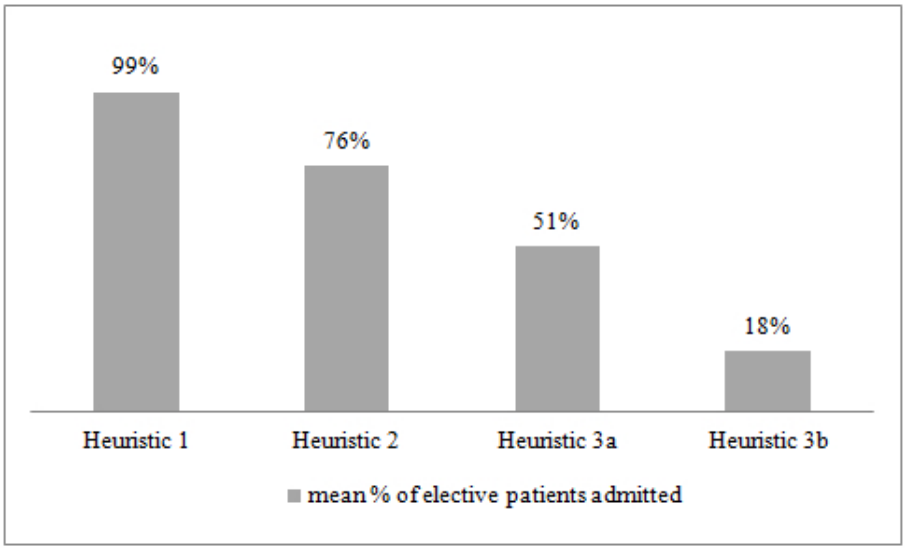

Figure 5.4: *The proportion of elective patients admitted to the hospital out of all that requested admission.*

they might be admitting those patients. Heuristic 3.B does not make as a strong selection of patients as heuristic 3.A. Therefore, the reason for the low admission proportion is probably because it reserves too much of the resources for emergency patients.

When looking at the mean number of days overbooked, figures 5.5 and 5.6, it is interesting to see that the only time regular beds are overbooked on more days than the ICU beds is in heuristic 1. This might be because of no reservations of resources for emergency patients, the hospital is able to accept more elective patients, and it is likely that they do not need as much stay in the ICU units as emergency patients.

Heuristic 3, both version A and B, overbook the ICU units on more days than the other resources. This might be because penalty for each overbooked ICU unit is lower than the penalty for overbooking one resource of the other resources, and that the procedure in the heuristic picks up on that. Also, it might be because it manages to forecast the need for elective patients better than the other heuristics

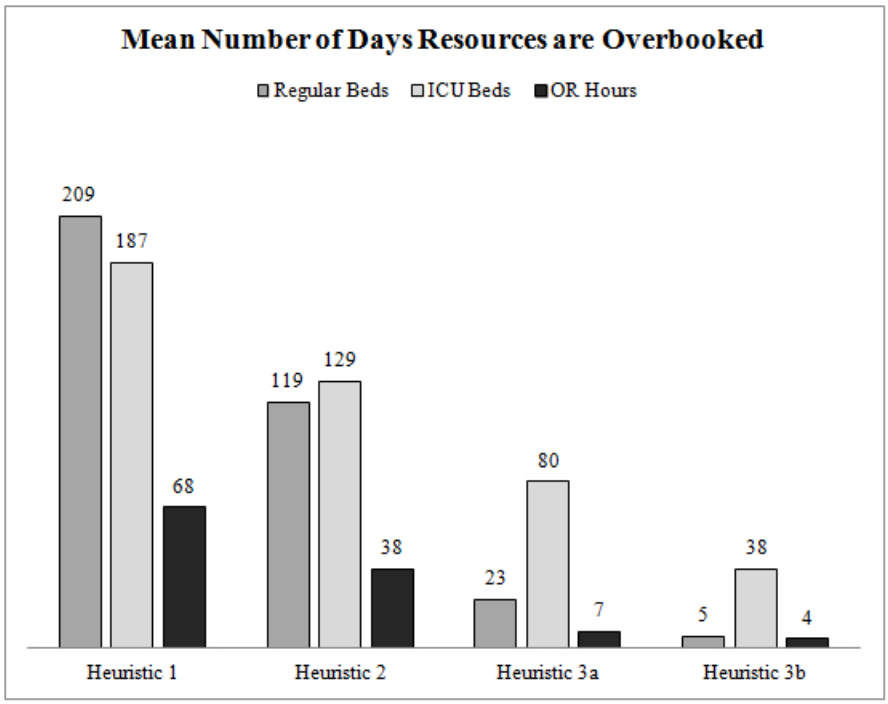

Figure 5.5: *Shows the mean number of days that each resource was overbooked over a period of one year.*

do. Emergency patients are more unpredictable, and have a higher usage of ICU beds than elective patients. Thus, the method might not capture the emergency need for ICU beds accordingly.

Looking at the number of days the OR slots are overbooked , it is interesting to see that they are by far overbooked on less days than beds. It seems that the OR is not as big of a bottleneck in the hospital's patient treatment as the beds.

Taking all the results together it looks like a revenue based heuristic does increase the total reward the hospital receives over the simulated period. It is recommended to reserve resources for patients, although careful consideration should be made about how much to reserve.

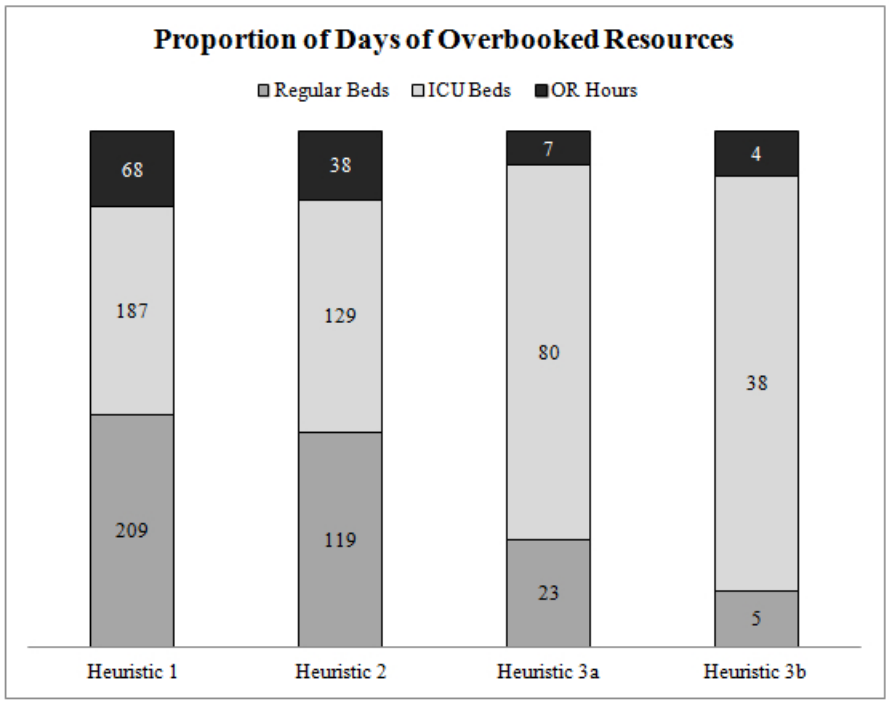

Figure 5.6: *Shows the proportion of days that each resource was overbooked out of the total count of days where resources were overbooked.*

CHAPTER 6

Data Analysis

The data used in the simulations was provided by the Neurosurgery Department at the UCLA Ronald Reagan Hospital. The data covered the period from January 1st 2011 to June 30th 2011. It mainly contained information about resource capacity of the department, patient admission information and treatment information, along with revenues and cost for treating patients. This data was analyzed in order to find patient arrival processes, patient groups, given various characteristics of patients, including resource requirements and contribution to the hospital.

It should be mentioned that due to patient confidentiality, the revenue and cost figures were scaled in such a way that they are not recognizable, but are still in proportion to each other.

6.1 Hospital's Resources

The hospital's data said that the number of regular beds for the unit were 46, and that there were 24 beds for the ICU unit. However, conversation with employees from the hospital suggested that the beds might also be used by patient from other departments than the neurosurgery department. Analysis of the data and admissions of elective patients suggests a 50% utilization of regular beds and 75% utilization of ICU beds would be a good approximation for the simulation. Therefore we ended up having the number of regular beds 23, and ICU beds 18.

There are four ORs available for the neurosurgery department. The assump-

tion here is that the OR is used for 8 hours a day, all days of the week. This results in a 32 hour availability per day. OR time is charged by every 15 minutes. Therefore, in the simulation, we use OR slots, and they are $4 \times 32 = 128$ per day.

6.2 Patient Arrival Processes

Patients' admissions over the first 20 weeks of 2011 (from the data provided by UCLA Ronald Reagan Hospital) were used to estimate the arrival process for both elective and emergency patients. It was translated into number of arrivals per day. The data suggested that emergency patients arrive on both weekdays and weekends, but elective patients are not admitted on weekends (there was only one admission of an elective patient on a weekend in the whole period, which I consider an exception and do not include). Therefore, it is assumed here that the elective demand is zero over the weekend.

It can happen that patient demand to different units in hospitals varies between time periods (e.g. before or after noon, between weekdays) as Swartsman (1970) mentions. The data used here did not include patient arrival time during the day, only the day the patient was admitted, and thus, it was not possible to see if patient demand was different between periods during the day. However, it was possible to check whether there was a significant difference in patient demand between weekdays. Here, I used the Kolmogorov-Smirnov test (K-S test), to see if there was a statistically significant difference between the arrival process each day. One of the advantages of the K-S test is that the data does not have to follow any specific distribution for the test to be applicable.

The null hypothesis is that there is not a statistically significant difference between demand on any weekday compared to another. The test was done with a significance level of $\alpha = 0.01$.

The results from comparing emergency demand between weekdays are shown

in table 6.1. They show that in most cases the p-value is very high. There are two p-values lower than 0.05. Both happen when comparing arrivals on a Friday to other days. However, there are no p-values lower than the α chosen, and thus, the conclusion is that there is not a significant difference between the arrival distribution of emergency patients between weekdays.

Results from elective patient demand comparison between weekdays were similar to the comparison of emergency patient demand. They are shown in table 6.2. There is one p-value below 0.05, but apart from that they are rather high. There was no p-value below the α level we chose. Since I cannot find strong evidence against the hypothesis that the distribution of elective patient admission is the same between weekdays, I used that the arrival rate was the same on all weekdays (weekends excluded).

It should be mentioned that the comparison of elective demand between weekdays had to be done by using the censored data for elective demand. The censored data does in fact tell us how many patients were admitted per day, but not how great the elective patient demand was. However, this was the only data available to apply the comparison to, and thus it was used to see if there is a difference in elective demand between weekdays.

The results from the analysis of patient arrival process was to have the arrivals of patients distributed the same on all days, with the exception of weekends for elective patients. Elective patients are not admitted on the weekends and thus the arrivals of elective patients is zero on those days.

6.3 Categorization of Patients

The data received had various important information about each patient, but patients were not categorized in patient groups that were helpful in order to estimate resource usage. Included in the information about patients were, in no specific

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	0.7710					
Tuesday	0.4973	0.999				
Wednesday	0.7710	0.7710	0.7710			
Thursday	0.2753	0.4973	0.4973	0.7710		
Friday	0.0232	0.4973	0.7710	0.0591	0.0232	
Saturday	0.9655	0.9655	0.7710	0.9655	0.4973	0.0591

Table 6.1: *p-values from comparing emergency arrival distribution between weekdays.*

.

	Monday	Tuesday	Wednesday	Thursday
Tuesday	0.771			
Wednesday	0.2753	0.771		
Thursday	0.9655	0.771	0.2753	
Friday	0.9655	0.1349	0.0232	0.9655

Table 6.2: *p-values from comparing elective arrival distribution between weekdays. There are no elective arrivals during Saturdays and Sundays and therefore those days are not included.*

order, admission diagnosis, main diagnosis, DRG group, surgery procedure, length of stay, OR slots, priority, insurance information, revenue, cost, etc. All of those mentioned (and more) are possible candidates for being used to somehow estimate a new patient's resource usage that the hospital is considering to admit.

The data needed careful analyzing to see what would be a good way to group patients with regards to their possible resource usage and contribution. Main diagnosis, patient priority, DRG group, severity, and insurance coverage of patient were likely candidates. However, some of this information might not be known at the time of admission. Thus, to make the grouping of patients more realistic, I decided to see if patients could somehow be grouped by three of the things that are known at the time of admission, priority, admission diagnosis and insurance coverage of the patient.

The results showed that an applicable way to group patients is by priority and main diagnosis, where in some cases patients with the same main diagnosis have to belong to different patient groups. The following sections explain in detail how the groups are defined, how the analysis took place, shows tests applied, and examples of patient grouping.

6.4 Patient Grouping

The main purpose of grouping patients in this work is to find patients that have similar resource usage and their characteristics in order to estimate future resource usage. Patients' resource usage is stored in a matrix where each row represents a resource and each column represents the number of the day of a patients stay.

The resource matrix is written as

$$u = \begin{bmatrix} u_{1,1} & \dots & u_{1,n} & \dots \\ \vdots & \ddots & & \\ & & & \\ u_{r,1} & & u_{r,n} & \\ \vdots & & & \ddots \end{bmatrix} \tag{6.1}$$

where an element $u_{r,n}$ represent the resource usage of resource r on day n.

By finding the characteristics of patients that have similar resource matrices, the hospital can better estimate the possible resource usage at the time of admission. By having a better estimate of resource usage, it becomes more likely that the hospital schedules elective patients in a way such that overbooking does not occur while still serving as many elective patients that bring contribution to the hospital as possible.

6.4.1 Priority of Patients

The hospital uses four different admission types; elective, emergency, trauma and urgent patients. Elective patients are not admitted unless it is estimated that the hospital has enough resources available for them. Emergency, trauma and urgent patients all have to be admitted to the hospital, even though the hospital's resources have reached full capacity. Therefore, trauma, urgent and emergency patients in our analysis are all treated the same, and are referred to as emergency patients.

Using the Kolmogorov-Smirnov (K-S) test to compare the resource usage of elective patients versus emergency patients, we get that there is a significant difference. An example of comparison is shown in table 6.3. In this case there is a significant difference between the length of stay of each patient priority. Since most other tests for the difference in resource usage between priorities were along the similar lines, it was decided to use the priority of patients when defining patient types.

Length of Stay	Days in ICU	OR slots
0.0073	0.2695	0.2129

Table 6.3: *p-values from comparing underlying distributions of resource usage for emergency patients and elective patients using K-S test at significance* $\alpha = 0.01$.

6.4.2 Admission Diagnosis

Admission diagnosis can reveal some clues about a patient's future resource usage. Patients with the same condition tend to have similar surgery procedures and similar amount of recovery time. However, each individual has their own response to treatment, and thus, there is always some chance of complication and almost certain that there will be some randomness between patients resource usage, even if they have the same diagnosis.

Admission data revealed in total 212 possible diagnoses at the time of patient's admission. The admission diagnoses are classified by the International Classification of Diseases (IDC). Each diagnosis has a main category, which describes the general condition of the patient, and a sub group, which explains in more detail the nature of the patient's condition. An example of this would be the diagnoses of Cervical Disc Displacement (Medical Code 722.0) and Lumbar Disc Displacement(Medical Code 722.2). Both fall under the main group of Intervertebral Disc Disorders (Medical Code 722), but they affect different areas in the spine. Even though they do affect different areas they might have similar treatment, and thus, it is worth to see if patients with these diagnoses can belong to the same patient type in our analysis.

As an example it is possible to look at main group which we will call X1, within elective patients. According to the data there are three different sub diagnosis for

Diagnosis	Length of Stay p-value	Days in ICU p-value	OR slots p-value
X1.a	0.2073	N/A	0.1828
X1.b	0.7201	0.0514	0.4402
X1.c	0.0036	3.63×10^{-9}	3.65×10^{-16}

Table 6.4: *Results from K-S test where the null hypothesis is that the usage of each resource is normally distributed*

this condition, X1.a, X1.b and X1.c, and the resource usage of patients belonging to those groups varies to some degree both within them and between them. To decide what kind of statistical test to use to test similarities between resource usage of the different diagnoses we have to figure out if the data follows a specific probability distribution. The most probable distribution for this kind of data would be the normal distribution. We use Kolmogorov-Smirnov (K-S) test for each of the resources, for all three admission diagnosis to see if the usage can be said to follow a normal distribution. The results can be seen in table 6.4, and they show that we cannot assume a normal distribution for all types of the resource usage.

When looking at patients' estimated resource usage, these patients can in some cases be said to belong to the same patient type. We wanted to incorporate this into the data usage to try to better determine the amount of resources that the patient will need for his/her treatment. However, it is not possible to merge all the subgroups of all main diagnosis into one patient type. Therefore, we used statistical tests to figure out which sub groups within a main diagnosis are similar when it comes to patient's resource usage, and which one are not.

Since the test shows that the resource usage does not follow a normal distribution for all of the three different diagnoses, it would not be advisable to use

Admission Diagnoses Compared	Length of Stay	Days in ICU	OR slots
X1.a, X1.b	0.8894	0.5134	0.5971
X1.a, X1.c	0.0462	1	6.59×10^{-5}
X1.b, X1.c	0.3957	0.4061	3.49×10^{-4}

Table 6.5: *p-values from comparing underlying distributions of resource usage for different admission diagnoses.*

the two-sample t-test to compare difference in resource usage between them. The two-sample t-test assumes that the underlying data follows a normal distribution. Instead we want to use a nonparametric test and the two-sample Kolmogorov-Smirnov test is one and it was used to compare the different resource usage. The null hypothesis in the test is that the underlying distributions of the two samples are the same. It is rejected if the p-value is lower than the significance level α. The resulting p-values are shown in table 6.5.

When we compared X1.a and X1.b we got fairly high p-values. Thus, we are never able to reject the assumption that the underlying distributions are the same. For groups X1.a and X1.c the p-values for days in ICU and for OR slots are high. But for length of stay we have a p-value of $0.0462 \leq \alpha = 0.05$. Therefore we reject that the length of stay for these samples have the same distribution. X1.b and X1.c; the p-value for OR slots is $3.49 \times 10^{-4} \leq \alpha = 0.05$ and thus we reject that the OR slots are distributed the same for both admission diagnoses. The result would thus be to merge groups X1.a and X1.b into the same patient type, but keep X1.c as a separate patient type.

The K-S test for comparing two different samples is said to be valid for sample sizes $n \geq 4$. There were cases where the sample was of size lower than four. In

Admission Diagnoses Compared	Length of Stay	Days in ICU	OR slots
996.A, 996.C	0.0577	0.5941	0.0018

Table 6.6: *p-values from comparing underlying distributions of resource usage for admission diagnoses 996A and 996C.*

those cases the samples were joined based on inspection. An example of this is for subgroups of main diagnosis which shall be called X2. Group a has 21 data points, group b has 6 data points and group c has 1 data point. We could use the K-S test for comparing groups a and b but the K-S test wouldn't give us a reliable result if we use it to compare group c with the other groups, since group c only has one data point. We start by checking if there is a significant difference between group a and b using the K-S test. The results are in table 6.7, and show that there is a significant difference between the distribution of OR slots. Now we have to figure out what to do with group c, it could belong to either the types for a or b, or it should be a separate type. In this case we look at the data point in group c, and see that, for example, there is a great difference between the length of stay when comparing to group b. However, the length of stay is within the lines of group a, and thus, we group a and c together, and have group b as a separate type.

6.4.3 Contribution from patients

Contribution is defined as net revenue minus the direct cost from treating a patient. When looking at expected contribution from admitting a patient, patients are categorized differently from when we are looking at the resource requirements.

	Group X2.a			Group X2.b			Group X2.c		
n	LOS	ICU	OR	LOS	ICU	OR	LOS	ICU	OR
1	2	0	2.5	98	0	10	6	0	1.5
2	2	0	3.25	37	24	8			
3	12	4	7.25	1	0	2.75			
4	1	0	0	20	5	8.25			
5	1	1	1.5	6	0	4.75			
6	2	2	1.75						
7	5	0	2.5						
8	1	0	0						
9	4	0	0						
10	2	0	3						
11	5	0	2.5						
12	28	19	1.25						
13	5	0	0						
14	7	0	2.25						
15	5	0	1.5						
16	2	0	0						
17	4	0	0						
18	14	0	1.75						
19	2	2	0						
20	2	1	1.75						
21	1	0	1.25						

Table 6.7: *Length of stay, ICU usage and OR slots used by patients from all patient entries for diagnoses X2.a, X2.b, and X2.c.*

Contributions from patients are dependent on the insurance plan they have and therefore it should be checked if patients should be categorized by that as well.

Applying similar tests as for the priority and admission diagnoses of patients, the result was to not divide patients between types based on the insurance plan they are under. The difference was not great enough to justify such a distingyishon. One of the reasons for that might be that there was not enough data points for the different insurance plans within the admission diagnosis. If the data would be collected over a longer period, the difference in contribution between patients based on insurance plan might be seen.

6.4.4 Patient Types Defined

The results above suggest that a way to categorize patients would be by priority and main diagnoses, with some types divided further by sub diagnoses. Even though it is suggested by the hospital that the insurance coverage has an effect on the contribution from patients, the data did not provide strong enough support for it, and thus, it was not taken into account when the patient types were formed.

After going though all main categories of diagnoses and merging applicable sub diagnoses within them, the results were that patients could be divided into 139 types, 57 of the types were for elective patients, and 82 types for emergency patients. Unfortunately, here were still a number of few groups which had only one entry to base the distribution of resources on, but the number reduced from 252 to 57. A bigger proportion of the types with a single entry belonged to emergency patients, or 36 types.

6.5 Constructing Resource Matrices

For every patient record in the data there was a matrix constructed to describe the patients resource usage. The resource matrix is of size $R \times N$, where R is

the total number of resources, and N is the length of the patients stay. When constructing a patients resource matrix the data columns that were used were; admission date, length of stay, days in ICU unit, OR slots, and surgery date.

An example of how a resource matrix was constructed from an admission record where admission date was 2/1/2011, length of stay was 5 days, days in ICU units were 2 days, OR usage was 3.75 hours and surgery date was 2/3/2011 is

$$u = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 15 & 0 & 0 \end{bmatrix}.$$

The first line of the matrix stands for the regular bed usage of the patient, the second line for the ICU bed usage of the patient, and the third line for the OR usage. Columns stand for the day of the patient's stay. If there is no usage of a resource on some day, that element in the matrix gets the value of zero.

Patient's bed usage is split between the first line and the second line of the resource matrix. If the patient uses a regular bed on a specific day, that element gets a value of 1. A patient cannot use both a regular bed and an ICU bed over the same night. Thus, if there is the value 1 for a regular bed usage on a specific day of stay, the ICU usage is 0 for that same day, and vise versa. Since the data did not include on which days the patient stays in the hospital, the assumption was made that ICU stay starts on the day of surgery.

The hospital charges for every 15 minutes of OR usage. Therefore, an OR slot is defined to be 15 minutes. The surgery took 3.75 hours, and changing this into OR slots, the result is $3.75 \times 4 = 15$ slots. The surgery took place on the third day of the patients stay, and thus, the 15 OR slots the patient uses are placed in the third column.

Patients are categorized by patient type, denoted with $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of all patient types. The total number of possible resource matrices is $\sum_{\forall i \in \mathcal{I}} J_i$,

where J_i is the number of possible resource matrices for patient type i.

6.6 Probabilities of Resource Matrices

Each patient type $i \in \mathcal{I}$ has at least one, or more, different resource matrices that are derived from the data provided by the hospital. When a patient arrives to the hospital, the hospital wants to find an estimation of the future resource usage of that patient. The idea is to use the hospital's knowledge about prior patients belonging to the same patient type to estimate future care needed for arriving patients.

Suppose the hospital has admitted 5 patients that belong to type i, of which 2 had resource matrix $u^{1|i}$, 2 had resource matrix $u^{2|i}$, and 1 had resource matrix $u^{3|i}$. Then, if a a patient of type i arrives, there is a probability of

$$p_{1|i} = \frac{\text{Number of patients of type } i \text{ with resource matrix } j = 1}{\text{Total number of patients of type } i} = \frac{2}{5} = 0.4,$$

that the patient has a resource matrix $u^{1|i}$. Similarly, for this case, the probabilities of the other resource matrices are $p_{2|i} = 0.4$, and $p_{3|i} = 0.2$.

6.7 Expected Total Resource Usage and Contribution of Patient Types

A patients total resource usage of resource r is the sum of the resource usage of a patient over all days of the patients stay. The expected total resource usage of resource r for a patient type i is

$$\sum_{n=1}^{N_i} \mathbb{E}[U_{r,n}^i] = \sum_{j=1}^{J_i} \sum_{n=1}^{N_i} p_{j|i} u_{r,n}^{j|i}, \qquad \forall i \in \mathcal{I}, r \in \{1, \ldots, R\}. \tag{6.2}$$

Each patient type i has an expected contribution of

$$\mathbb{E}[G_i] = \sum_{j=1}^{J_i} p_{j|i} g_{j|i}, \qquad \forall i \in \mathcal{I}.$$

6.8 Demand by patient types

Patients belong to specific patient types $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of all patient types. Define $\mathcal{I}_{EL}$ as the set of all patient types that elective patients can belong to, and $\mathcal{I}_{EM}$ as the set of all types emergency patients belong to. Then $\mathcal{I}_{EM} \cup \mathcal{I}_{EM} = \mathcal{I}$.

If D_i is an exponentially distributed random variable for the demand of patients of type i per day, and q_i is the probability that a patient that arrives at the hospital belongs to type i, then the expected value of patients belonging to type i per day can be found by using equation

$$\mathbb{E}[D_i] = q_{i|EL}\frac{1}{\lambda_{EL}}, \tag{6.3}$$

where λ_{EL} is the parameter of the distribution of all elective patient arrivals. The probability that an elective patient belongs to patient type i is found by

$$q_{i|EL} = \frac{\text{Number of type } i \text{ patient arrivals}}{\text{Total number of elective arrivals}}, \tag{6.4}$$

and by using the data provided by the hospital to find the probabilities. The probabilities for emergency patients are found using the same method. Since $q_{i|EL}$ and $q_{i|EM}$ are probabilities it must hold that $\sum_{i\in\mathcal{I}_{EL}} q_{i|EL} = 1$, and $\sum_{i\in\mathcal{I}_{EM}} q_{i|EM} = 1$.

6.9 Maximum demand a day per patient type

To find this, the admissions of each patient type i per day was summed up for each day of the period the data covered. Then the maximum number of admissions was found for each type.

6.10 Penalties for Overbooking Resources

The simulation model aims to maximize reward, which is defined as the contribution from patient care less the penalties from overbooking resources. If any of

the hospital's resources are overbooked, there will be some additional cost for the hospital, which could be both because of cost directly related to the hospital having to take measures to accommodate the patient, or lost goodwill from patients because they get worse service than otherwise. Therefore, penalties are included in the reward function if any of the resources have been overbooked.

The data had information about the cost of treating patients and the revenues they bring to the hospital. In some cases the cost of patient's care was greater than the revenue, which leads to a negative reward for the hospital. Therefore, the rewards from treating patients were adjusted, by adding the most negative reward to all of the rewards, resulting in no negative reward values in the simulation. Patients' identity and personal information must be protected. Therefore the total rewards from patients were adjusted in a manner such that they cannot be recognized, but are still proportional to each other.

To find the penalties used in the simulation we used the data based on patient types that included the insurance coverage, since both resource requirements and rewards have to be taken into account. In the simulation there were $R = 3$ resources, and thus, let π_1 be the penalty for one overbooked unit of resource 1 (regular beds), π_2 be the penalty of one overbooked unit of resource 2 (ICU beds), and π_3 the penalty for one overbooked unit of resource 3 (OR slot). To find suitable values for the penalties, a linear model was formulated. Let $\vec{u}_r^{j|i} = (u_{1,1}^{j|i}, \ldots, u_{1,N_i}^{j|i})$ be the resource requirement of resource r for patient type i with resource usage case j, and $g^{j|i}$ be the contribution received from resource case j,where $i \in \mathcal{I}, j \in \{1, \ldots, J_i\}$. The minimum penalties for each resource can be found by solving:

Resource	Penalty units
Regular beds	712
ICU beds	194
OR slots	318

Table 6.8: *Penalties for all resources used in simulation for all heuristics.*

$$\begin{aligned} \min \quad & \sum_{r=1}^{R} \pi_r \\ \text{s.t.} \quad & \sum_{r=1}^{R} \pi_r \sum_{j=1}^{J_i} p_{j|i} u_r^{j|i} \geq \sum_{j=1}^{J_i} p_{j|i} g^{j|i}, \qquad \forall r \in \{1, \ldots, R\}, i \in \mathcal{I}, \\ & \pi_r \geq 0, \qquad \forall r\{1, \ldots, R\}. \end{aligned} \tag{6.5}$$

The program is a simple linear program and was solved in AMPL. The results, the penalty for overbooking one unit of each resource, can be seen in table 6.8.

6.11 Identification of Censored Data

There are no indications in the hospitals records when the hospital has more elective patient demand than it is available to admit. Therefore assumptions had to be made in order to find which observed data points are censored and which ones are not.

To find the combined resource usage in the data period - The total number of days each patient stayed in the hospital was given, along with the number of days they spent in the ICU unit. However, there was no indication of when during the stay the patient is in the ICU unit. Thus, assumptions had to be made

We have information the number of days each patients stays in ICU, but not on which days of the total stay the patient was in the ICU unit. Therefore the following assumptions had to be made about the ICU stay:

a. Patients start ICU bed stay on the day they have surgery. If a patient does not have surgery on the first day of admission they stay in a regular bed until the day of surgery.

b. If patient is in an ICU bed for a longer period than the period from surgery day to discharge day, the ICU stay is at the end of a patient's total stay in the hospital.

c. If a patient does not have surgery the ICU bed stay begins on day of admission and is later transferred to a regular bed.

Using this we were able to find the number of beds taken for each day in the given period. The data has records of when the surgery is carried out and from that we found the data points that are censored because of fully booked ORs.

We assume that the hospital rejects admission when one or more of the following conditions hold; i) all beds in the ICU unit are taken, ii) all beds in unit with regular beds are taken, or iii) all ORs are fully booked. When this happens it is assumed that the elective demand was greater than or equal to elective admissions, and those are the censored data points.

CHAPTER 7

Remarks

The estimation of patient arrival distribution is limited to the fact that the only data given are the number of patients admitted per day. The rate of patient arrivals might change over the day, and thus, it might be interesting to look at the patient arrivals by time periods shorter than a day. This would also influence how the censored data is transformed.

The results from this work shows that there is a vast difference in the hospital's reward between using the different heuristics. It is enjoyable to see that the method Barz and Rajaram (2012) develop gives the highest reward for version A of the heuristic. The main difference between the simulation here and the approach they take is the estimation of patients' future resource usage. They use the patient's state at the given time to find the expected value of the future resource usage, and thus the expected value is updated every time the patient moves into another state. Here however, the expected resource usage is only found at the day of patient's arrival, and not updated again. It might be better to use the patient states the same way as Barz and Rajaram do.

From the results of the simulation, it is definitely recommended to reserve some amount of the hospital's resources for emergency patients. The recommended proportion of reservation is not necessarily the same for all resources, and it should be analyzed how many units of each resource should be reserved separately. Care must be taken to not reserve too many resource units for emergency patients, as that might result in too few patient admissions.

REFERENCES

Barz, C., Rajaram, K. 2012. Elective Admissions under Multiple Resource Constraints. Working paper.

Bekker, R., Koeleman P. M. 2011. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, **14**, pp. 237249.

Cam, L. 1990. Maximum Likelihood: An Introduction. *International Statistical Review*, **58**(2), pp. 153-171.

Collart, D., Haurie, A. 1976. *IEEE Transactions on Automatic Control*, **21**(2), pp. 233-238.

Dempster , A.P., Laird, N.M., Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), pp. 1-38.

Duda, Catherine. Telephone interview. 6 January 2012. 24 February 2012.

Gerchak, Y., Gupta, D., Henig, M. 1996. Reservation Planning for Elective Surgery Under Uncertain Demand for Emergency Surgery. *Management Science*, **42**(3), pp. 321-344.

Greene, W. 2011. *Econometric Analysis, 7th ed.*, Prentice Hall, New Jersey.

Huh, W.T.,Levi, R., Orlin, J. B., Rusmevichientong, P. 2011. Adaptive Data-Driven Inventory Control with Censored Demand Based on Kaplan-Meier Estimator. *Operations Research*, **59**(4), pp. 929-941.

Kaplan, E. L., Meier, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, pp. 457-481.

Kolesar, P. 1970. A Markovian Model for Hospital Admission Scheduling. *Management Science*, **16**(6), pp. 11-18.

Litvak, N. and van Rijsbergen, M. and Boucherie, R.J. and van Houdenhoven, M. 2008. Managing the overflow of intensive care patients. *European journal of*

operational research, **185**(3). pp. 998-1010.

McGill, J. I. 1995. Censored regression analysis of multiclass passenger demand data subject to joint capacity constraints. *Operations Research*, **60**, pp. 209-240.

McLachlan, G., Krishnan, T. 1997. *The EM-algorithm and Extensions*. Wiley, New York.

Ross, S.M. 2004. *Introduction to Probability and Statistics for Engineers and Scientists, 3rd ed.*, Elsevier Inc., USA.

Ross, SM. 2007. *Introduction to Probability Models, 9th ed.*, Elsevier Inc., USA.

Ross, S.M. 1996. *Stochastic Processes, 2nd ed.*, John Wiley & Sons, Inc., New York.

Ryzin, G.V., McGill, J. 2000. Revenue Management without forecasting of optimization: An adaptive algorithm for determining airline seat protection levels. *Management Science*, **46**(6). pp. 760-775.

Swartsman, G. 1970. The patient Arrival Process in Hospitals: Statistical Analysis. *Health Services Research*, **5**(4), pp. 320-329.

Stefanescu, C. 2012. Multivariate Demand: Modeling and Estimation from Censored Sales. Working Paper.