University Medical Center Utrecht
University Of Twente

Bachelor Thesis Industrial Engineering and Management

# Determining Bed Capacity using Mathematical Optimization

*Author*
Hayo Bos
h.bos@tilburguniversity.edu

*Supervisors*
Dr.ir.A.G. Maan-Leeftink(UT)
Dr.D.Demirtas (UT)
Drs. R.H.A. Maas (UMCU)

November 22, 2018

# Contents

# Preface

The report in front of you concludes my bachelor assignment which I conducted at Neurology and Neurosurgery department of the University Medical Center of Utrecht. It also concludes my bachelor study Industrial Engineering and Management and therewith my period at the University of Twente. It is the result of five months of hard work, in which I have learned how it is to be working for a large, complex organization with people who look from a totally different angle to certain subjects as I do. I have experienced my time at UMCU as fun and a valuable addition to my track record.

I want to thank both my supervisors Gréanne and Derya for the feedback they have provided, their flexibility and their ideas to further improve my thesis. Besides that, I want to thank Rob and Janet for their approachability and willingness to help me when I got stuck. I want to thank Michel for his comments and readiness to be a sparring partner when needed and my boyfriend Quinten for his feedback and support during the writing process. Lastly, I want to thank my sister Nienke for her inspiration and time to make my thesis to what it has become.

I hope you will enjoy reading my thesis.

Hayo Bos, November 22, 2018.

# Management Summary

## Introduction and Problem Analysis

University Medical Center of Utrecht's (UMCU) brain division houses three clinical wards and two Medium Care Units (MCUs). Two of these clinical wards and both MCUs are central in this research. In these wards there is too much variability in bed utilization, and too many patients are being declined due to a fully occupied ward. Besides that, a lot of incident reports are due to communication errors between the professionals of the different wards. To solve these problems, the management has decided to merge both MCUs and clinical wards. This intervention alone would not solve all problems. After a thorough analysis, we have discovered that the management lacks understanding of the (quantitative) relation between the bed occupancy and bed blocking probability. In order to solve the problems mentioned, we thus need to provide information on the optimal bed capacity and its relation with relevant Key Performance Indicators (KPIs). The central problem is therefore formulated as:

*"What is the optimal number of beds for the new wards, taking the relevant KPIs into account?"*

## Method

The foundation of this research is a thorough data analysis. Using UMCU's data cube, we investigated the arrival and length of stay data of the year 2017 for all four wards. With the results of the data analysis, we derived the current performance of the wards with respect to bed occupancy and blocking probability. According to the literature, the blocking probability should be the central KPI when determining the bed capacity. The target blocking probability for the new clinical ward is set to 5%, and for the MCU to 3%. The current performance of the clinical wards is 3.4 % (D340&D370) and 10.9% (D350). For the MCUs this is 7.5 % (D351) and 23.5% (D361). To solve the central problem, we apply a variety of mathematical optimization techniques. We use queueing theory to derive the relation between the blocking probability and bed occupancy for both the new MCU and clinical ward. Furthermore, we use mathematical programming to alter the elective arrivals of the clinical ward such that the variability is reduced, and use simulation modelling to see the possible gain of this intervention.

## Results

Using the models of the previous section, we calculated the relation between the blocking probability and bed occupancy given a number of beds. This relation is pictured in Figure 1.

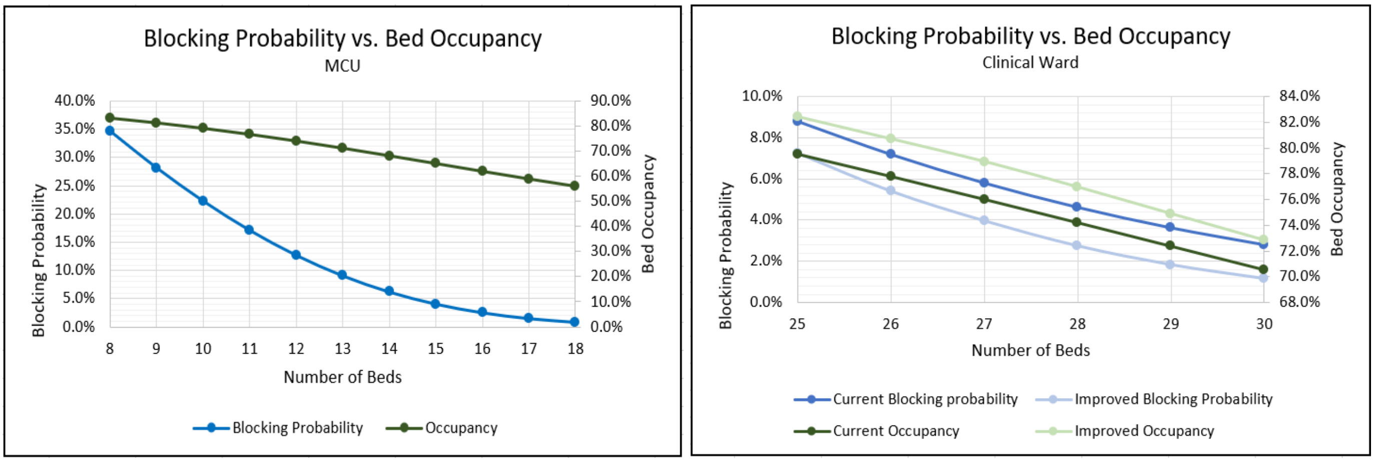


Figure 1.: The relation between the blocking probability and bed occupancy for a given number of beds for the MCU (left) and clinical ward (right).

From this figure we can derive that, if we want to meet the target, we need at least 16 beds for the MCU resulting in a bed occupancy of about 62%. For the clinical ward we pictured both the current situation (dark colors) as well as the situation after altering the elective arrivals (light colors). To meet the target blocking probability we now need 28 beds with a bed occupancy of 74%. If however the management is able to stabilize elective arrivals through the week (including the weekend), we need 27 beds for a lower blocking probability resulting in a bed occupancy of about 79%. Furthermore, we remark that for all set-ups altering the elective arrivals results in a higher bed occupancy and lower blocking probability.

## Recommendations

We have translated the described analysis into workable recommendations which are presented below.

• We recommend to base the number of required beds for the new ward and MCU on a quantitative method as ours rather than solely using a target occupancy and averages. Given the desired rejection rate of 3% for the MCU and 5% for the clinical ward, we advise a bed capacity of 16 and 28 beds respectively.

• The variability in elective arrivals should be reduced as much as possible. Ideally, the number of elective arrivals should be equal per day per week.

• Capacity decisions should be based on the entire distribution of the relevant variables rather than averages (to avoid the well-known "the flaw of averages"). This means that UMCU should use decision models (e.g., from Operations Research) for capacity problems, that use the underlying variability of the data instead of simple rules of thumb that only use averages.
• Improvements should be made regarding the available data. The number of rejections should be recorded. This will make future capacity decisions more reliable. Furthermore, data should be entered more accurately. It could however be wise to explain to the employees why correct data matters, and that (capacity) decisions based on correct data can actually reduce the workload. Besides that, a possible solution is to automatize registration of e.g., admission and discharge time.

# Definitions, Abbreviations and Notation

## Definitions

• **Clinical ward**: A room or group of rooms with beds to provide regular care to patients.
• **MCU/Step Down Unit**: Medium Care Unit; a special ward providing care of a level between the Intensive Care and the clinical ward. Within the Neurology & Neurosurgery department, the MCU also has a stroke unit for emergency patients.
• **Bed blocking probability / Probability of refusal**: The probability that a patient can not be treated where he/she should be treated because all beds are occupied.
• **The management**: The main group of stakeholders within UMCU who are entitled to make the decisions, for example about bed capacity, the decision for which this thesis gives advice.

## Abbreviations

• **ALOS**: Average Length Of Stay
• **CI**: Confidence Interval
• **CV**: Coefficient of Variation
• **ICU**: Intensive Care Unit
• **KPI**: Key Performance Indicator
• **LOS**: Length of Stay
• **MCU**: Medium Care Unit
• **MOL**: Modified Offered Load
• **MSER**: Marginal Standard Error Rule
• **N&N**:Neurology and Neurosurgery
• **OR**: Operations Research
• **PCIR**: Patient Care Incident Report
• **SMEs**: Subject Matter Experts
• **UMCU**: University Medical centre of Utrecht

# Notation

| Symbol | Definition |
|---|---|
| $n$ | The number of replications in a simulation study. |
| $\bar{X}$ | Sample mean. |
| $S$ | Sample standard deviation. |
| $\tau$ | Lexis ratio. |
| $k, K$ | Subgroup and total number of subgroups respectively. |
| $t, T$ | Moment in time (e.g., day of the week) and total time span under consideration respectively. |
| $s, c$ | Number of beds under considerations, and maximum number of beds respectively. |
| $B_t, \bar{B}$ | Blocking probability on $t$, average blocking probability. |
| $\mu_k$ | Parameter of hyperexponential distribution, $\mu_i^{-1}$ is the mean of subgroup $k$. |
| $p_k$ | Parameter of hyperexponential distribution, $p_k$ is the fractional size of subgroup $k$. |
| $\lambda_t$ | Arrival rate on time $t$. |
| $\Lambda$ | Total arrival rate over $T$. |
| $m^k(t)$ | Offered load (i.e., patients present) of subgroup $k$ on time $t$. |
| $m(t)$ | Total offered load on time $t$. |
| $m^*(t)$ | Target load on t. |
| $c_t^i$ | Constant. |
| $\bar{X}$ | Sample mean. |
| $S$ | Sample standard deviation. |
| $t_{n-1,1-1/2\alpha}$ | The Student-t distribution with $n-1$ degrees of freedom, and a significance level of $\alpha$. |

Table 1.: Definitions of mathematical symbols.

# 1. Context Analysis and Problem Formulation

In this chapter we describe the problem statement. We use a problem cluster by which we analyse what the core problem is, which will be the common thread throughout thesis.

## 1.1. The University Medical centre of Utrecht

University Medical centre of Utrecht (UMCU) is one of the eight university medical centres of the Netherlands. UMCU employs more than eleven thousand people, and provides care to more than thirty thousand patients a year. Within UMCU, there are twelve divisions corresponding to their function. One of these divisions is the brain division. The brain division is divided into four departments: Psychiatry, Neurology and Neurosurgery, Rehabilitation Physiotherapy Science & Sport and Translational Neuroscience. This assignment takes place in the Neurology and Neurosurgery department (N&N). The N&N department is divided into seven care units which groups patients according to their clinical status of diagnoses, see Figure 1.1.

### 1.1.1. Wards

Within the N&N department, there are three clinical wards and two Medium-Care Units (MCUs). For this assignment, two clinical wards and both MCUs are considered. These wards correspond to one or multiple cost centre(s)[1]. The distribution of care units over the different wards and the wards' cost centres are given in Table 1.1.

### 1.1.2. Merger

The management of the N&N department has encountered several problems. There is a deficit of 2.3 million euros, nurses experience a lot of variability in bed demand and patients are being refused due to lack of beds at certain moments. Furthermore, there are a lot of Patient Care Incident Reports (PCIR)[2] regarding communication between (medical) professionals. As an attempt to solve these problems, the management has decided to merge two of its wards (C3 East and D3 West) and join its MCUs. The required capacity is yet unknown, and is one of the reasons for this research.

[1] kostenplaats
[2] Meldingen Incidenten Patintenzorg

| | **Cost Centre** | **Care Units** |
|---|---|---|
| **Clinical Ward C3 East** | D340&D370 | General Neurology (GN)<br>Neuro Muscular Diseases (NMD)<br>Functional Neurosurgery and Epilepsy (FNE) |
| **Clinical Ward D3 West** | D350 | General Acute Neurosurgery (GAN) |
| **MCU C3 West** | D361 | Cerebro Vascular Diseases (CVD) |
| **MCU D3 West** | D351 | Primarily General Acute Neurosurgery |

Table 1.1.: The link between the wards, the care units and the cost centres of the Brain division.

The relation between and context of the mentioned problems is further investigated within the next section.

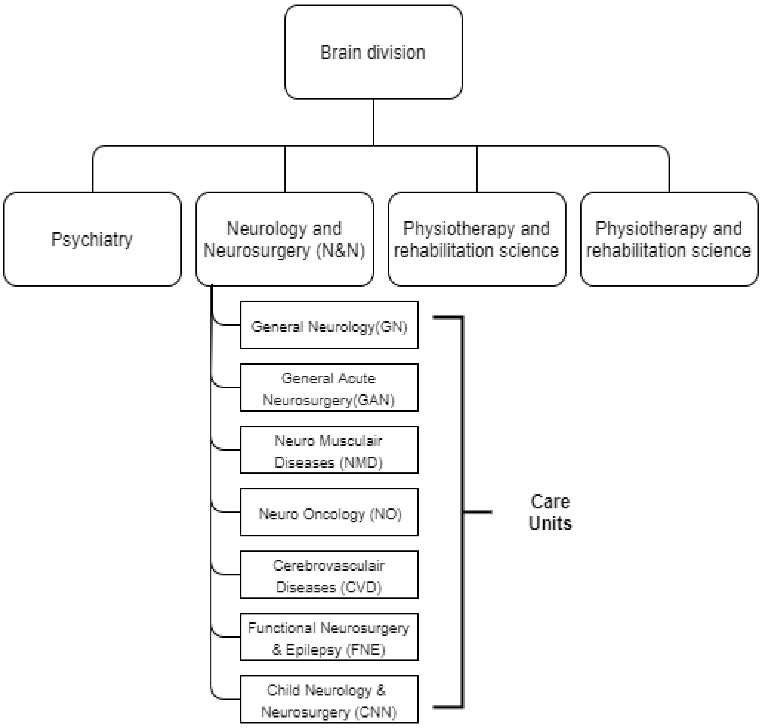


Figure 1.1.: The organigram of the Brain division of UMCU.

## 1.2. Problem Definition

This section explains the managerial problems that triggered this research. Subsequently, it will be clarified how these problems lead us to the core problem. The two management problems that are central in this assignment are

*"There are patients who are declined or placed within another division"* -and-
*"There is too much variability in bed demand"*

These problems signaled the management that the care giving process is not efficient enough. Since the causes of these managerial problems are yet unclear, we need to distract the core problem. To do so, we will use the method provided of Heerkens and Van Winden (2011:44-50). Roughly this method consists of analyzing the problems' context by means of a problem cluster, and then select the core problem based on several criteria. The graphical illustration of the problem cluster can be found in Appendix A. The explanation of this figure is given in the text below. To start, we analyse the management problems and subsequently how they share partly the same cause, which happens to be the core problem of this assignment.

### 1.2.1. Rejection of Patients

The fact that patients are declined is caused by having no unoccupied beds which is in turn caused by the fact that on a certain moment, the number of beds is too low. This does not necessarily mean that the number of beds should be higher, since the bed demand fluctuates over time. However, to correctly determine the number of beds needed good understanding of the relation between the fraction of patients that are declined and the bed utilization is required. The management has indicated that they lack this knowledge. For instance, the management does know that it does not want too many patients to be refused but they can not tell the desired and current number of refusals. After all, having no declined patients is not reasonable due to the stochastic nature of patient arrivals. The above boils down to an absence of knowledge about the trade-off between bed utilization and patient refusals. Due to this lack of insight, it could be that the number of beds is too high because the number of rejections is actually (very) low. This would lead to overcapacity which is a waste of money. This in turn leads to an increase of the deficit. All in all, the management needs understanding of the bed utilization and capacity and its relation with patient refusals, because now it is likely that the number of beds is either too high or too low.

### 1.2.2. Variability in Bed Demand

The experienced variability of bed demand makes it hard to determine the right number of nurses, and therefore the department is likely to have either too many or too few nurses. The former is unwanted because then the hospital wastes money, and the latter because having too few nurses increases their work pressure which reduces the quality of care (Lang et al.,2004).

We observed three causes of the variability of bed demand. The first is that admissions take place before discharges. This causes peak demand for beds, which increases the variability. The second is that the current wards and MCUs are too small. This is an additional reason that led to the decision to merge both wards and MCUs. The third reason is that it is not entirely clear how to measure bed occupancy and it means. One can thus say that the management lacks knowledge about bed capacity management. There is also no knowledge about which type of patients (elective or scheduled) exactly cause the variability in bed demand.

The lack of insight into the bed capacity has in turn several causes. One of them is that there is no registration of beds that are closed that day. This means that the utilization of beds as reported by the information system is much lower than in reality. Another cause is that there is variability in the arrival of patients which makes the process complex.

### 1.2.3. The Core Problem

The next step is to select the core problem. We will do so following the guidelines of Heerkens and Van Winden (2011:48). The main idea is to go "back" in the problem cluster to the first problem that can be solved and has no external causes. This leaves us two candidates: "There is no registration of beds that are closed" and "There is a lot of variability in the arrivals of patients". The first is not suitable as this is something the department is already working on. The latter is not suitable because we cannot change the way most patients arrive. Therefore this problem is not really solvable which is why these are not suitable candidates (Heerkens and Van Winden, 2011:48). The variability of elective patients could be decreased when using a better way of scheduling based on the pattern of emergency patients, but the department would first prefer to have better insight in what would be the optimal number of beds (where these patterns probably also play a role). This argument, together with the fact that the first candidate is not suitable, results in moving forward one step in the cluster. Since, as said before, the management is eager to gain this knowledge and it is the first suitable problem in the cluster we decided let this be the core problem. Resembling the above, the core problem is formulated as:

*"The management does not have enough understanding of the different variables that play a role when determining the bed capacity"*

Solving this problem will help the management to decide on the bed capacity for the merged ward and joined MCU.

### 1.2.4. Scope

This section serves to define the scope of this research which is determined using the framework of Hans et al.(2011). A picture of this framework is given in Figure 1.2. The research focuses on determining the bed capacity for the new ward and the MCU. Following the framework this corresponds to "Resource capacity planning" on strategic level. Using the framework, we underpin further scope conditions. First of all, the implementation and evaluation is not part of this assignment. The management has indicated that the advice will (most likely) be taken into account when the rebuilding of the nursing wards and MCUs takes place. This will not be until 2019, which is outside the time window of this assignment. Secondly, the scheduling of personnel is not taken into account. This corresponds well to the framework of Hans et al.(2011), since personnel planning belongs to the "Offline Operational" level. Third, the allocation of care units to wards is assumed to be fixed. The management has indicated that the decision making regarding the assignment of care units to wards has been completed and should not be revised. Fourth, the research most likely will not entail admission scheduling which is planning on the tactical level (Hans et al.,2011). The number of admissions allowed heavily depends on other parts of the chain, like operation room planning. Studying this aspect is not feasible within the given time frame as it is a fairly complex process. The intended deliverable is, as stated in the previous subsection, insight into the bed utilization and its relation with relevant variables given historical data. It will not be a real time tool to plan day-to-day bed capacity.

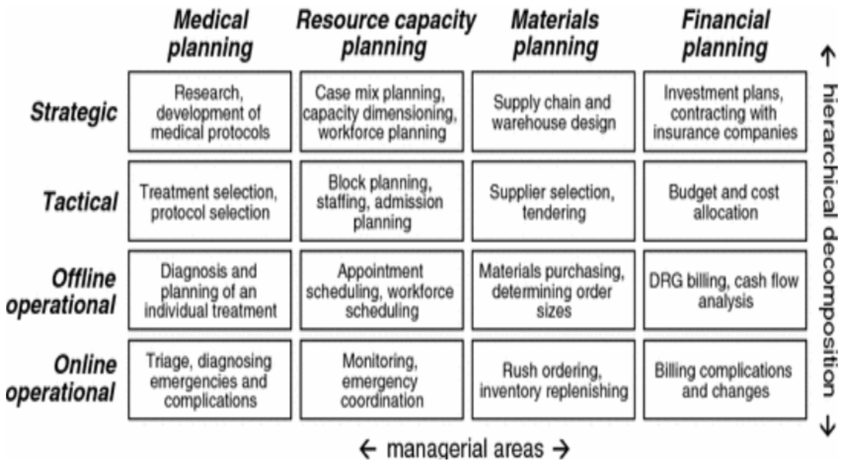


Figure 1.2.: The framework of Hans et al. (2011).

## 1.3. Knowledge problem and Research Questions

### 1.3.1. Research Aim

The aim of this research is to provide understanding of the bed utilization and capacity. After using the results presented in this thesis, the management can underpin their decision on the number beds. Solving the core problem will result in less variability in bed demand and a better understanding of how many patients are declined.

### 1.3.2. Knowledge problem

Solving the problem stated in the previous section requires knowledge. This knowledge will be acquired by following the information strategy when one faces a knowledge problem. The knowledge problem in this assignment is formulated as:

*"What is the optimal number of beds for the new wards taking the relevant KPIs into account?"*

Answering this question will provide the required understanding of the bed capacity and occupancy.

### 1.3.3. Research Questions

To solve the knowledge problem, we break the problem down into smaller sub problems which are presented below research questions. These research questions are the foundation of this thesis.

#### Current situation

First, we have to get insight about the current situation. The management indicated that there is no consensus over how many beds are optimal. A first step to provide insight is to determine what "optimal" in this context exactly means. The first research question serves to shed some light on this:

***1: What is the current performance with respect to bed capacity on both wards and MCUs? (Ch.2)***

***a:*** *What are the most important Key Performance Indicators (KPIs) regarding bed capacity, and how do both bed wards and MCUs perform now?*

***b:*** *What are the desired levels of the KPIs ?*

Then it should investigated what the arrival process looks like. What are the care pathways? Is there a difference between elective and emergency patients? One should also think of the statistical properties of the arrival time- and service time distribution. The next question deals with these issues.

***2: What are the characteristics of the patient related processes? (Ch.2)***

***a:*** *What are the patient care pathways?*

***b:*** *What are the statistical properties of the arrival and service process according to historical data?*

***3. What are the characteristics of the clinical ward and MCU after the merger? (Ch.3)***

#### Literature

To prevent ourselves from reinventing the wheel, we look at work from other researchers about this topic. The one question we answer in this section is:

***4: How to correctly determine the right number of beds on wards and MCUs according to the literature?(Ch.4)***

### Desired situation

To apply the theory and models, we will try to estimate their performance. To do so, we will test different configurations. Thereafter we will summarize the findings of this research with the final research question. Therefore, the last two research questions are:

***5: What is the quantitative relation between the relevant KPIs when determining the required bed capacity? (Ch.5)***

**a:** *What are the scenarios to test?*
**b:** *How do these scenarios perform?*

***6: What are the practical insights gained from this research? (Ch.6)***

# 2. Analysis of the Current Situation

In this section we investigate the current situation on the discussed wards and MCUs. First, the patient flows towards and in between the wards and MCUs will be discussed. Then, relevant concepts and Key Performance Indicators (KPIs) will be defined. Subsequently we will use UMCU's database, called the data cube hereafter, to analyse the arrival process in more depth and develop understanding about the length of stay of the patients. The descriptives of the data used from the data cube can be found in the caption of the relevant figures and tables.

## 2.1. Patient Flows and Context

This section serves to shed light on the various patient flows. This will be done by considering three dimensions: the type of admissions, whether an admission is planned or not and whether a patient is assigned to the proper ward. Furthermore the arrival process is discussed. As a last subject, the typical patients and ward characteristics are discussed.

### 2.1.1. Type of Admissions

Patient admissions can be split up into new and transfer arrivals. Transfer arrivals encompass patients that are admitted on the concerning ward after an admission on another ward. Suppose a patient is admitted on the Intensive Care Unit (ICU) after receiving surgery and after some time (e.g., when the patient's physical condition is more stable) the patient is transferred to the clinical ward. In this example, the new arrival is registered at the ICU, and the transfer arrival is registered at the clinical ward. Both the clinical wards and MCUs have both type of admissions. Moreover, part of the transfer admissions of the wards are caused by transfers from the MCUs to the clinical ward and vice versa. In the sequel, both transfer- and external arrivals are considered.

### 2.1.2. Elective and Emergency Patients

Another common way to distinguish arrivals of patients is to consider whether their admission was planned or not. These are oftentimes called elective and emergency arrivals. This stands for a scheduled and unscheduled admission respectively. Examples of scheduled admissions are day treatments for certain neuro muscular diseases (e.g., Multifocal Motor Neuropathy) or a sleep deprivation in case of epilepsy. An example of an emergency patient is an arriving patient with head trauma caused by a car accident. The data cube keeps track of this distinction for both new and transfer admissions.

### 2.1.3. Diverted Patients

A third way to look at patient inflow, is to consider whether the ward of admission is the one where the patient should be admitted. Consider the patient with head trauma again. After arrival, the patient needs observation of the neurologist. Ideally, the patient is placed on the D340&D370 ward. Now suppose that this ward is occupied, and that the patient is diverted to the D350 ward, or even worse a ward in another division or hospital. The patient is then admitted on a “wrong” bed. The number of patients on wrong beds should be kept as small as possible both from a cost and a quality of care point of view. This means that a ward should not be fully occupied when a patient arrives.

### 2.1.4. The Arrival Process

Before admission, a patient goes through several steps. Most of the emergency patients arrive at the hospital via the Emergency Department (ED) or the outpatient clinic[1]. If a patient needs to be admitted, a doctor of the regarding department is contacted. This doctor in turn contacts the coordinating nurse who tries to find a free bed on the correct clinical ward or if this is not possible on another ward within the department. If it turns out that all of the above mentioned beds are occupied UMCU’s bed coordinator tries to find a bed somewhere else within the hospital. If this also fails, the patient is redirected to another hospital. Elective patients are directly admitted on the ward of destination. An elective patient makes an appointment and the staff reserves a bed. If, due to an unforeseen event (such as an unusual number of emergency arrivals) there is no bed available anymore, the elective patient is canceled.

### 2.1.5. The Clinical wards

As mentioned in Chapter 1 there are two clinical wards under consideration. The maximum bed capacity on these wards is 18 beds for the D340&D370 ward and 12 beds for the D350 ward. Most of the days however, some of these beds are “closed”. As said before, the number of beds that is closed per day is not available in the data cube. This decision is made every morning when a group of nurses discuss the expected arrivals, discharges and transfers of that day. The beds are distributed over rooms, which are in turn located close to each other. A typical aspect of the first ward is the relatively high number of short stays. This is caused by the fact that there are many day treatments on this ward. The second ward has many (neuro) surgical patients which are e.g., admitted on the ward, receive surgery and subsequently are transferred back. In the desired future situation these two wards become one ward.

### 2.1.6. The Medium Care Units

Besides the clinical wards, also two Medium Care Units (MCUs) are under study: the D361 MCU and the D351 MCU. The MCUs have a capacity of 7 and 6 beds respectively.

[1] Polikliniek

Noteworthy is that the first MCU is a stroke unit, which consists of beds reserved for patients with a stroke. Because the level of care is higher than clinical wards (hence the prefix medium), there are more nurses per patient. Similar to the clinical wards, the number of beds in use varies per day. Similar to the clinical wards, the MCUs will be combined.

## 2.2. Performance Indicators

There are many indicators for performance in healthcare regarding bed capacity. A selection of them is used in this thesis, on which we elaborate below. For each KPI we first explain what is measured and then how it is calculated.

### 2.2.1. Length Of Stay

The length of stay (LOS) is the time that passes between the moment that a patient is assigned to a hospital bed on a ward and the moment that patient is discharged. It is important to keep in mind that discharge in our context could also mean the transfer to e.g., the operation room, the MCU, the ICU etc.,contrary to the commonly used definition of discharge which is to be discharged from the hospital. To measure the LOS per patient, we do consider transfers. For example, suppose that a patient stays 3 days on Ward A and 4 days on Ward B. According to our intuition, this results in two LOS data points, being 3 days for ward A and 4 days for ward B (contrary to summing them). The LOS is an important KPI because it can be used used in the calculation of bed occupancy as we will see later, and thus is related with decisions regarding the number of beds. The financial controller of the department calculates the average length of stay (ALOS) as:

$$ALOS = \frac{Total\ LOS}{Number\ Of\ New\ Admissions} \tag{2.1}$$

This overestimates the “real” LOS because the transfer patients are not taken into account. UMCU’s Business Intelligence division also provides information on the LOS. They do take the transfers into account, but again only provide averages. It is well known that (capacity) decisions in general should not be based purely on averages (also known as “The flaw of averages”) but rather on the underlying distribution. This will be further investigated in Section 2.4. The equation used to determine the ALOS in the sequel of this thesis is:

$$ALOS = \frac{Total\ LOS}{Number\ Of\ New\ Admissions\ +\ Transfer\ Admissions} \tag{2.2}$$

### 2.2.2. Rejection Rate

Another interesting KPI is the rejection rate, i.e., the percentage of patients that cannot be admitted due to a lack of capacity. This occurs when a patient arrives and finds all beds occupied. The rejection rate is defined as follows:

$$Rejection\ Rate = \frac{Number\ of\ Rejected\ Patients}{Total\ Number\ of\ Arrivals} \tag{2.3}$$

This is an important KPI in bed capacity decisions. Unfortunately, this KPI is not (yet) being monitored because it is hard to keep track of the number of patients that is rejected. We should therefore analyse the data to calculate the current performance.There is no log of patients that are rejected. The same goes for the total number of arrivals, because only the *admissions* are being registered (i.e., the number of patient that do not find the ward fully occupied). The best thing that can be done is to estimate the current performance by estimating the fraction of time the wards are fully occupied.

### 2.2.3. Bed Occupancy

The last KPI that will be discussed is the bed occupancy. This is probably the hardest KPI to define, because there is no consensus both in the literature and at UMCU on how to calculate this ratio. It is an important KPI because in many hospitals it plays a central role in determining the number of beds needed (which in fact is strongly advised against by e.g., Green (2002)). Intuitively, one would say that bed occupancy is calculated as the capacity in use divided by the total capacity available (as one would usually define occupancy or utilization). Following this reasoning, the occupancy then can be defined as:

$$Bed\ Occupancy_t = \frac{Number\ of\ Patients\ Present_t}{Number\ of\ Beds\ Available_t} * 100\% \tag{2.4}$$

Where the $t$ indicates the time at which we evaluate the KPI (e.g., hour of the day). Following this definition, two difficulties arise. The first is that it is hard to measure the number of patients present at a given time. Time is a continuous parameter, and giving a precise measurement of bed occupancy would require continuous measurements which are not available in the data cube. The data cube does provide data on the number of patients present per 15 minutes. This data however is highly sensitive for input errors. An example of an input error is when a patient is already discharged but not registered as such, because the person in charge of this first had other things to do. Besides that, sometimes a patient is not physically present whereas he is registered as such. This can be the case when someone has completed his take-in for surgery, and can wait the days at home. The system then registers the patient as being present, whereas he does not occupy a bed. Reasoning this, the bed occupancy can be higher than 100%. The second difficulty is that beds are some days "closed" but for the previous years this has not been registered. It should be noted that the department is working on this, and does monitor the beds that are open/closed since a couple of months. The previous analysis underpins the need for another, more objective, estimator for the bed occupancy.

A well known result often used in Operations Research is Little's Law (Little, 1961). It is formulated as follows:

$$L = \lambda W \tag{2.5}$$

With $L$ being the number of customers present, $\lambda$ being the arrival rate per time unit and $W$ being the average length of stay. This equation is used in the same context as ours by e.g., De Bruin et al. (2009) and Cochran and Roche (2008). We could define $\lambda$ as the arrival rate of patients per day and $L$ as the number of patients present in steady state. Define $c$ as the average open beds on day. Combining equation 2.4 and 2.5 then results in:

$$Bed\ Occupancy = \frac{\lambda ALOS}{c} \tag{2.6}$$

## 2.3. Arrivals

In this section the patient arrival process will be analysed in depth. We do so by splitting the admissions into emergency and elective admissions (see Section 2.4). This is a useful distinction because especially unexpected arrivals (i.e., emergency arrivals) tend to be well described by a Poisson process (Young, 1965). The Poisson process will be explained in more depth below. Furthermore De Bruin et al. (2009) discovered that in a hospital situation also the planned arrivals (i.e., elective arrivals) are well described by a Poisson process. First the Poisson process will be explained as it plays a central role in this section. Subsequently we look at the arrival patterns of patients per ward (both elective and emergency), at the arrivals per day of the week and finally at the arrivals per hour of the day. It should be noted that the D351 ward was closed the first few weeks of 2017. This has been taken account for during all further analyses in the sequel of this thesis.

### 2.3.1. Poisson Process

Let $N(t)$ be the number of arrivals up to time $t \geq 0$ (so $N(3)$ is the number of arrivals up to and including $t = 3$). If the following (mild) assumptions hold, number of arrivals in a specified interval $N(t + s) - N(t)$ tend to have several interesting properties.
**Assumptions:**
1. The probability that 2 or more customers (patients) arrive at exactly the same time is 0.
2. The number of arrivals in non-overlapping time intervals is independent
3. In case of a stationary Poisson process: the distribution of $N(t+s)-N(t)$ is independent of $t$ (i.e., the probability of 6 arrivals today is equal to the probability of 6 arrivals eight days from now).

If the above assumptions hold, then $N(t + s) - N(t)$ (the number of arrivals in an interval of length $s$) follows a Poisson distribution with parameter $\lambda s = E[N(s)]$ and $\lambda = E[N(1)]$. Moreover, the inter-arrival times (the time between subsequent arrivals) tend to have follow an exponential distribution (Law, 2015:380-384). In other words, using the theory above, we can draw inferences about the arrival patterns. It is not hard to conclude that this property is very useful in capacity related problems.

If the arrival rate is time dependent (i.e., the average number of customers arriving between e.g., 3:00 AM and 5:00 AM differs from the number of customers arriving between 7:00 AM and 09:00 AM), assumption 3 is violated. We then could be dealing with a non-stationary Poisson process (Law, 2015:380-384) (which turns out to be the case in our situation). In this case, we should specify the arrival rate as a function of the time (e.g., day of the week). Techniques for doing this will be discussed in Chapter 4.

### 2.3.2. Arrivals Per Ward

#### Elective Arrivals

To analyse the arrival pattern per ward, we counted the number of emergency and elective arrivals on each day of the year 2017. Both new arrivals and transfers are taken into consideration, since also transfers are patient arrivals. Subsequently, the data was used to calculate summary statistics, and to draw a histogram. A histogram can be seen as a draft of the underlying probability distribution. Because a Poisson distribution (see Appendix B.1) is suspected, this distribution is "fitted" through the data to check whether Poisson arrivals are reasonable on first sight. The estimator of the parameter of the Poisson distribution used is the sample average $\overline{X}$. This is also the Maximum Likelihood Estimator (MLE) and thus has several desirable properties (for more information see Appendix B.4). For both the clinical wards (D340&D370 and D350) the histograms can be found in Figure 2.1. It turns out that the frequency of 0 elective arrivals on a day seems to be a bit high. By visual inspection it is clear that the Poisson model underestimates this frequency and therefore we do not need a statistical test to verify this. The high frequency of 0 arrivals could be caused by the fact that

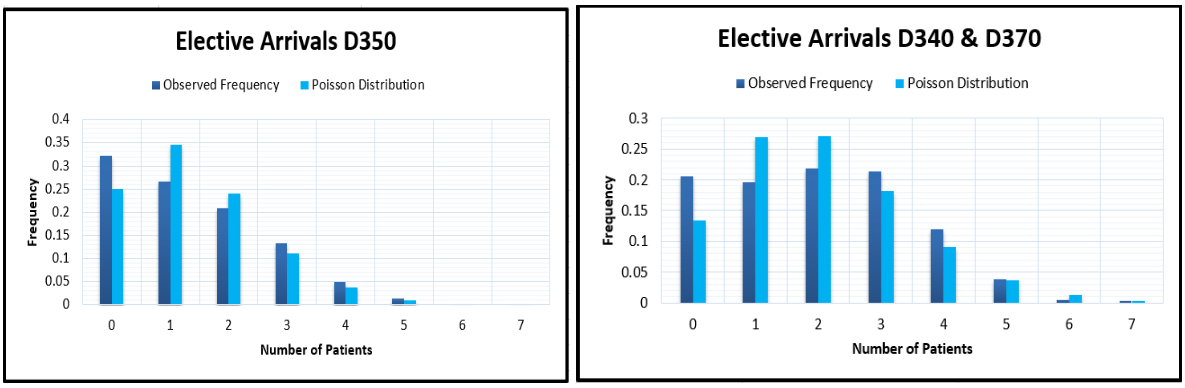


Figure 2.1.: Histogram of the elective arrivals on the clinical wards in 2017. *Data retrieved from data cube, n=363 (days) and n=362 respectively.*

there are (basically) no planned admissions on weekend days and holidays[2]. Splitting the elective arrivals of the clinical wards into week- and weekend arrivals (where holidays on **weekdays** are excluded) gave the histograms of Figure 2.2.

[2] First- and second day of Christmas, Easter and Pentecost, New Years day, Ascension day and the day after, Good Friday and Kingsday

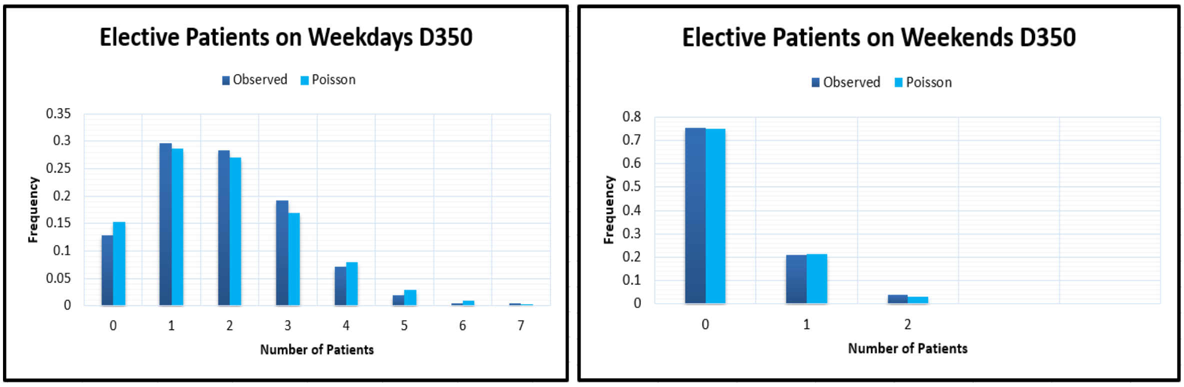


Figure 2.2.: Daily elective arrivals of the ward D350 after being split into week- and weekend days in 2017. *Data retrieved from the data cube, n=250 and n=105 respectively.*

This provides a much better fit (see Table 2.1 for the p-values the $\chi^2$ goodness-of-fit test) for the elective arrivals for the D350 ward. The histograms for elective arrivals split into week- and weekend days for the D340& D370 ward can be found in Appendix C in Figure C.1. For the MCUs, such behavior did not occur. This is confirmed by the fact that the MCUs keep running on the same level (do not scale down which is the case with the clinical wards) during weekends. Furthermore, even though the elective arrivals of the MCUs are labeled "elective", their urgency is often still quite high and therefore admissions during the weekend continue. The histograms for the elective arrivals for the MCUs can be found in the Appendix C in figure C.2.

### Emergency Arrivals

For the emergency arrivals the same analysis has been made, see Figure 2.3. We did not split the arrivals into week and weekend days since emergency arrivals are not likely to change in number during weekends or holidays. This is confirmed by Figure 2.6, where the emergency arrivals of both MCUs are pictured. The emergency arrivals of the clinical wards can be found in the Appendix C in Figure C.3.

### Summary statistics per ward

In Table 2.1 the summary statistics regarding patient arrivals are portrayed. The mean is just the sample mean $\overline{X} = \frac{1}{n}\sum_i x_i$. The standard deviation is an (often) used measure for the "spread" of the data, and is estimated with the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \overline{X})^2}$. For discrete data (as is the case), another useful statistic is the lexis $\tau$ ratio which is the variance divided over the mean, and estimated by $\hat{\tau} = \frac{s^2}{\overline{X}}$. This last statistic is helpful, since it can be used to choose between distributions. $\tau \approx 1$ indicates a Poisson distribution because the variance of this distribution equals the mean (Appendix B.1). For each ward the p-value for the $\chi^2$ Goodness-of-Fit test is given. Note that the

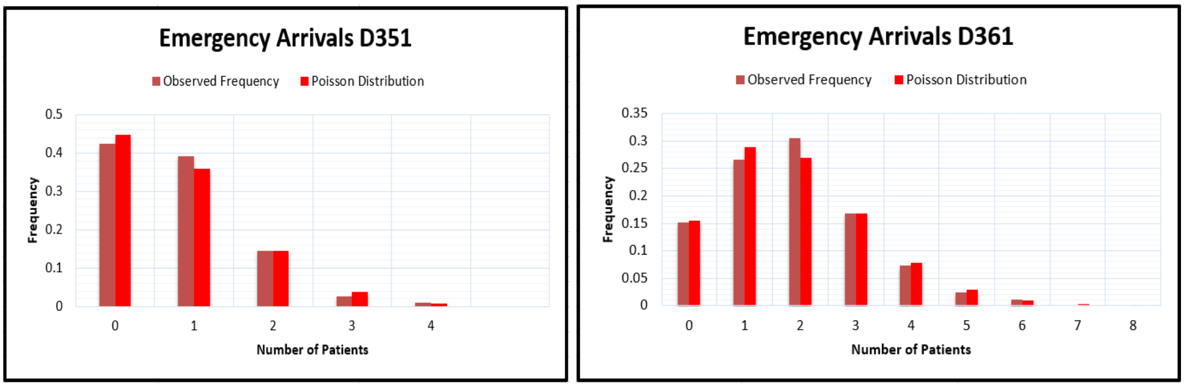


Figure 2.3.: Daily arrivals of emergency patients for both MCUs in 2017. *Data retrieved from the data cube, n=310 and n=357 respectively.*

$H_1$ of this test is to reject the fit. Since rejection occurs for low p-values, we have that p-values $< \alpha = 0.05$ advocate a bad fit. In the table these cases are boldfaced.

| | Mean | Standard Deviation | Lexis | p-value |
|---|---|---|---|---|
| **D361** | 0.36 | 0.54 | 0.81 | 0.48 |
| **D340&D370 Week** | 2.64 | 1.32 | 0.66 | **0.00** |
| **D340&D370 Weekend** | 0.62 | 0.85 | 1.19 | 0.92 |
| **D350 Week** | 1.88 | 1.27 | 0.86 | 0.66 |
| **D350 Weekend** | 0.29 | 0.53 | 0.99 | 0.98 |
| **D351** | 0.31 | 0.51 | 0.82 | 0.49 |

Table 2.1.: Summary statistics of the arrivals of elective patients, year 2017.

With respect to the elective patients, we see that the lexis ratio varies between 0.81 and 1.19. The Poisson distribution "passes" the $\chi^2$ test for all wards except the arrivals for the D340&D370 ward. However, by visual inspection of the observation vs. the Poisson fit we see that the variance is over all well captured by the Poisson fit. Especially in large samples, the $H_1$ is quite easily accepted (i.e., not much deviation is required). Combining the above, we conclude that a Poisson distribution is an appropriate fit for all wards. The summary statistics w.r.t. emergency patients are given in Table 2.2.

| | Mean | Standard Deviation | Lexis | p-value |
|---|---|---|---|---|
| **D361** | 1.87 | 1.31 | 0.92 | 0.74 |
| **D340&D370** | 1.69 | 1.56 | 1.43 | **0.00** |
| **D350** | 1.02 | 1.00 | 0.98 | 0.81 |
| **D351** | 0.80 | 0.85 | 0.90 | 0.52 |

Table 2.2.: Summary statistics of the arrivals of emergency patients, year 2017.

With respect to the emergency arrivals, the Poisson model provides a good fit for all wards except the D340& D370 ward (like in the elective case). By visual inspection we see that especially the frequency of days with 0 elective patients is “too high”. However, especially for emergency arrivals, the Poisson process assumptions oftentimes hold. Moreover, again the variance of the observations is well captured by the Poisson model. Therefore, albeit the fact that the fit is not perfect, we still assume a Poisson distribution for this ward. Hence the overall conclusion is that the Poisson distribution provides a suitable fit for all wards regarding emergency arrivals.

### 2.3.3. Weekly Patterns

#### Elective Patients

The previous analysis gives the underlying distribution for the arrivals per day. It could be that this pattern is suitable for all days of the week. However, it could very well be the case that certain patients do not arrive during certain days. One can think of the absence of planned patients during weekends. This subsection serves to investigate this.

In Figure 2.4 the number of elective arrivals in the year 2017 can be found per day for the wards D340 & D370 and D350.

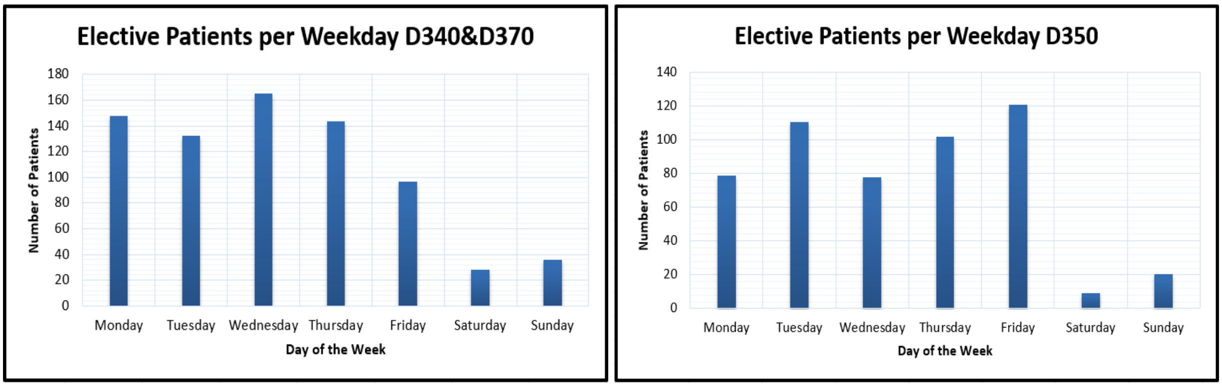


Figure 2.4.: Weekly pattern of elective patients for the clinical wards D340& D370, and D350 in 2017. *Data retrieved from data cube, n=729 and n=500 respectively.*

The first thing that strikes is the significantly lower number of elective admissions during weekends on the clinical wards. This is perfectly in line with the expectations of the management that planned patients are not admitted during weekends. Besides that, it underpins the choice to split week from weekend days in the previous subsection.For the D350 ward (the surgical ward), we see furthermore that there is a peak on Friday and a trough on Mondays. This is confirmed by the fact that surgeries need some preparation which is not currently done during the weekend (hence the trough on Monday), and that scheduled admissions during the weekend should be prevented (hence the peak on Friday).

For both MCUs the graphs are given in Figure 2.5. Although there seem to be peaks and troughs for the MCUs, these are not easily explained by the management. Furthermore it should be noted that the total number of elective admissions on these wards is much lower than on the clinical wards, making it more likely that these peaks and troughs are just caused by randomness.

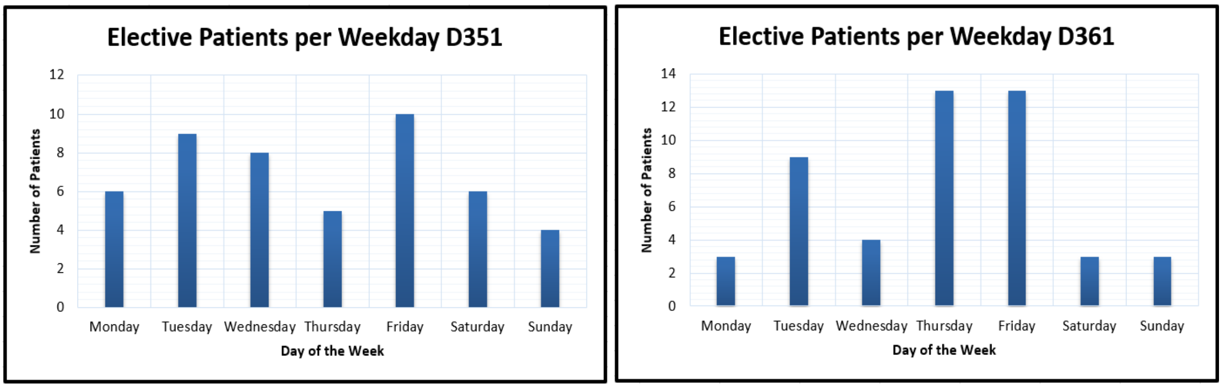


Figure 2.5.: Weekly pattern elective patients for the MCUs D351 and D361 in 2017. *Data retrieved from data cube, n=48 for both wards.*

### Emergency Patients

The same analysis is made for emergency patients. There is no reason to assume that there is such a pattern as with elective patients, because there is no reason to think that there are e.g., fewer stroke cases during weekends. This absence of a pattern is confirmed by the data of all wards. The emergency arrivals per weekday for one of the clinical wards and one MCU are pictured in Figure 2.6. For the other clinical ward and MCU the graphs can be found in the Appendix D and these show similar results.

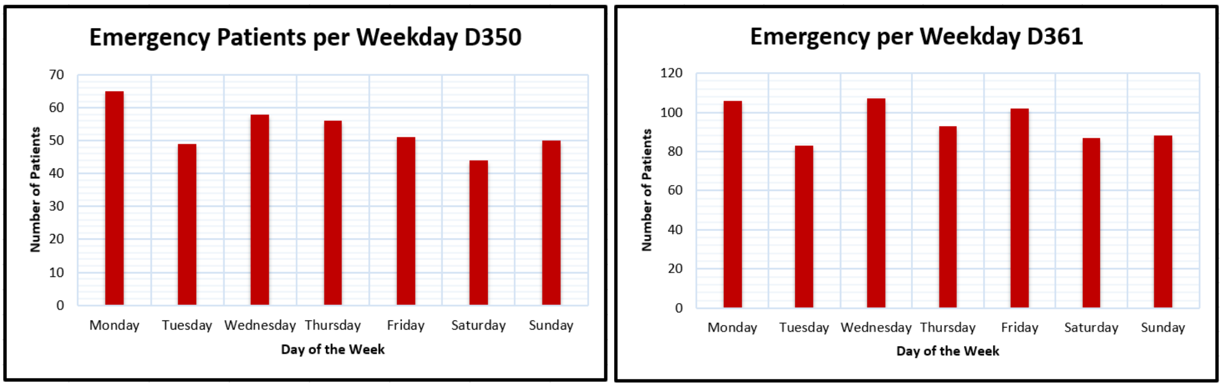


Figure 2.6.: Weekly pattern emergency patients for D350 ward and D361 MCU in 2017. *Data retrieved from data cube, n=373 and n=666 respectively.*

### Conclusion

Both the data and the management tell us that there is a strong indication for a different number of arrivals per weekday for the clinical wards D340&D370 and D350. For the MCUs this is less obvious, and due to the low number of elective arrivals it is hard to draw inferences about this. A summary of the variation in the number of elective arrivals per weekday is given in Figure 2.7.

| Elective Admissions | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| D340&D370 | 20.9% | 19.5% | 24.8% | 20.8% | 14.0% |
| D350 | 15.7% | 22.6% | 16.2% | 20.9% | 24.7% |

Figure 2.7.: Distribution of elective admissions per day of the week in 2017.

Using this section's analysis and the Poisson assumption of Subsection 2.3.1, the arrival rates per day of the week for each ward are given in Table 2.3.

| | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| **D350** | 2.76 | 3.08 | 2.63 | 3.04 | 3.30 | 1.02 | 1.34 |
| **D340&D370** | 4.32 | 3.82 | 5.04 | 4.47 | 4.17 | 2.04 | 2.20 |
| **D351** | 1.21 | 1.11 | 1.11 | 1.15 | 1.32 | 1.06 | 0.84 |
| **D361** | 2.25 | 2.06 | 2.35 | 2.41 | 2.59 | 1.87 | 1.91 |

Table 2.3.: Poisson arrival rates per day of the week per ward, including both elective and emergency admissions. *Data from 2017, retrieved from data cube, n=3219.*

## 2.3.4. Arrivals per hour

### Elective Patients

It could also be the case that arrivals differ per hour of the day. Again, one would expect that this is especially the case for planned admissions. For instance, patients suffering from epilepsy sometimes need to stay awake during for a night for examination purposes. These patients are logically admitted in the evening which could cause a (small) peak in these hours. Another, more significant pattern that one would expect is the absence of planned admissions during the night. For all four wards the hourly elective arrivals of the year 2017 are given in Figure 2.8 and 2.9 (Note: when hours are missing on the x-axis, it means there are no admissions registered on that time in 2017). As expected, the number of arrivals during night is near zero for all wards. Most wards plan their elective admissions between 08:00 AM and 01:00 PM. The huge peak for the D350 ward is confirmed by that ward's staff, because planned surgical patients are oftentimes admitted around 01:30 PM.

### Emergency Patients

For emergency arrivals a similar analysis has been made. Although the differences throughout the day are less significant, there still is some pattern. The number lowest number of patients arrive around 08:00 AM. This number is slowly increasing with midnight as peak hour. After that, the numbers are decreasing again till sunrise. This is not counter intuitive because e.g., when a patient has a stroke, he or she could remark this by not being able to move certain extremities when he or she wakes up (and thus not during the night). There are some strange peaks for some wards around midnight. The management has explained that when the nursing staff starts the night shift around 11:00 PM, they first evaluate the evening shift with the evening nurses. After that, they meet with patients, provide them their medicine etc. When most patients sleep and the nurses got some spare time (around 01:00 AM) they update the admissions of the last two hours. This could cause a peak around 01:00 AM, whereas the admissions are actually spread out more evenly. This tend to happen more with emergency patients, due to their unexpected arrival. The same phenomenon happens in the morning and afternoon. The graphs picturing the hourly emergency arrivals can be found in Appendix E.

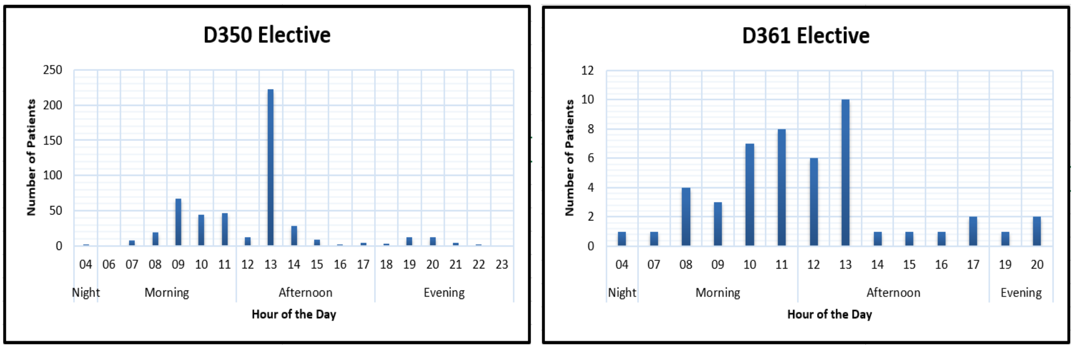


Figure 2.8.: Hourly admissions of elective patients in 2017. *Data retrieved from data cube, n=503 and n=48 respectively*

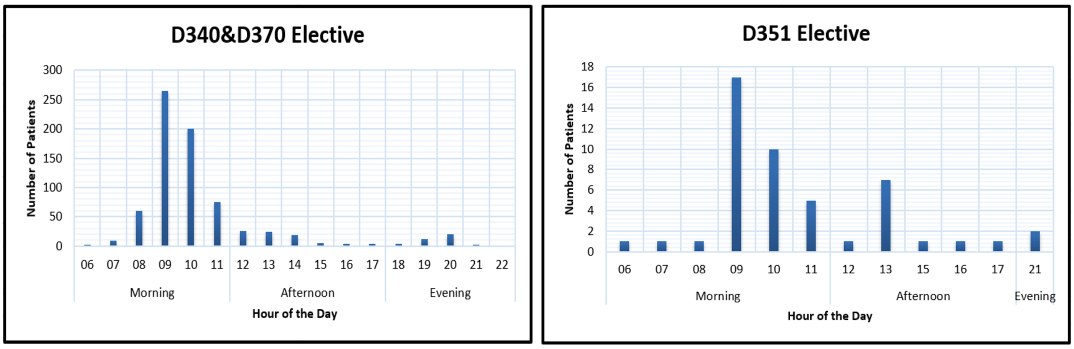


Figure 2.9.: Hourly admissions of elective patients in 2017. *Data retrieved from data cube, n=729 and n=48 respectively*

### Conclusion

Both the elective and emergency admissions do show a difference throughout the day. To quantify this, a heat map has been made which is pictured in Figure 2.10. This

| Emergency | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D351 | 4.84% | 5.65% | 2.02% | 4.44% | 1.21% | 3.23% | 3.23% | 2.02% | 2.02% | 1.61% | 3.63% | 4.03% | 4.03% | 3.63% | 6.45% | 4.03% | 4.44% | 6.85% | 8.06% | 5.24% | 4.44% | 5.24% | 5.24% | 4.44% |
| D361 | 4.20% | 7.51% | 3.45% | 3.30% | 2.25% | 1.65% | 1.95% | 1.65% | 1.05% | 1.50% | 3.60% | 4.20% | 3.60% | 4.80% | 6.46% | 5.71% | 5.41% | 5.56% | 6.61% | 4.95% | 7.36% | 4.65% | 5.26% | 3.30% |
| D340&D370 | 5.68% | 5.19% | 3.90% | 4.22% | 2.76% | 2.11% | 1.62% | 1.95% | 0.97% | 2.11% | 2.60% | 2.44% | 3.25% | 4.71% | 4.87% | 3.73% | 5.52% | 6.17% | 6.17% | 8.28% | 4.55% | 6.33% | 5.19% | 5.68% |
| D350 | 2.14% | 6.97% | 2.95% | 2.68% | 1.34% | 1.88% | 1.88% | 1.61% | 1.34% | 2.68% | 2.95% | 5.09% | 1.61% | 4.29% | 4.29% | 4.56% | 5.36% | 6.17% | 7.51% | 7.51% | 6.97% | 7.24% | 6.43% | 4.56% |

| Elective | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D351 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 2.08% | 2.08% | 2.08% | 35.42% | 20.83% | 10.42% | 2.08% | 14.58% | 0.00% | 2.08% | 2.08% | 2.08% | 0.00% | 0.00% | 0.00% | 4.17% | 0.00% | 0.00% |
| D361 | 0.00% | 0.00% | 0.00% | 0.00% | 2.08% | 0.00% | 0.00% | 2.08% | 8.33% | 6.25% | 14.58% | 16.67% | 12.50% | 20.83% | 2.08% | 2.08% | 2.08% | 4.17% | 0.00% | 2.08% | 4.17% | 0.00% | 0.00% | 0.00% |
| D340&D370 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.27% | 1.23% | 8.23% | 36.35% | 27.43% | 10.29% | 3.43% | 3.29% | 2.61% | 0.69% | 0.55% | 0.41% | 0.41% | 1.65% | 2.74% | 0.27% | 0.14% | 0.00% |
| D350 | 0.00% | 0.00% | 0.00% | 0.00% | 0.40% | 0.00% | 0.20% | 1.59% | 3.78% | 13.32% | 8.95% | 9.34% | 2.39% | 44.14% | 5.57% | 1.79% | 0.40% | 0.99% | 0.60% | 2.58% | 2.39% | 0.99% | 0.40% | 0.20% |

Figure 2.10.: Distribution of elective and emergency arrivals per hour of the day in 2017.

heat map can be used to determine the busy moments on a day, i.e., the moments with a lot of admissions. This is of course useful for answering the main question of this research, but can also help with e.g., nurse scheduling. Emergency patients tend to arrive between 1:00 PM and 10:00 PM, with a concentration around 6:00 PM. Elective patients arrive roughly between 7:00 AM and 1:00 PM. The hourly variability discovered in this subsection has to be taken into account in our model which will be introduced in Chapter 4. Therefore, we already give the arrival rates incorporating the hourly variability (i.e., a slight adaptation of the ones given in Table 2.3). Since giving hourly arrival rates will result in 168 different rates per ward which is a bit cumbersome, we assume equal emergency arrivals over the day. Furthermore we assume that all elective patients arrive equally distributed over daytime hours, and are forbidden during night. This somehow represents reality. Not taking into account hourly variability in detail might seem an unrealistic simplification, but this will be justified in Chapter 5 in Section 5.1. The adapted arrival rates are given in Table 2.4.

| | *Monday* | | *Tuesday* | | *Wednesday* | | *Thursday* | | *Friday* | | *Saturday* | | *Sunday* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** |
| **D340& D370** | 7.15 | 1.48 | 6.37 | 1.27 | 8.21 | 1.87 | 7.23 | 1.71 | 6.03 | 2.31 | 2.58 | 1.50 | 2.89 | 1.51 |
| **D350** | 4.27 | 1.25 | 5.20 | 0.96 | 4.12 | 1.14 | 5.00 | 1.08 | 5.62 | 0.98 | 1.19 | 0.85 | 1.74 | 0.94 |
| **D351** | 1.51 | 0.91 | 1.52 | 0.70 | 1.44 | 0.77 | 1.37 | 0.93 | 1.80 | 0.84 | 1.32 | 0.80 | 0.40 | 0.64 |
| **D361** | 2.41 | 2.08 | 2.48 | 1.63 | 2.60 | 2.10 | 3.01 | 1.82 | 3.18 | 2.00 | 2.04 | 1.71 | 3.45 | 2.76 |

Table 2.4.: Daily Poisson arrival rates (both elective and emergency) split up into day (D) and night (N). *Retrieved from data cube, year 2017, n=3219.*

### 2.3.5. Arrivals throughout the year

The last type of pattern that will be examined is the pattern throughout the year. It could for instance be that the number of admissions is lower during summertime due to holidays. In Figure 2.11 the admissions for the entire department have been pictured for both emergency and elective arrivals. The choice to display the entire division at once instead of splitting into wards is made deliberately because it is not expected that this will differ considerably per ward.

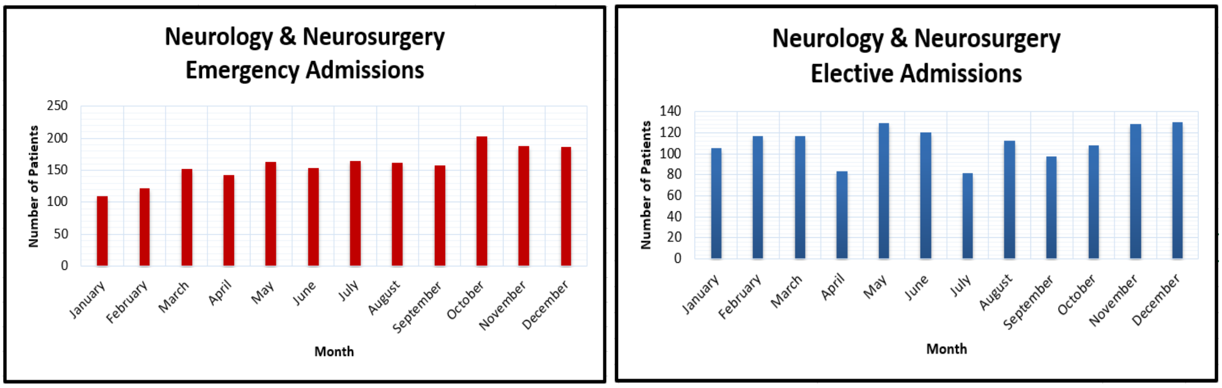


Figure 2.11.: Arrivals of respectively emergency and elective patients within the *N&N* department in 2017. *Data retrieved from data cube, n=1903 and n=1328 respectively.*

As can be seen there are no major differences throughout the year for emergency arrivals; we see that the number of admissions in January is much lower than e.g., in October. There is no clear explanation for this. For elective arrivals however, we see that the number of admissions is a bit lower during the months April and July. For April, this can be explained by the large number of holidays within this month. For July this can be explained because this is a typical month that people go on vacation.

## 2.4. Length of Stay

### 2.4.1. Gini Coefficient

The length of stay is defined as the time that a patient occupies a bed (see Subsection 2.2.1). As a Poisson distribution is a distribution often used to model customer arrivals, there are also distributions that are often used for modelling service times (the LOS can be regarded as such). One can think of the Gamma, Weibull and Lognormal distribution (Law, 2015:286-305). The LOS distribution is analysed per ward. Contrary to the arrivals, there is no clear pattern in this data. This unfortunately was the case for all wards. Several distributions have been tried, but none of them provided a reasonable fit. This is underpinned by analysis of e.g., De Bruin et al. (2009) and Costa et al (2003). Most likely this "fuzzy" data is caused by the fact that the LOS should not be considered as being equally distributed for all patients. We present several summary statistics for all wards. In addition to the regular summary statistics, we also present the Gini coefficient (Gini, 1912). This is suggested by De Bruin et al. (2009), and based on the Lorentz curve often used in economics. The Lorenz curve can be used to picture the concentration of wealth (Lorenz, 1905) (i.e., what percentage of the total wealth belongs to what percentage of the population). See Figure 2.12 for an example.

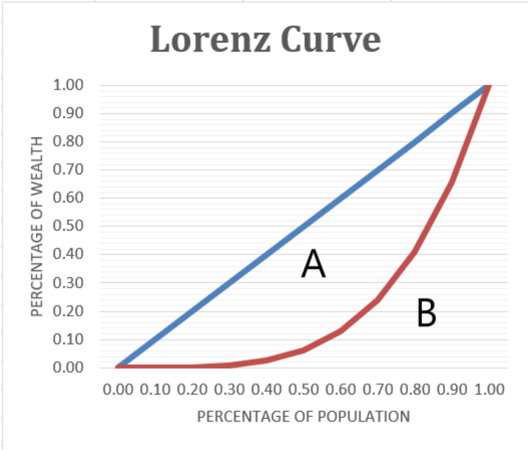


Figure 2.12.: A Lorenz curve.

The bigger the "belly" i.e., area A, the more unevenly wealth is distributed. The blue line means perfect equality. The Gini-coefficient ($G$) quantifies this dispersion as:

$$G = \frac{A}{A+B}$$

Now this idea can be used to analyse the LOS. If the distribution of the LOS would be equal for all patients, $G$ would be close to 0 (i.e., there would be no "belly" at all because of perfect equality). If $G$ is close to one there is a lot of dispersion and the distribution is more variable. It is expected that for some wards $G$ will be quite close to 1, the difference between the medians and means is big. This means that a small part of the patient population has a disproportional long LOS. For the calculation of $G$ per ward, we use the formula provided by De Bruin et al. (2009):

$$G = \frac{1}{n}(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}) \tag{2.7}$$

With $n$ being the total number of observations (i.e., registered LOS), $y_i$ the LOS of arrival $i$ and $y_i \leq y_{i+1}, i = 1, ..., n-1$.

### 2.4.2. Summary Statistics

The summary statistic per ward w.r.t. the LOS are given in Table 2.5.

| | D361 | D340&D370 | D350 | D351 |
|---|---|---|---|---|
| **Mean** | 2.85 | 3.26 | 3.93 | 3.17 |
| **Standard Deviation** | 4.32 | 5.45 | 4.28 | 4.82 |
| **Coefficient of Variation** | 1.52 | 1.67 | 1.09 | 1.52 |
| **Median** | 1.45 | 1.30 | 2.94 | 1.50 |
| **Skewness** | 4.32 | 5.13 | 4.21 | 3.98 |
| **Gini** | 0.58 | 0.64 | 0.45 | 0.59 |

Table 2.5.: Summary statistics LOS in 2017. *Retrieved from data cube, n=3226.*

As can be seen, patients on average stay the longest on the D350 (surgical) ward, and the shortest on the D361 MCU. The LOS data does not really have a pattern which is, as explained before, most likely caused by different LOS distributions for different patient groups. Striking are the big differences between the means and medians. This could indicate very skewed data or outliers. The latter cannot be easily confirmed since for instance on the D340&D370 ward there are many patients who stay very short, and many patients who stay much longer. This also explains the large standard deviations. The best thing to do would be splitting to search for the underlying patient groups having a homogeneous distributions.

### 2.4.3. Distribution Fitting

As said, the LOS data is not as easily described by a probability distribution as the arrival data. This is also encountered by other studies, see for example De Bruin et al. (2009) or Costa et al. (2003). Following the reasoning of Subsection 2.4.1 and our intuition, it might very well be the case the group of patients should be split into patients who stay long and short, each having their own distribution. According to Adan and Resing, we should fit a hyper exponential distribution with two exponentials if the coefficient of variation is greater than or equal to 1 (2015:17). This happens to be the case for all wards (see Table 2.5). The hyperexponential distribution sums several exponential distributions, each with a different parameter (see Appendix B.3). This corresponds to our intuition because e.g., the long stay patients could have their "own" exponential distribution as do the short stay patients. For a hyperexponentially distributed variable the notation $H_k(p_1, ..., p_k; \mu_1, .., \mu_k)$ is common. In this notation, $k$ is the number of groups, $p_i$ stands for the proportion (size) of group $i$ and $\mu_i$ for the parameter of the exponential distribution of group $i$ (i.e., the ALOS of group $i$ is $\frac{1}{\mu_i}$). In order to apply this distribution to our case, we must estimate the parameters $p_1, p_2, \mu_1$ and $\mu_2$ ($k = 2$).

Oftentimes, we assume equal weighted means for both subgroups, i.e., $\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$ ("balanced means assumption"). Furthermore, since we divide the total population into subgroups according to a proportion $p_i$ we have $p_1 + p_2 = 1$ and $\frac{p_1}{\mu_1} + \frac{p_2}{\mu_2} = ALOS$. It turns out that the Gini coefficient $G$ can be of help in this matter. De Bruin and Bekker

(2010) indicate that when $0.5 \leq G \leq 0.75$ we can estimate the parameters using the Gini coefficient as follows:

$$\widehat{p_1} = \frac{1}{2} - \sqrt{G - \frac{1}{2}} \qquad \hat{\mu_1} = \frac{ALOS}{2\hat{p_1}} \tag{2.8}$$
$$\widehat{p_2} = 1 - \hat{p_1} \qquad \hat{\mu_2} = \frac{\hat{p_2}\hat{\mu_1}}{\hat{p_1}}$$

This applies to three out of four wards (see Table 2.5). For the other ward, we need some additional analysis provided by Adan and Resing (2015:18). If $G$ does not fulfill the requirement of De Bruin and Bekker (2010), we can estimate the parameters with:

$$\hat{p_1} = \frac{1}{2}\left(1 + \sqrt{\frac{\widehat{cv}^2 - 1}{\widehat{cv}^2 + 1}}\right) \qquad \hat{\mu_1} = \frac{ALOS}{2\hat{p_1}} \tag{2.9}$$
$$\hat{p_2} = 1 - \hat{p_1} \qquad \hat{\mu_2} = \frac{\hat{p_2}\hat{\mu_1}}{\hat{p_1}}$$

Where $\widehat{cv} = \frac{s}{\bar{X}}$ denotes the coefficient of variation. The results are given in Table 2.6, together with the p-value of the $\chi^2$ goodness-of-fit test.

| | **D361** | **D340&D370** | **D350** | **D351** |
|---|---|---|---|---|
| $\hat{p_1}$ | 0.79 | 0.88 | 0.65 | 0.80 |
| $\hat{p_2}$ | 0.21 | 0.12 | 0.36 | 0.20 |
| $\frac{1}{\hat{\mu_1}}$ | 1.81 | 1.85 | 3.05 | 1.99 |
| $\frac{1}{\hat{\mu_2}}$ | 6.78 | 13.46 | 5.54 | 7.80 |
| p-value $\chi^2$ | **0.01** | **0.00** | **0.00** | 0.05 |

Table 2.6.: Results of fitting the hyperexponential distribution to the LOS of the four wards.

The p-values suggest a bad fit for three out of four wards (i.e., p-value$< \alpha = 0.05$). We made a histogram together with the probability density function (pdf) of the hyperexponential distribution to compare the data with the hypothesized distribution for all wards. See for an example Figure 2.13. After eyeballing this figure, we conclude that the hyperexponential distribution does approach the LOS data fairly good. It does capture most of its variance, and has more or less and identical shape. Besides that, we see that our intuition about the subgroups (long and short stay patients) was right, all four wards have a group of long and short stay patients. It should be noted that the LOS data is hard data to find a suitable distribution for, which again has also been noted by many other researchers before. Both the figure and our intuition therefore let us conclude that it is reasonable to assume the hyperexponential distribution for the LOS for all four wards.

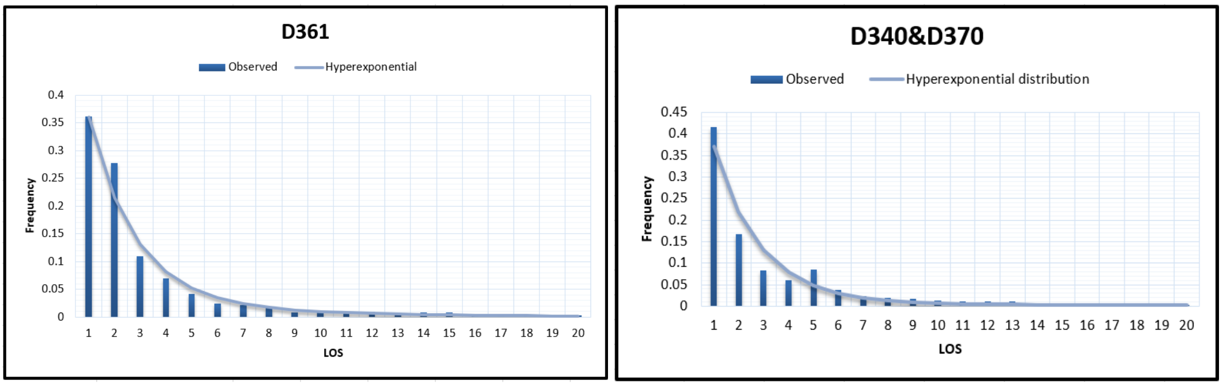


Figure 2.13.: LOS 2017 data vs. hyperexponential distribution. *Note: LOS over 20 days have been left out for illustratory purposes. Data retrieved from data cube, n=714 and n=1345.*

## 2.5. Performance

This section evaluates the current performance with respect to the defined KPIs. The KPI Length of Stay will not be considered again, since it has been extensively analysed in the previous section. First, we discuss the bed occupancy, and after that the rejection rate.

### 2.5.1. Bed Occupancy

To evaluate the current performance with respect to bed occupancy, Equation 2.6 is used. The input needed for Equation 2.6 is given in Sections 2.3 and 2.4. Besides Little's Law, we also calculated the bed occupancy using the hourly presence data (note again, this data is very sensitive for input errors). In Table 2.7 the average bed occupancy for the year 2017 is given, using both Equation 2.6 and 2.4.

| | **Little's Law** | **Data** |
|---|---|---|
| **D340&D370** | 66.67% | 69.38% |
| **D350** | 79.03% | 82.98% |
| **D351** | 58.87% | 51.64% |
| **D361** | 90.73% | 82.20% |

Table 2.7.: Bed occupancy per ward in 2017.

First of all, the bed occupancy for both the D361 and D350 ward are much higher compared to the other wards. Together with the variable arrival- and LOS process, this results in a relatively high rejection rate (see Table 2.8). Little's Law and the data pretty much coincide for the clinical wards, but differ a bit for the MCUs. Especially for the MCUs (a lot of) data was missing, what could have caused this result. See Chapter 6 for a thorough discussion about the available data. Although controversial (see Chapter 4), the management strives for a bed occupancy of 85% for all wards.

### 2.5.2. Diverted Patients

As discussed in Section 2.1.3, diverted patients are patients that are assigned to a bed on another ward other than the ward where he belongs. In the worst case, a patient is diverted to another division or even hospital but this rarely happens. Ideally we would prefer to know the frequency of this phenomenon (for instance to be able to say something about the *true* arrivals of a ward) , but the data cube does not provide the required data. This complicates using this Chapter's analysis for mathematical models since the measured admissions do include diverted patients from other wards. Since for the MCUs diversions almost exclusively happens between the MCUs (i.e., not to other divisions or hospitals) this is not a problem because these wards are merged. For the clinical wards this is more troublesome since there is also a third clinical within the division ward (D360) to which patients can be diverted. However, this ward is left out of consideration in this thesis. Due to the lack of data, we assume that all patients diverted from one of our clinical wards (e.g., D350) are placed on the other (e.g., D340&D370) and v.v, cf. Section 5.1.

### 2.5.3. Rejection Rate

The rejection rate can be calculated using Equation 2.3. Unfortunately, there is no data available about the number of rejections. One could think of considering the fraction of time that the wards are fully occupied considering the same data used for calculating the bed occupancy. However, this data is as stated before highly sensitive for input errors (e.g., more patients than beds present). Another way of calculating the current rejection rate is by considering the queueing model suggested in Chapter 5, which also is used for the calculation of the future situation (see Chapter 3). Because of the sensitivity to input errors we decided to use the model to estimate current performance. The results are depicted in Table 2.8.

| | **Model** |
|---|---|
| **D340&D370** | 3.4% |
| **D350** | 10.9% |
| **D351** | 7.5% |
| **D361** | 23.5% |

Table 2.8.: Estimations of current rejection rate, calculated using the data (first column) and the model of Ch. 5 (second column).

The management has indicated to strive for an average blocking probability of 3% for the MCU and 5% for the clinical ward.

## 2.6. Conclusion

This chapter serves to answer the first and second research question. The relevant KPIs are the bed occupancy, the LOS and the rejection rate. Since the data available is erroneous (see Chapter 6) we decided to calculate the current bed occupancy using Little's Law and the current rejection rate using the model of Chapter 5. The current bed occupancy for the clinical wards is 69.38% (D340&D370) and 82.98% (D350). For the MCUs, the current occupancy is 51.64% (D351) and 82.20% (D361). The current rejection rates for the clinical wards are 3.4% (D340&D370) and 10.9% (D350). For the MCUs, the rejection rates are 7.5% (D351) and 23.5% (D361). The management now primarily bases bed capacity decisions on a target occupancy of 85%. Besides that this target is not met, many studies show that bed capacity decisions should rather be based on the rejection rate. The target rejection rates for the new clinical ward and MCU are 5% and 3% respectively.

For modelling purposes, we wanted to find a suitable distribution for the LOS data and to discover the underlying arrival pattern. The LOS data is known to be difficult, since there are different subgroups within the patient population with different LOS distributions (e.g., short and long stay patients). We fitted a hyperexponential distribution to the LOS data for all wards, which sums various exponential distributions and thus copes with the different subgroups. The ALOS in 2017 was around 3 days for all wards. The arrival data turned out to be well approximated by a non-stationary piecewise constant Poisson process with a cycle length of 1 week. The rates change each day part of 12 hours (i.e., day and night).

# 3. Analysis of the New Situation

## Introduction

As explained in Section 1.3.2 we are primarily interested in the number of beds on the clinical ward and MCU *after* the merger. The analysis in Chapter 2 shed light on the current situation, and can be used to compare with the estimations of the performance of the future situation. This chapter serves to answer the third research question, and has roughly the same set up as the aforementioned chapter. First, the arrival process is discussed, and thereafter the LOS is examined.

## 3.1. Physical Changes

As discussed, the merger will result in both clinical wards and MCUs being combined into one clinical ward and one MCU respectively. The management hopes to gain economies of scale and to improve the communication between the medical professionals. The merger will eliminate the flow in between the clinical wards and MCUs, since the clinical wards and MCUs are no longer separated. This results in a lower arrival rate and higher LOS. The former is caused by the absence of transfers, and the latter by the transfer patients now stay on a single ward. Figure 3.1 illustrates the physical change.

## 3.2. General approach

Since there is no data available of the situation after the merger, we have to deal with this in another way. Intuitively, one would suggest merging the data of the clinical wards and the MCUs respectively. This implicitly assumes that the arrival process and LOS are independent for the separate wards and MCUs. This does not seem to be unreasonable for most patients, except for those who were transferred once or multiple times from e.g., one clinical ward to the other. For example, consider patient X who stays 3 days at the D340&D370 ward. After that, he is transferred to the D350 ward where he stays another 5 days. In the current situation, this would boil down to two separate admissions (one at each clinical ward) and two separate LOS entries. To generate data for the new situation, one would ideally delete one admission, and add up both LOS (i.e., one admission at the new clinical ward with a LOS of 8 days). Unfortunately, the data cube does not provide the LOS data per transfer as is required (see Section 6.3.1). Since these transfers from clinical ward to clinical ward and MCU to MCU caused only a relatively small part of the total admissions in 2017 (1.12% and 4.46% respectively), we ignored this phenomenon and merge the arrival and LOS data as given.

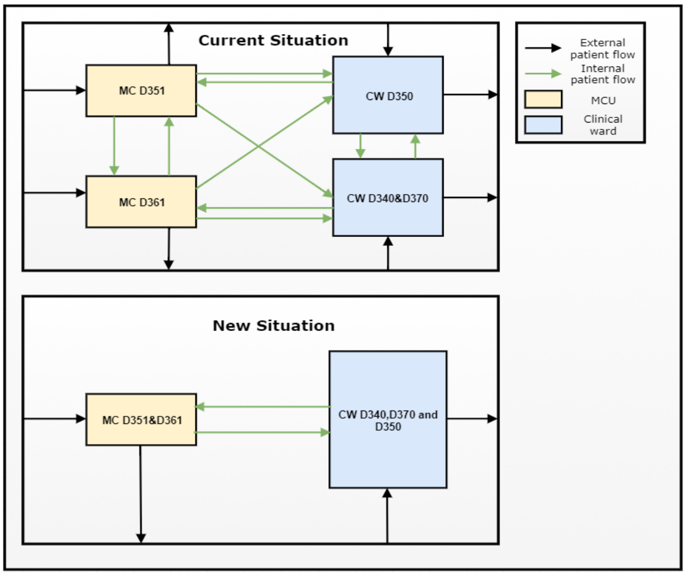


Figure 3.1.: A schematic view of the current and future situation.

We thus assume independency of the arrivals for both clinical wards and MCUs.

## 3.3. Arrivals

The arrival process will again be analysed for the clinical ward and MCU separately. First, we look at the arrivals per day. Subsequently patterns throughout the day and week respectively are investigated. As shown in Section 2.3, the arrival processes for all wards tend to be well described by a (non-stationary) Poisson process. One of the properties of a Poisson process is the *merging* property: if we have two independent Poisson processes $X$ and $Y$ with parameters $\lambda_1$ and $\lambda_2$ respectively, we can describe the joined process $Z = X + Y$ again with a Poisson process, with parameter $\lambda_1 + \lambda_2$. This directly applies to our situation, since we have just assumed an independent arrival process for the separate wards. Using the merging property, the arrival rates for the new wards are given in Table 3.1.

| | *Monday* | | *Tuesday* | | *Wednesday* | | *Thursday* | | *Friday* | | *Saturday* | | *Sunday* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** |
| **CW** | 7.15 | 1.48 | 6.37 | 1.27 | 8.21 | 1.87 | 7.23 | 1.71 | 6.03 | 2.31 | 2.58 | 1.50 | 2.89 | 1.51 |
| **MCU** | 4.27 | 1.25 | 5.20 | 0.96 | 4.12 | 1.14 | 5.00 | 1.08 | 5.62 | 0.98 | 1.19 | 0.85 | 1.74 | 0.94 |

Table 3.1.: Predicted Poisson arrival rates for the new clinical ward and MCU per day of the week, split into day (D) and night (N) time.

## 3.4. Length of Stay

Similar to the current situation, we also analyse the LOS for the future situation. We generated the LOS data for the new wards as described in the section "General Approach". The summary statistics for the LOS data of the new wards can be found in Table 3.2. For modelling purposes we again have fitted a statistical distribution through the data. For both wards again a hyperexponential seems to describe the underlying data properly and is therefore assumed to be a good fit. The parameters of both hyperexponential distributions together with the p-values of the $\chi^2$ goodness-of-fit test are given in Table 3.2. Note that the p-values suggest a bad fit. However, because the variability of the LOS data is well captured and taking into account that it is hard to find a suitable distribution for the LOS data, we ignore this.

| | **CW** | **MCU** |
|---|---|---|
| **Mean** | 3.52 | 2.94 |
| **Standard Deviation** | 5.03 | 4.47 |
| **Coefficient of Variation** | 1.43 | 1.52 |
| **Median** | 2.10 | 1.45 |
| **Skewness** | 4.92 | 4.21 |
| **Gini** | 0.57 | 0.59 |
| $\hat{p}_1$ | 0.77 | 0.79 |
| $\hat{p}_2$ | 0.23 | 0.21 |
| $\frac{1}{\hat{\mu}_1}$ | 2.29 | 1.86 |
| $\frac{1}{\hat{\mu}_2}$ | 7.69 | 7.10 |
| p-value $\chi^2$ | **0.00** | **0.02** |

Table 3.2.: Summary statistics and parameter estimation of the LOS of new clinical ward and MCU.

## 3.5. Conclusion

In this chapter, we answer the third research question. Similar to Chapter 2, we considered to arrival and LOS process. To generate data for the new wards, we merged data of the separate wards, and ignored transfers in between these separate wards. Using the merging property of the Poisson process, we (again) can assume a non-stationary Poisson process with a cycle length of one week and varying arrival rates per day part of 12 hours. The LOS data is well approximated by a hyperexponential distribution, with an ALOS of 3.52 days for the clinical ward and 2.94 days for the MCU.

# 4. Modelling Approach

This chapter discusses the modelling approaches available to answer the central research question, and to provide argumentation to make a valid choice amongst them. This chapter is supported by the existing literature on the subject. Therefore we first define the theoretical perspective used for the literature search. Subsequently we provide an overview of the techniques available to determine the number of beds, and conclude with the method of our choice.

## 4.1. Theoretical Perspective

The theoretical perspective used in this chapter is that of optimizing hospital bed capacity along the edges of Operations Research (OR). Optimizing in this context means establishing a perfect trade off between bed occupancy and patient rejections (cf. Section 2.2), while keeping the variability in bed demand in mind. Operations Research is an umbrella term consisting of mathematical techniques to support decision making.

## 4.2. General notions

There are many different approaches to optimize bed capacity. Roughly, most research either is based on queueing theory and/or simulation. Most authors agree that current practices of determining bed capacity used by healthcare managers are not correctly underpinned. First of all, now the bed occupancy is oftentimes leading in the decision making process (e.g., the "85% occupancy rule"). Green (2002) points out that it is better to base the decision regarding the number of beds on the availability of care that is the fraction of declined patients. Secondly, performance indicators which are used to determine the number of beds, are often not well defined (e.g., Green, 2002; Cochran and Roche, 2008). And finally, many managers underestimate the importance to incorporate the variability of e.g., LOS and patient arrivals. Decision makers now primarily use averages to make their decisions, resulting in the underestimation of beds required (De Bruin et al., 2009). There is rich literature about solving the issues mentioned. This chapter serves to shed some light on this, and ultimately providing an (preliminary) answer the research question. We will discuss methods from Queueing Theory, Discrete Event Simulation and Mathematical Programming.

## 4.3. Queueing Theory

Queueing theory is a branch of OR and can be used to describe and predict behavior of waiting lines. It also entails the study of the arrival process (input process) and the service process (output process). In case of a hospital ward, the beds can be seen as servers, the patients' length of stay as service time, the admission of patients per time as arrival rates and a possible waiting list as queue. Queueing systems are oftentimes categorized using the Kendall-Lee notation : $A/S/c/K/N/D$ (Kendall,1951). In this notation, A stands for the arrival process, S for the service process, c for the number of servers, K for the queue capacity, N for the population size and D for queueing discipline. Oftentimes the last two symbols are omitted, because of their irrelevance to the system on hand.

Regarding the problem under examination, several queueing related methods are proposed. Green (2002) and Cochran and Roche (2008) suggest the $M/M/c/\infty$ system to model the situation. This model assumes exponentially distributed inter arrival times (i.e., a Poisson process, see Subsection 2.3.1), exponentially distributed service times, c servers (beds) and an infinite queue capacity. Exponentially distributed inter arrival times are proposed for both elective and scheduled admissions. As discussed before, it is reasonable to approximate both scheduled and unscheduled arrivals with a Poisson process. Since two merged Poisson processes again form a Poisson process, the "M" assumption for patient arrivals is not unreasonable on itself, but it also assumes stationary. As is pointed out in Chapter 2, the arrival process for elective patients varies across the week. The model its main advantage is the relatively simple calculations of KPIs. It is not hard to point out the restrictions of this model. For instance, this model assumes infinite queue capacity. It is probably more realistic to assume a finite queue capacity, or even no queue capacity at all since patients that arrive when all beds are occupied, are diverted to another ward. Besides that, the model assumes homogeneous LOS and arrival rates, which might very well be a wrong assumption as is pointed out in Section 2.6.2. Green (2002) does consider the refusal of patients, by examining the probability of delay. Cochran and Roche (2008) optimize capacity with respect to target utilization, which is not the way to go as explained earlier.

De Bruin et al.(2009) model a hospital ward as a $M/G/c/c$ queueing system (also known as Erlang-Loss model). In this model, the patient arrival process is assumed to be a Poisson process, the service distribution is not pre-specified (G means general) and that the number of beds equals the queue capacity i.e., patients that arrive when all beds are occupied are "blocked". According the De Bruin et al., blocked in this context could mean sent away to another ward, or even another hospital. This seems to be a more accurate approach, because in our situation there also is no waiting room and the LOS distribution is hard to specify. A drawback is that also this model does not take non stationarity of the arrival process into account. Besides that, there is still some generalization of the LOS needed, albeit to a lesser extent than the previous approach.

Gorunescu et al.(2002) and Belciug and Gorunescu (2014) provide an even more sophisticated approach because they also take several cost aspects into consideration. The method is based on the $M/PH/c/c$ model. This model is identical to the one De Bruin et al. use except for the distribution of the LOS. The PH stands for phase type distribution, which means that different type of patients are allowed to have different LOS parameters. An example of a phase type distribution is the Erlang distribution. Together with the blocking probability they use inventory theory to draw up a cost function. More specifically, slow moving expensive items have demand processes that are often modelled with a Poisson process (due to the infrequent random arrivals of customers). The goal here is to optimize the trade off between stock out cost and holding cost. Gorunescu et al. (2002) use this approach, where the stock out cost can be seen as the cost for diverting a patient, and holding cost the cost for having an empty bed. The resulting cost function contains a factorial which makes the optimization not straightforward. In more recent work Belciug and Gorunescu append the described approach by suggesting the use of a genetic algorithm to optimize the cost function (2014). A drawback of these approaches is the stationarity assumption, and the fact that it is really hard to properly determine the costs involved.

The last queueing model that will be discussed is the $M_t/H_2/s/s$ model as proposed by Bekker and de Bruin (2010) and Bekker and Koeleman (2011). This model assumes a non-stationary piece wise constant Poisson process, i.e., a different arrival rate per time interval $t$ (e.g., a different rate per day of the week or hour of the day). The LOS is assumed to be hyperexponentially distributed with 2 exponential distributions. This model does not need many restrictive assumptions, and its non-stationarity assumption is distinctive. However, exact analysis of this model and even for the more general $M_t/G/s/s$ is known to be difficult. To this end, there are two famous approximations available, being the Modified Offered Load (MOL) (Jagerman, 1975) approximation and the Pointwise Stationarity Approximation (PSA) (Green and Kolesar, 1991). Massey (2002) argues that the MOL approximation is superior to the PSA approximation, which is also the approximation that Bekker and de Bruin (2010) base their analysis on. Bekker and Koeleman (2011) base their approximation on the Hayward approximation with non-integer input for the Erlang-Loss function. The drawbacks are the requirement of a piece wise constant arrival rate, where this is likely to fluctuate smoothly over time. Furthermore, the model requires a (hyper)exponential LOS distribution, where this is obviously not always the case. All in all, this last model requires only a few assumptions and captures most of the variability discovered in Chapter 2.

## 4.4. Simulation Modelling

Nearly all parts of critical care have a certain variability because many processes are of a stochastic nature. Although the models in the previous section do take this variability into account to some extent, all models have to obey several restriction assumptions. For example, homogeneous arrival patterns were assumed for most models. As Costa et al.(2003) point out, it is nearly impossible to encompass all this variability through mathematical equations (i.e., queueing theory). Therefore, the use of simulation modelling is suggested. Simulation modelling is another branch of Operations Research, that is often used to provide insight in the effect of variables in complex systems. An advantage of simulation modelling is that generally fewer assumptions are required than when one uses an analytic model. Besides that, it is also a suitable tool to perform experiments or what if scenarios as is done by Devapriya (2015) because the real system is not used (which would oftentimes be very costly). A drawback of simulation is the fact that one is actually performing experiments and uses this experiments to draw inferences about the system in real life. This is obviously less "precise" than when using an analytic solution of which queueing theory is an example. Furthermore, building a simulation model is time consuming, and case specific. A last drawback is that also with simulation one cannot perfectly represent reality which also is the case for queueing theory. All discussed articles use a (sometimes healthcare) specific simulation tool. Because mainly the model development process is relevant, the used software is not discussed.

Costa et al. provide a systematic approach of how to develop a simulation model to provide insight into bed capacity (2003). This approach consists of four steps. The first is to describe the patient flows in the hospital compartment under study. This could be done by e.g., a flowchart. Subsequently, one has to analyse the statistical information available. This encompasses fitting statistical distributions to e.g., the arrival rate and LOS. According to Costa et al. one needs three types of data, being demographic data about the patient group that are likely to affect the LOS, data about arrival patterns and lastly data describing the LOS. The third step is to repetitively perform experiments where after one should validate the results with the "real world" to ensure correctness of the model.

Once the model has been built, it is important to draw up useful performance indicators. Harper and Shahani suggest the bed occupancy (which they define as the ratio of bed days used to bed days available) and the refusal rate (2002). Devapriya et al. add patient wait time (if applicable) and the number of patients arrived per arrival source (2015). Finally, Costa et al. (amongst others) suggest the total number of admissions and the admissions split into elective and emergency patients (2003).

## 4.5. Mathematical Programming

Up to this point, we have primarily discussed the optimal number of beds based on some care process related characteristics. Closely related to the discussion above is how to deal with scheduled- and unscheduled arrivals. One of the issues belonging to this area is the impact of scheduled admissions on variability in bed demand. It turns out that the blocking probability increases as the peakedness of inter arrival times increases (Bekker and Koeleman, 2011). In other words, smoothing the elective arrivals has a positive effect on the number of patients that are blocked, on the variability of bed demand and thus lower number of beds required.

Bekker and Koeleman (2011) propose the use of quadratic programming with linear constraints to deal with this. The objective function is the sum over a time period of the quadratic difference between a target load[1] and the offered load. The constraints are primarily concerned with the calculation of this offered load. Bekker and Koeleman

[1]Load=Arrivals per time unit * Time required per patient

provide models to calculate this load in case of non-homogeneous arrival patterns per day. Based on this model a guideline for the number of elective admissions per day can be derived. The described model does not need a lot of restricting assumptions; the arrival process may be non-homogeneous and the LOS can be different for different patient groups. However, the LOS still needs to be (hyper) exponentially distributed. Furthermore, the calculations involved are quite cumbersome. This could be a reason not to apply this method given the limited time available. Another aspect that limits the applicability of the described model is that it does not take into account that (a part of the) elective arrivals are dependent on e.g., operation room schedules, which are in turn dependent on availability of personnel. Leaving this out of consideration could be a non-realistic simplification.

## 4.6. Conclusion of the Literature Study

Both simulation and queueing models can provide the answer for the central research question. The analyses of Chapters 2 and 3 show that we are dealing with a non-stationary Poisson process, and that the processing times are hyperexonentially distributed with two exponentials. Furthermore, the system can be seen both as two separate "stations" and as a network as a whole. There are no closed form queueing models available for the latter point of view that entails the discussed variability in arrivals and service times. This could be a reason to consider simulation. On the other hand, analytic techniques as queueing theory are less time consuming and are more suitable for creating a tool such that the management is able to replicate the calculations in this thesis even after the project has been completed. Furthermore, using queueing theory and mathematical programming it might be possible to also take a look at how the management should organise the elective arrivals and what that would yield in terms of bed capacity by using quadratic programming.

Taking all of the above into consideration, the modelling approach of this thesis will be based on the $M_t/H_2/c/c$ model discussed by Bekker and Koeleman (2011) and Bekker and de Bruin (2010) in Section 4.3 because it encompasses most of the discovered variability. Besides that, it also shares part of the analysis required for the quadratic programming model discussed in Section 4.5. To predict the effect of the new admission schedule generated by the quadratic program, we built a simulation model. This is the answer to the fourth research question.

# 5. Application of models

In this chapter we apply the models of Chapter 4 to our situation. The results are presented and form the foundation of the recommendations of Chapter 6. We first discuss the assumptions needed for the sequel of this chapter. Subsequently we present the queueing model which gives the relation between the bed occupancy and blocking probability. The quadratic programming model that follows shows us how to alter the elective arrivals such that a target load (i.e., number of patients present) per day is met and there is no variability in arrivals. With a simulation model we then see what the effects are of this reduction in variability. The calculations for the queueing model were done using Matlab. The quadratic program is solved using the CPLEX plug-in for Matlab. The simulation model is built in Siemens Tecnomatrix Plant Simulation, version 13.

## 5.1. Assumptions

It is nearly impossible to perfectly represent reality using (mathematical) modelling. However, mathematical modelling can provide useful insight into the relations amongst the variables on hand. To this end, we need to make assumptions under which we can model the system and calculate several characteristics. This section provides these assumptions:

- We assume Poisson arrivals, as is substantiated in Chapter 2 and 3.
- We assume hyperexponential LOS, which seems reasonable according to Section 2.4.3 and Section 3.4.
- We ignore the interaction between the MCU and the clinical ward.
- We assumed that fluctuations in arrivals *within* the time span of 12 hours do not influence our analysis. This is underpinned by the rule of thumb derived by Bekker and Koeleman (2011), who state that if the time span of fluctuating arrivals is less than 5 times the ALOS, these fluctuations average out.
- We assume that the admission data of Chapter 2 and Chapter 3 represent the real arrival data (i.e., refused admissions are not taken into account). This is justified by the fact that many refused D351 patients are admitted on D361 and v.v. (which causes the data cube's arrival rate to be higher on both MCUs), and many refused D340&D370 patients are admitted on the D350 ward and vice versa. Besides that, the number of patients that is diverted to another division or hospital is negligible (cf. Section 2.5.2). We ignore the patients diverted from and to the D360 ward, since this ward is not part of this study.

• We assume the time of arrival to be independent of the LOS, whereas this in reality might not be the case (e.g., a patient arriving at 11:00 PM is not likely to leave before 07:00 AM the next day).

## 5.2. Queueing Model

We will first discuss some section specific notation. For the main definitions we refer to Table 1. Following the analyses of Chapter 2 and 3, we consider time intervals of twelve hours which results in 14 intervals per week (day and night for each day of the week), i.e., $t = 1, .., 14 = T$. As can be seen in Table 1, $m(t)$ is the load during interval $t$ from which it follows that $m(t)$ is piece wise constant. $\mu_1, \mu_2, p_1, p_2$ correspond to the parameters of the hyperexponential LOS distribution with $k$ exponentials, see Section 2.4.3 and Appendix B.3.

First of all, $m(t)$ needs to be calculated. This is done in two steps. First the $m(t)$'s are calculated per "hyperexponential group" (Equation 5.1), and then the total $m(t)$ is calculated (Equation 5.2).

$$m^k(t) = \frac{1}{\mu_k} \frac{1 - e^{-\mu_k}}{1 - e^{-T\mu_k}} \sum_{i=0}^{T-1} \lambda_{t-i} e^{-\mu_k i}, \ t = 1, .., 14, \ k = 1, 2 \tag{5.1}$$

$$m(t) = p_1 m^1(t) + p_2 m^2(t) \tag{5.2}$$

For the derivation of Eq. 5.1 and 5.2 we refer to Bekker and Koeleman (2011) and Bekker and de Bruin (2010) and references therein. Using the calculated load we can estimate the probability of $c$ customers present (i.e., the blocking probability) using the MOL approximation of Jagerman (1975) discussed earlier, and besides that also estimate the bed occupancy given a number of beds. Let $B_t$ denote the blocking probability during $t$, and let $\bar{B}$ denote the average blocking probability over all $t$. Then $B_t$ and $\bar{B}$ are calculated using:

$$B_t(s, m(t)) = \frac{m(t)^s / s!}{\sum_{i=0}^{i=s} m(t)^i / i!} \tag{5.3}$$

$$\bar{B} = \frac{1}{T} \sum_{t=1}^{t=T} B_t \tag{5.4}$$

The bed occupancy on $t$ can be calculated by dividing the product of the average acceptance rate (i.e., $1 - \bar{B}$) and the offered load $m(t)$ over the number of beds available $c$, i.e.,

$$Bed\ Occupancy_t = \frac{(1 - \bar{B}) m(t)}{c} \tag{5.5}$$

### 5.2.1. Input data

We need input data regarding the arrivals ($\lambda_t$) and the processing times ($\frac{1}{\mu}$). This input is directly provided by the analysis of Chapter 3. Regarding the arrivals Table 3.1 will be used. For the processing times we will make use of Table 3.2.

### 5.2.2. Mean Load across the week

In this subsection we predict the mean number of patients present per day part of the week (day and night for each day). These figures are presented in Table 5.1.

| | *Monday* | | *Tuesday* | | *Wednesday* | | *Thursday* | | *Friday* | | *Saturday* | | *Sunday* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** | **D** | **N** |
| **MCU** | 10.5 | 10.0 | 10.5 | 9.5 | 10.2 | 9.7 | 10.6 | 9.9 | 11.2 | 10.4 | 10.3 | 9.5 | 10.1 | 10.1 |
| **CW** | 22.0 | 19.0 | 24.3 | 20.1 | 25.7 | 21.9 | 26.9 | 22.6 | 27.0 | 23.1 | 20.8 | 18.0 | 17.9 | 15.9 |

Table 5.1.: The predicted number of patients present per day part per ward.

We can easily conclude that the mean number of patients present (i.e., mean load) on the MCU is rather constant, whereas the mean number of patients on the clinical ward fluctuates heavily. Combining the fact that the mean load only depends on the LOS- and arrival distribution and that the LOS distribution is constant throughout the week, we can derive that this fluctuation in load is caused by the fluctuating arrival rate (cf. Section 3.3).

### 5.2.3. Bed Occupancy vs. Blocking Probability

For both the new clinical ward and the MCU, we calculated the average and maximum blocking probability (i.e., the *day part* with the highest rejection rate) across the week for a different number of beds, using Equations 5.3 and 5.4. These numbers are denoted with $\bar{B}$ and $B_{max}$ respectively. Besides that, we calculated the corresponding average bed occupancy. From the average blocking probability, we can easily derive the mean number of days that the regarding ward is fully occupied (i.e., $\bar{B} \cdot 365$). These numbers have been included for illustratory purposes. For the new MCU, the results are given in Table 5.2. For the CW, the results are given in Table 5.3.

| Nr. of beds | $\bar{B}$ | $B_{max}$ | Occupancy | Days fully occupied |
|---|---|---|---|---|
| **11** | 17.1% | 21.7% | 76.7% | 62.5 |
| **12** | 12.7% | 16.9% | 74.1% | 46.4 |
| **13** | 9.1% | 12.7% | 71.2% | 33.2 |
| **14** | 6.2% | 9.3% | 68.2% | 22.7 |
| **15** | 4.1% | 6.5% | 65.1% | 14.9 |
| **16** | 2.6% | 4.4% | 62.0% | 9.3 |

Table 5.2.: Trade off between the blocking probability and bed occupancy for the new MCU.

| Nr. of beds | $\bar{B}$ | $B_{max}$ | Occupancy | Days fully occupied |
|---|---|---|---|---|
| **25** | 8.8% | 18.5% | 79.5% | 32.1 |
| **26** | 7.2% | 16.1% | 77.8% | 26.2 |
| **27** | 5.8% | 13.9% | 76.0% | 21.2 |
| **28** | 4.6% | 11.8% | 74.2% | 16.9 |
| **29** | 3.6% | 9.9% | 72.4% | 13.3 |
| **30** | 2.8% | 8.2% | 70.6% | 10.3 |
| **31** | 2.2% | 6.6% | 68.8% | 7.9 |
| **32** | 1.6% | 5.3% | 67.0% | 5.9 |
| **33** | 1.2% | 4.2% | 65.2% | 4.4 |
| **34** | 0.9% | 3.2% | 63.5% | 3.2 |
| **35** | 0.6% | 2.4% | 61.9% | 2.3 |

Table 5.3.: Trade off between the blocking probability and bed occupancy for the new clinical ward.

The capacity decision based on the blocking probability can be made by starting at the bottom of the table and move up till $\bar{B}$ or $B_{max}$ "hits" the pre-specified target. Especially for the CW the differences between the occupancy and the max occupancy respectively $\bar{B}$ and $B_{max}$ are large. This is due to the fluctuating load across the week, see Section 5.2.2. Particularly for the CW, we can alter this load by improving the way elective arrivals are scheduled. If we eliminate (unnecessary) variation in the elective arrivals, most likely the rejection rate will drop and the occupancy will rise (Bekker and Koeleman, 2011). In the next subsection we further analyse this issue.

## 5.3. Quadratic Programming Model

To derive an admission schedule, we use a quadratic program based on Bekker and Koeleman (2011). Since we are interested in a number of allowable arrivals per day, the model is on a day-to-day basis (hence does not consider day and night). In this model, we minimize the sum of quadratic differences between the offered load on day $t$, $m(t)$, and a target load $m^*(t)$ (Eq. 5.6) . A possibility is to redistribute the calculated load of Table 5.1. This target load can either be based on a target blocking probability (i.e., Eq. 5.3), or a target occupancy. The choice of using the quadratic instead of the absolute deviation is based on the intuition that larger differences in load cause bigger problems for the staff. The model is given below. For the definitions we refer to Table 1.

$$\textit{Minimise} \qquad \sum_{t=1}^{7}[m(t) - m^*(t)]^2 \tag{5.6}$$

$$\textit{Subject to} \qquad m^k(t) = \frac{1}{\mu_k}\frac{1-e^{-\mu_k}}{1-e^{-T\mu_k}}\sum_{i=0}^{6}\lambda_{t-i}e^{-\mu_k i}, \qquad t = 1,..,7,\ k = 1,2 \tag{5.7}$$

$$m(t) = p_1 m^1(t) + p_2 m^2(t) \tag{5.8}$$

$$\lambda_t = c_t^1 \qquad t = 6,7 \tag{5.9}$$

$$\sum_{i=1}^{i=7}\lambda_i \geq \Lambda \tag{5.10}$$

$$\lambda_t \geq c_t^2 \qquad t = 1,..,7 \tag{5.11}$$

The objective function 5.6 minimizes the sum of the quadratic differences between the load as a function of the new arrival rates and the target load per day. Constraints 5.7 and 5.8 calculate the load as a function of lambdas, and are equivalent to equations 5.1 and 5.2. Constraint 5.9 is optional, and can be used to forbid elective arrivals during the weekend (e.g., scenario 2 of the next section). In that case, $c_6^1$ and $c_7^1$ are equal to the arrival rate of emergency patients during the weekend. Constraint 5.10 ensures that the *total* number of arrivals is equal or greater than it is now. Constraint 5.11 can be used to ensure that the $\lambda$s are at least equal to the emergency arrivals per day (since these cannot be altered).

Solving the model results in a vector of $\lambda$s (one for each day of the week). Subtracting the emergency arrivals results in the allowable elective arrivals. Non-integer arrival rates are rounded to the nearest integer as is suggested by Bekker and Koeleman (2011). As the target load is calculated using the current arrival rates, the total new arrival rate may not be lower than the total current arrival rate since that would lead to an unfair comparison. If the rounding procedure results in a total number of elective arrivals lower than it is now ($\approx 24.4$), we have fixed this by rounding *up* the highest fractional rate(s) which was/were initially rounded *down*.

### 5.3.1. Implementation

To implement the quadratic programming model, we have to determine the target load per day of the week. We do so by redistributing the load per week as given in Table 5.1. The total load for the clinical ward is 153.13. The management has indicated that the target load during weekends should be 4 patients less during than on weekdays, because then one nursing shift can be saved. Besides that, we try to level the load throughout the week as much as possible. This results in a load of $m^*(1) = ... = m^*(5) = 23.019$ on weekdays and $m^*(6) = m^*(7) = 19.019$. We will evaluate two scenarios with this target load. In the first scenario, elective arrivals during weekends are allowed, whereas these are not allowed in the second scenario. A third scenario that will be evaluated is when we try to level the load over the week *including* weekends. This results in a target load of $m^*(1) = ... = m^*(7) = 21.88$ The parameters $\mu_1, \mu_2$ are given in Table 2.6 and $\Lambda = 43.22$. The results are given in Table 5.4.

| | **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
|---|---|---|---|---|---|---|---|
| **Scenario 1** | 7 | 4 | 4 | 4 | 3 | 0 | 3 |
| **Scenario 2** | 9 | 4 | 4 | 4 | 4 | 0 | 0 |
| **Scenario 3** | 4 | 4 | 3 | 3 | 3 | 4 | 4 |

Table 5.4.: Target admissions to achieve specified load (rounded off).

The table provides guidelines about the number of admissions per weekday, given the constraints per scenario. If no elective arrivals during weekends are allowed (Scenario 2), the "missed" patients should roughly be admitted on Monday. If the load should be equal for all days, we see that the number of arrivals allowed per weekday are nearly equivalent. One might expect to have a schedule of 4-3-4-3-4-3-4 for Scenario 3, since that seems to be more equal throughout the week. The actual schedule looks different due to the fact that the number of emergency arrivals is not perfectly level through the week. The number of elective arrivals on Monday is very high for Scenario 2, which might lead to a high blocking probability. It would be interesting to know what the effects of the schedules in Table 5.4 would be on the occupancy and blocking probability. Equation 5.3 cannot be used to this end, since we no longer have Poisson arrivals for the elective patients. We therefore need a model that takes a mix of Poisson and deterministic arrivals. To the best of our knowledge, there is no closed form analytical expression to solve this problem. Therefore, we built a simulation model which is substantiated in the next section.

## 5.4. Simulation Model

In this section, we perform a simulation study to analyse the performances of the admission schedules derived in the previous section for the clinical ward. To develop a sound model and draw proper inferences, we use the guidelines of Law (2015:67). We first describe the system, collect data and set the required assumptions. This has been done extensively in Chapters 2 and 3 and Section 5.1. Subsequently the computer program is made, and one we have taken care of validation and verification issues. Thereafter the warm-up period and run length discussed. Lastly the experiments are performed and inferences are drawn.

### 5.4.1. The Model

#### Input and Output Variables

The input of the simulation model consists of a description of both the arrival- and service process. This analysis is provided in Chapter 3. We assume Poisson emergency arrivals, deterministic elective arrivals and hyperexponential service times. The output variables of interest are the fraction of time that the clinical ward is fully occupied (i.e., blocking probability) and the bed occupancy. The arrival rates are again split up into day (6:00 AM to 6:00 PM) and night (6:00 PM to 6:00 AM) per day of the week. We assume that the (deterministic) elective arrivals occur uniformly between 6:00 AM and 6:00 PM. The occupancy and blocking probability (fraction of time that all servers are occupied) are registered on an hourly basis (i.e., number of patients present at the start of each hour vs. the number of beds available). Because the arrival rate fluctuates over time and the simulation software generates inter arrival times (cf. Section 2.3.1), a thinning algorithm is used (Law, 2015:477-478). In this algorithm, we generate arrivals according to the maximum arrival rate and accept the arrival with a probability equal to the fraction of the arrival rate of that day part over the maximum arrival rate.

#### Flow chart

To visualize the building blocks of the simulation, an event flow chart is used. The flow chart is given in Figure 5.1.

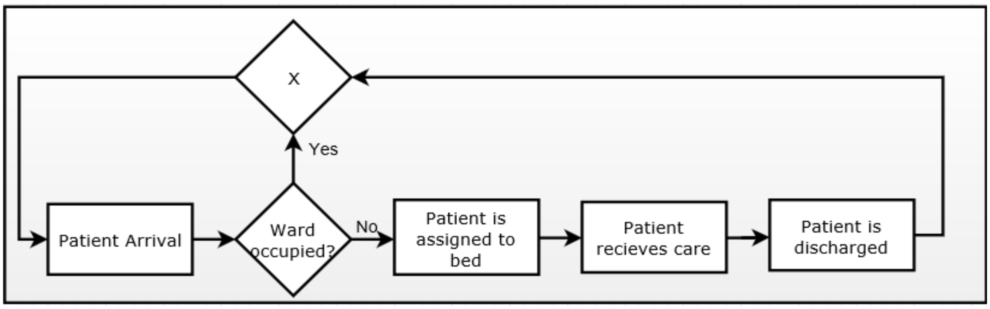


Figure 5.1.: Event flow chart of the clinical ward.

### 5.4.2. Verification & Validation

#### Verification

According to Law, “Verification is concerned with determining whether the “assumptions document” has been correctly translated into a computer “program”...” (2015:246). This boils down to translating the assumptions in Sections 5.1 and 5.4.1 correctly into a model that works. Since the required input distribution can easily be configured in Plant Simulation, this should not cause any trouble. Furthermore, the thinning algorithm takes care of the fluctuating arrivals. If we run the simulation model it runs without errors and produces reasonable output, justifying the claim that the model is verified.

#### Validation

Validation is concerned with ensuring that the simulation is a “good” representation of reality. That is, “Validation is the process of determining whether a simulation model is an accurate representation of the system, for the particular objectives of the study” (Law,2015:247). In our context, this means that the simulation should be a credible representation of reality and able to predict the mentioned KPIs for the new situation. A commonly used technique is “black-box validation” where simulation output is compared to real world output and statistical techniques are used to judge the (possible) differences. Since this real world output is not yet available (we are simulating a future situation), we are designated to other techniques. We therefore compared the simulation output with the output of Table 5.3 given the same input assumptions. This comparison is given in Table 5.5. On eyeballing we see that these values do not differ significantly. Furthermore, Law suggests having conversations with Subject-Matter Experts (SMEs) who are able to judge whether the simulation model on hand is valid (2015:267-268). In our case, the SMEs will be the management. Furthermore, Law states that thorough data analysis and an assumptions document contribute to validity (2015:258-259). Together with the management we consider the simulation model as being valid.

### 5.4.3. Output analysis

#### Type of simulation

Since simulation modelling essentially boils down to using a computer program to model a real world situation and performing experiments to draw inferences, we need statistical techniques to do this in a correct manner. We are dealing with a non-terminating simulation since there is no natural endpoint. We could either have transient, steady state or cyclic steady state output. Since our arrival rate varies over the week, cyclic output seems to be the case (Law, 2015:493-497). To deal with this, we need cyclic steady state output parameters, and thus calculate the blocking probability and occupancy as averages per week.

| | *Occupancy* | | *Blocking Probability* | |
|---|---|---|---|---|
| **No of Beds** | **Simulation** | **Queueing** | **Simulation** | **Queueing** |
| **25** | 78.65% | 79.51% | 8.34% | 8.79% |
| **26** | 77.20% | 77.79% | 6.52% | 7.19% |
| **27** | 75.73% | 76.02% | 5.37% | 5.81% |
| **28** | 74.16% | 74.22% | 4.08% | 4.63% |
| **29** | 72.54% | 72.40% | 3.13% | 3.65% |
| **30** | 70.81% | 70.58% | 2.42% | 2.83% |
| **31** | 68.99% | 68.77% | 1.86% | 2.16% |
| **32** | 67.30% | 66.99% | 1.31% | 1.63% |
| **33** | 65.62% | 65.24% | 0.93% | 1.20% |
| **34** | 63.93% | 63.53% | 0.62% | 0.88% |
| **35** | 62.28% | 61.87% | 0.44% | 0.63% |

Table 5.5.: Comparison of queueing model with simulation model.

### Warm-Up

As stated above, we are dealing with non-terminating simulation and thus have steady state output. However, it usually takes some time till the steady state is achieved (Law, 2015:512). The data retrieved during this *warm-up* period should not be taken into account when drawing inferences. There are two ways of calculating the warm-up period suggested by Robinson (2014:175-178) and Law (2015:511-523). First graphical inspection, second the Marginal Standard Error Rule (MSER). The MSER seems more objective, which is why we decided to use it. For the formal definition of the MSER rule we refer to Robinson (2014:177-179). We apply the MSER to the elective arrivals according to scenario 2 from Table 5.4 with a bed capacity of 25. Using data from 5 replications (initial guess) with a run length of 52 weeks and the maximum warm-up period of the average bed occupancy and blocking probability per week, we have a warm-up period of 1 week. However, the hyperexponential distribution of the LOS of the clinical ward has as parameter values $\frac{1}{\mu_1} = 2.29$ days with $p_1 = 0.79$ and $\frac{1}{\mu_2} = 7.69$ days with $p_2 = 0.21$ respectively, and therefore one can conclude that 21% of the patients has a LOS of around 8 days. Therefore, a warm-up of 3 weeks seems to be more safe.

### Number of Replications

Next, we need to determine the number of replications (i.e., the size of our sample) such that we have enough data to properly draw inferences. We fix the run length to a period of 52 weeks excluding the warm-up period. For the calculations in this section we use the average blocking probability per week as input data, since this KPI seems empirically to be the most unstable. To determine the number of replications, we use the confidence interval method as proposed by Robinson (2014:184-186). This is an algorithm which chooses a number of replications such that the interval bounds around the mean deviate at most a pre-specified percentage (say $d$). Without going into details, the method boils down to running experiments until the width of the confidence interval is smaller than $d$ i.e., :

$$\frac{t_{n-1,1-1/2\alpha}S}{\sqrt{n}|\bar{X}|} < d \tag{5.12}$$

See Table 1 for the definition of the mathematical symbols. If we set $d = 0.05$, the number of replications is 36.

### 5.4.4. Results

In this section we present the results. We calculated the average bed occupancy and blocking probability over the week for all three scenarios, using a net run length of 52 weeks and 36 replications per experiment. For both KPIs, the 95%-confidence intervals are given, indicated with CI-Left (left interval bound) and CI-Right (right interval bound). For each scenario, we evaluated a set-up of 25 to 30 beds. The output for Scenarios 1 up to 3 is given in Tables 5.6, 5.7 and 5.8 respectively.

*Scenario 1*

| **Nr of Beds** | $\bar{B}$ | **CI-Left** | **CI-Right** | **Occupancy** | **CI-Left** | **CI-Right** |
|---|---|---|---|---|---|---|
| **25** | 7.83% | 7.53% | 8.12% | 81.39% | 81.15% | 81.63% |
| **26** | 6.12% | 5.83% | 6.40% | 79.93% | 79.69% | 80.16% |
| **27** | 4.50% | 4.21% | 4.80% | 78.22% | 77.94% | 78.50% |
| **28** | 3.21% | 2.98% | 3.44% | 76.47% | 76.20% | 76.75% |
| **29** | 2.32% | 2.09% | 2.55% | 74.60% | 74.30% | 74.89% |
| **30** | 1.57% | 1.38% | 1.76% | 72.65% | 72.34% | 72.95% |

Table 5.6.: Simulation output for Scenario 1.

*Scenario 2*

| **Nr of Beds** | $\bar{B}$ | **CI-Left** | **CI-Right** | **Occupancy** | **CI-Left** | **CI-Right** |
|---|---|---|---|---|---|---|
| **25** | 8.37% | 8.13% | 8.61% | 80.74% | 80.50% | 80.98% |
| **26** | 6.48% | 6.22% | 6.75% | 79.37% | 79.13% | 79.62% |
| **27** | 4.86% | 4.56% | 5.15% | 77.79% | 77.52% | 78.05% |
| **28** | 3.59% | 3.35% | 3.83% | 76.10% | 75.82% | 76.39% |
| **29** | 2.50% | 2.27% | 2.73% | 74.36% | 74.08% | 74.64% |
| **30** | 1.72% | 1.55% | 1.89% | 72.46% | 72.15% | 72.76% |

Table 5.7.: Simulation output for Scenario 2.

*Scenario 3*

| Nr of Beds | $\bar{B}$ | CI-Left | CI-Right | Occupancy | CI-Left | CI-Right |
|---|---|---|---|---|---|---|
| **25** | 7.24% | 6.91% | 7.57% | 82.40% | 82.18% | 82.62% |
| **26** | 5.40% | 5.09% | 5.71% | 80.69% | 80.43% | 80.95% |
| **27** | 3.97% | 3.68% | 4.26% | 78.95% | 78.66% | 79.24% |
| **28** | 2.76% | 2.51% | 3.01% | 76.97% | 76.66% | 77.29% |
| **29** | 1.85% | 1.62% | 2.07% | 74.91% | 74.61% | 75.22% |
| **30** | 1.18% | 1.01% | 1.34% | 72.86% | 72.54% | 73.19% |

Table 5.8.: Simulation output for Scenario 3.

Comparing these results with Table 5.3, we see that for all configurations and scenarios the blocking probability is lower (up to 1.88 percent point) and the bed occupancy higher (up to 2.93 percent point) if we remove the variability in elective arrivals. This proofs by example the hypothesis of Bekker and Koeleman (2011), and shows the value of reducing variability. The improvement is even better if elective arrivals during weekends are allowed (Scenario 3). In that case, we could in some cases reduce the number of beds required by 2. This is a serious improvement. Besides that, since the offered load is more equal through the week, it is expected that the workload is more even distributed.

## 5.5. Conclusion

This chapter answers the fifth research question. We used a $M_t/H_2/c/c$ queueing model to calculate the number of beds required given the target rejection rate of 3% for the MCU and 5% for the clinical ward. The required number of beds is 28 for the clinical ward, resulting in a bed occupancy and rejection rate of 74.2% and 4.6% respectively. For the MCU we need 16 beds resulting in a bed occupancy and rejection rate of 62.0% and 2.6% respectively.

Since elective arrivals can be altered, we investigated the effect of such an intervention for the clinical ward by means of a quadratic programming model and a simulation model. If we are able to level the elective arrivals over the week, we can reduce the number of beds required to 27 with a rejection rate of 3.97% and a bed occupancy of 78.95%.

# 6. Conclusion, Recommendations and Discussion

This chapter answers the research questions and solves the knowledge problem introduced in Chapter 1. Subsequently, we present our recommendations and we discuss difficulties and reflect on our work. Lastly, we underpin the relevance of our research and provide suggestions for further research.

## 6.1. Conclusion

In Chapter 1 we deducted the core problem, which was to be solved by answering several research questions. The research questions have been answered in the preceding chapters, and are repeated in this section together with a condensed answer per question. These answers together solve our core problem.

***1: What is the current performance with respect to bed capacity on both wards and MCUs?***
To answer this question, we have evaluated the bed occupancy and estimated the current rejection rate to judge the performance of both wards and MCUs. To evaluate the bed occupancy, we used the available data on hourly presence and because of erroneous data we also used Little's Law. The target occupancy is set at 85%. The bed occupancy for the MCUs in 2017 was 58.87% (D351) and 90.73% (D361), and for the clinical wards 66.67% (D340&D370) and 79.03% (D350) according to Little's Law. Due to a lack of data, the rejection rate is estimated using a queueing model. The target rejection rate is 5% for the clinical wards, and 3% for the MCUs. The rejection rate for the MCUs is estimated as 7.5% (D351) and 23.5% (D361), and for the clinical wards as 3.4% (D340&D370) and 10.9% (D350).

***2: What are the characteristics of the patient related processes?***
First, we distinguished the different type of patient arrivals. Subsequently we have performed an extensive analysis on both the patient arrival process as well as the LOS process. It turned out that the patient arrivals can be well approximated by a non-stationary Poisson process, and that especially the elective arrivals fluctuate over the day and over the week whereas the emergency arrivals seemed to be rather constant. We therefore assume piecewise constant arrival rates, varying per day part of 12 hours. We have illustrated by the Gini coefficient that the total LOS is unequally distributed over the patient population, and that the LOS distribution probably consists of multiple sub-distributions. This was confirmed by the fact that the LOS is well approximated by

a hyperexponential distribution.

***3: What are the characteristics of the clinical ward and MCU after the merger?***
To be able to predict the characteristics of the new wards, we have merged the arrival and LOS data respectively to "simulate" the new situation. Using the merging property of the Poisson process, the arrival process can be approximated by a non-stationary Poisson process, and the LOS by a hyperexponential distribution. The arrival rates again are piecewise constant, varying per 12 hours.

***4: How to correctly determine the right number of beds on wards and MCUs according to the literature?***
After reviewing a selection of articles in the field of Operations Research, we discovered that most methods used to determine bed capacity are based on Queueing Theory, Discrete Event Simulation and Mathematical programming. Besides that, almost all studies advocate that solely basing the capacity decision on a target occupancy is a bad idea, and that the focus rather should be on the blocking probability. Several models have been evaluated which led to the choice of a time depending, hyperexponential and finite server queueing model ($M_t/H_2/c/c$) based on the MOL approximation. Furthermore, a quadratic program is used to alter the elective arrivals such that a target load per day is met. Using simulation modelling one can calculate the effect of such an intervention.

***5: What is the quantitative relation between the relevant KPIs when determining the required bed capacity?***
To answer this question, we used a queueing model with time-dependent Poisson arrivals, a hyperexponential service time distribution and finite capacity. This showed us that there is a non-linear relation between the number of beds and both the bed occupancy and blocking probability. To meet the target blocking probability, we need 16 beds for the MCU and 28 beds for the clinical wards, resulting in a blocking probability and bed occupancy of 2.6% and 62% and 4.6% and 74.2% respectively. If however, the management is able to alter the elective arrivals (which only applies for the clinical ward) such that the variability in the number of arrivals per day of the week is reduced (e.g., always 3 elective arrivals on Monday), the number of beds can be reduced for the same blocking probability. If for instance we can level the elective arrivals through the week (including the weekend), we need 27 beds resulting in a blocking probability and bed occupancy of 3.94% and 78.83% respectively.

## 6.2. Recommendations

In this section we provide workable recommendations resulting from the conclusions of Section 6.1.

• We recommend to base the number of required beds for the new ward and MCU on the tables provided in Chapter 5, i.e., Table 5.2 and 5.3. These tables provide insight into the relation between the bed occupancy and rejection rate. Given the desired rejection rate of 3% for the MCU and 5% for the clinical ward, we advise a bed capacity of 16 and 28 beds respectively.
• The variability in elective arrivals should be reduced as much as possible. Ideally, the number of elective arrivals should be equal per day per week. We have shown by means of a simulation study that the number of beds on the clinical ward can be reduced to 27 for the same rejection rate and a higher occupancy.
• Capacity decisions should be based on the entire distribution of the relevant variables rather than averages (to avoid the well-known "the flaw of averages"). This coincides with the notion of many researchers that the bed capacity should be determined based on the rejection rate and not (primarily) on the bed occupancy.
• Improvements should be made regarding the available data. The number of rejections should be recorded, both within the division as for patients diverted to another division or hospital. This will make future capacity decisions more accurate. Furthermore, data should be entered more accurately. Given the already high workload of health care employees, it is not our intention to increase this even more. It could however be wise to explain to the employees why correct data matters, and that (capacity) decisions based on correct data can actually reduce the workload. Besides that, a possible solution is to automatize registration of e.g., admission and discharge time. Automatically registering data is more objective and saves valuable time.

## 6.3. Discussion & Limitations

In this last section, we critically reflect on the research process and the thesis as a whole. Possible errors and difficulties encountered will be discussed, and differences with comparable work will be underlined. We will first discuss the (in)availability of data. Subsequently we will discuss experienced difficulty when fitting statistical distributions to the arrival and LOS data. As a last subject we will advocate the practical and scientific relevance, and provide suggestions for further research.

### 6.3.1. Data Issues

#### LOS Data

The first data issue that will be discussed is the troublesome LOS data. The data provided by the cube is focused on averages rather than separate data points. For the analysis of Chapter 2, we needed separate LOS data per stay. Now however, if a patient is transferred $n$ times to the same ward within the same admission, the cube only provide the average LOS of these separate stays. To generate separate LOS points for these cases, we manually added $n-1$ extra data points with this average LOS. This most likely results in more centered data than actually is the case. This in turn results in a lower coefficient of variation and Gini coefficient which are used for the distribution fitting process of Chapters 2 and 3. Since these cases of missing data points did not occur too often, we do not think the results would be significantly different compared to the case of complete data.

Something similar occurred when analyzing the merger in Chapter 3. If for instance we consider the merger of D351 and D361, we would ideally sum the separate LOS of patients who were being transferred from the D351 to the D361 one or multiple times during the same admission. Because we do not have these separate LOS but only averages, we were unable to do so. Since the fraction of arrivals due to these transfers is relatively low, we do not expect the effect of this assumptions to be significant.

### Hourly Presence Data

The second data issue is about the presence data. This data provides the number of patients present per ward on a given hour. However, it turned out that on multiple occasions, there were more patients present than there was capacity. This does not seem to be correct, and the management has indicated that this is most likely due to input errors. This forced us to use models to evaluate e.g., the bed occupancy.

### Diverted Patients Data

The third data issue is that, as discussed in Chapter 2, there is no data on the number of patients being rejected or assigned to the wrong ward due to a fully occupied ward. This is the reason we had to assume that the number of admissions coincides with the total number of arrivals (i.e., the sum of admissions and rejected patients). Since patients are rarely diverted to other divisions or hospitals, this assumption is not too restrictive. However, patients are frequently assigned to another ward within the division (i.e., when a surgical patient is diverted to the non-surgical ward). Since we are primarily considering the situation after the merger (and the MCUs are merged), this does not cause problems for the MCUs. However, for the clinical wards this is slightly different since we are only considering 2 out of 3 clinical wards of the division. Since there is no data, we cannot say whether the number of diversions from and to this third ward cancel out against each other and thus we cannot say whether this assumption is completely justifiable or not.

### Missing Arrival Data

The last data issue encountered was the missing arrival data. Especially for the MCUs, a many were not registered. It is not clear if this means that there are no arrivals at the given day, or if the ward was closed that day. The former seems unlikely, because there are also days registered with 0 admissions. The missing data for the MCUs could be a cause for the gap between occupancy reported by the data and that calculated by the model in Table 2.7.

### 6.3.2. Distribution Fitting

Also regarding the fitting of statistical distributions several footnotes can be made. First of all, one could argue whether the Poisson assumptions for all wards is a valid one. We have tried to objectively underpin this claim by providing p-values of the $\chi^2$ goodness-of-fit test. However, especially for the D340&D370 ward one could question the validity of the Poisson assumption. It is important to realize that although the fit does not seem to perfectly follow the data, it does (roughly) have the same shape. This indicates that the Poisson model captures (most of) the variance of the arrival process, and therefore is not necessarily a bad assumption.

Secondly, the hyperexponential distribution does not seem to be a perfect fit for the LOS in both the current as well as the future situation. It should be noted that many authors experience difficulties when fitting a distribution to the LOS data. We have tried many distributions (Lognormal, Gamma, Normal) to the data, but the hyperexponential distribution provided the best fit. Besides that, it also coincides with our intuition about a separate distribution for both the short and long stay patients. For the clinical ward, the model seems to overestimates respectively underestimates the frequency of short stay and medium stay patients compared to the data. This could result in a slightly lower bed demand given the same number of admissions. For the MCU, the model underestimates short stay patients a bit but overestimates medium stay patients which could result in a slightly higher bed demand than actually is the case. Most likely the data is hard to fit because many patients groups within the population are likely to have a different LOS distribution. To discover all these groups, one could use e.g., Classification and Regression Trees as suggested by Costa et al. (2003). This however falls outside the time window of this assignment.

## 6.4. Relevance and Future Research

In this section we argue the scientific and practical relevance of this research and provide possible openings for future research.

### 6.4.1. Scientific Relevance

This research is based on a continuation of a line of research in bed capacity planning. Especially work of De Bruin (2010), Bekker and Koeleman (2011) and Bekker and De Bruin (2010) was used. We have modelled the arrival and service process as suggested by De Bruin (2010), and thus supported his analysis by another case. Besides that, we have applied the queueing model proposed by Bekker and Koeleman (2011) and Bekker and De Bruin (2010) and have shown that also within our case a time dependent queueing model seems appropriate. On top of that, we have applied the quadratic programming model suggested by Bekker and Koeleman (2011) and proved its value by means of a simulation study. To the best of our knowledge, a combination of the above techniques has not been applied before which directly confirms the scientific relevance. Buter (2017) shares part of our analysis but focuses on distributing medical disciplines over wards and hence uses a less sophisticated queueing model to determine capacity. Furthermore, Buter (2017) uses an extensive simulation study to evaluate different set-ups regarding the distribution of medical disciplines, whereas we used simulation modeling to show the possible gain of reducing variability in elective arrivals.

### 6.4.2. Practical Relevance

Using the analyses, the management now knows how to properly determine the number of beds required. The management has gained insight into the quantitative relation between the bed occupancy and blocking probability. We have shown that a target occupancy of 85% is unreasonable, and that the blocking probability should be the leading KPI in bed capacity decisions. This insight decreases the number of diverted patients, and thus increases the quality. Besides that, we have shown the value of altering the elective arrivals which could increase the bed occupancy (i.e., productivity) without increasing the blocking probability. Our research thus could lead to an increase both quality of care as well as productivity, from which we conclude that our research is practically relevant.

### 6.4.3. Future Research

This research quantifies a strategic decision (see Section 1.2.4); using the analyses provided the management can choose a (maximum) number of bed positions for the long term. However as indicated before, the operational management consisting of head nurses discuss on a day-to-day basis how many beds to “close” on a certain day. In future research, it could be helpful to develop a tool supporting this decision.

Besides that, in Section 5.4.4 we have showed the value of rearranging the elective admissions. A next step would be to investigate how such an intervention would look like in practice.

A more theoretical starting point for future research would be to develop an analytical model to optimize the bed occupancy for a multi-echelon chain of e.g., wards or even hospitals instead of single wards. In supply chain it is well-known that optimizing the chain instead of single echelons at a time is considerably more difficult, but could result in much better performance than optimizing each echelon sequentially. It is not unreasonable to assume that this will also be the case in healthcare.

# References

Adan, I & Resing, J. (2015) *Queueing Systems*, lecture notes, Queueing Theory, Landelijk Netwerk der Mathematische Besliskunde, March 26,2015

Bekker, R., & Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science, 14(3)*

Bekker, R. & De Bruin, A.M. (2010) Time-dependent analysis for refused admissions in clinical wards *Annals of Operations Research, 178:45-65*

Belciug, S., & Gorunescu, F. (2015). Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. *Journal of Biomedical Informatics, 53, 261-269.*

De Bruin, A. M., Bekker, R., van Zanten, L., & Koole, G. M. (2010). Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research, 178(1), 23-43.*

Buter, R. (2017). *Het reduceren van variatie in bedbezetting door het toewijzen van specialismen aan verpleegafdelingen* (Bachelor Thesis). Retrieved from `https://essay.utwente.nl/72981/`

Cochran, J. K., & Roche, K. (2008). A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society, 59(11), 1471-1482.*

Costa, A. X., Ridley, S. A., Shahani, A. K., Harper, P. R., De Senna, V., & Nielsen, M. S. (2003). Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia, 58(4), 320-327.*

Devapriya, P., Strömblad, C. T. B., Bailey, M. D., Frazier, S., Bulger, J., Kemberling, S. T., & Wood, K. E. (2015). *StratBAM: A Discrete-Event Simulation Model to Support Strategic Hospital Bed Capacity Decisions. Journal of Medical Systems, 39(10).*

Gini, C. (1912). Variabilità e mutabilità contributo allo studio delle distribuzioni e delle relazione statistiche. *Studi Economico-Giuredici dell'Università di Cagliari 3 (Part 2),i-iii,3-159.*

Gorunescu, F., Millard, P. H., & McClean, S. I. (2002). A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society, 53(1), 19-24.*

Green, L., & Kolesar, P. (1991). The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science*, 37(1), 84-97

Green, L. V. (2002). How many hospital beds? *Inquiry, 39(4), 400-412*

Harper, P. R., & Shahani, A. K. (2002). Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society, 53(1), 11-18.*

Hans, E. W., van Houdenhoven, M., & Hulshof, P. J. H. (2011). *A framework for health care planning and control.* (Memorandum / Department of Applied Mathematics; No. 1938). Enschede: Department of Applied Mathematics, University of Twente

Heerkens, J.M.G. & Van Winden, A. (2012). *Geen Probleem.* Nieuwegein: Van Winden Communicatie

Jagerman, D.L. Nonstationary blocking in telephone traffic (1975), *Bell System Technical Journal*, 54(3), 625-661

Law, A.M. (2015) *Simulation Modeling and Analysis. (5th ed.)* New York: McGraw-Hill education

Little, J. D. C. (1961). A Proof for the Queuing Formula: $L = \lambda W$. *Operations Research.* 9 (3), 383387

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association, 9, 209-219.*

Massey, W.A. (2002). The Analysis of Queues with Time-Varying Rates for Telecommunication Models. *Telecommunication Systems*, 21(2-4), 173-204

Robinson, S. (2014) *Simulation: The practice of model development and use. (2nd ed.)* Palgrave Macmilllan

Winston, W.L. (2004).*Operations Research: Applications and Algorithms. (4th ed.)* Belmont:Brooks/Cole Cengage Learning

Young, J.P. (1965). Stabilization of inpatient occupancy through control of admissions. *Journal of the American Hospital Association, 39,41-48*

# A. Problem Cluster

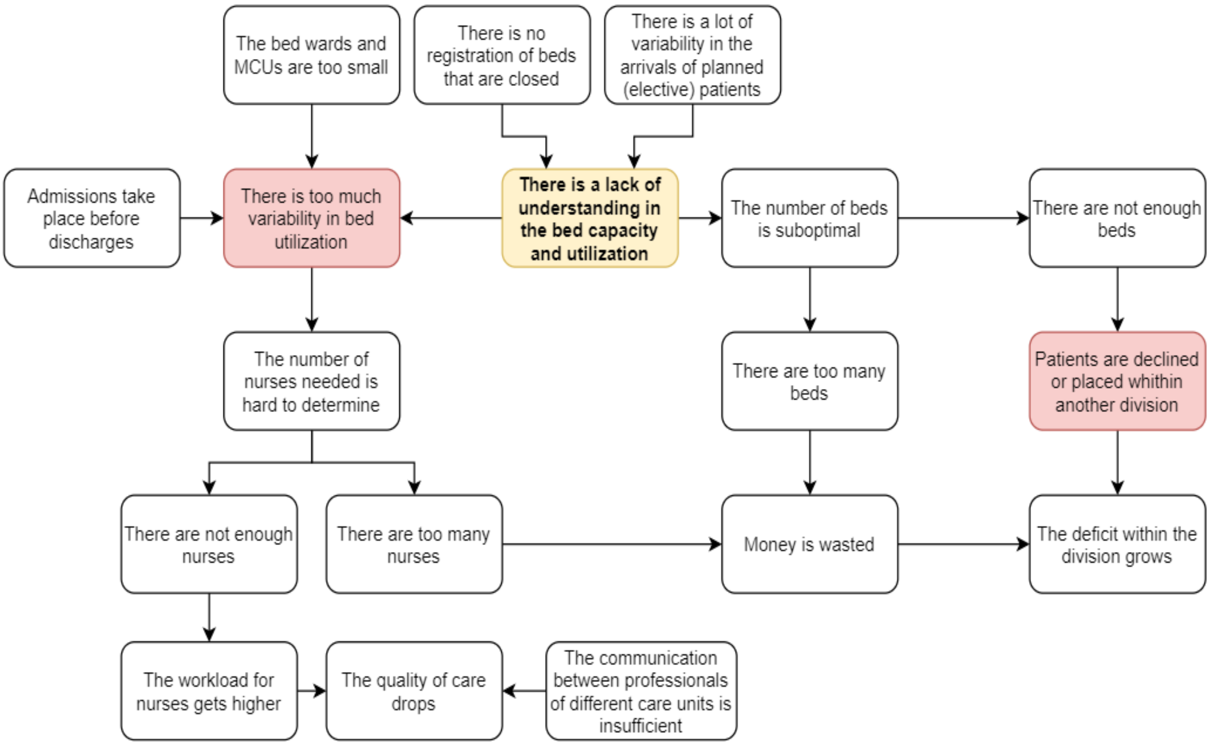


Figure A.1.: Problem cluster

# B. Probability distributions and MLE

## B.1. Poisson distribution

The Poisson dsitribution is a discrete probability distribution, with:

$$\begin{aligned} pdf \;&:\; P(X=k) = e^{-\theta}\frac{\theta^k}{k!} \\ mean &: \theta \\ variance &: \theta \end{aligned}$$

## B.2. Exponential distribution

$$\begin{aligned} pdf \;&:\; f(x_i,\theta) = \theta e^{-\theta x_i} \\ mean &: \frac{1}{\theta} \\ variance &: \frac{1}{\theta^2} \end{aligned}$$

## B.3. Hyperexponential distribution

$$pdf \;:\; H_k(p_1,...,p_k;\mu_1,..,\mu_k) = \sum_{i=1}^{k} p_i f(x_i,\mu_i),\; p_i \in [0,1], \sum p_i = 1$$

with $f(x_i,\mu_i)$ being the pdf of an exponential distribution with parameter $\mu_i$.

$$\begin{aligned} mean &: \sum_{i=1}^{k}\frac{p_i}{\mu_i} \\ variance &: \left[\sum_{i=1}^{k}\frac{p_i}{\mu_i}\right]^2 + \sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j \left(\frac{1}{\mu_i}-\frac{1}{\mu_j}\right)^2 \end{aligned}$$

## B.4. MLE Poisson

In maximum likelihood estimation, we try to optimize the likelihood function $L(\theta_1, ..., \theta_m, x_1, ..., x_n)$. This answers the question "What is are the values of the parameters $(\theta_1, .., \theta_m)$ that best explain the data $(x_1, ..x_n)$?". In case of the Poisson distribution, this boils down to:

$$\max_{\theta} L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} f(\theta, x_i) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \qquad \iff$$

$$\max_{\theta} ln(L(\theta, x_1, ..., x_n)) = -n\theta + ln(\theta) \sum_i x_i - \sum_i ln(x_i!) \qquad FOC$$

$$n = \frac{\sum_i x_i}{\theta} \iff \hat{\theta}_{MLE} = \overline{X}$$

# C. Arrivals per day

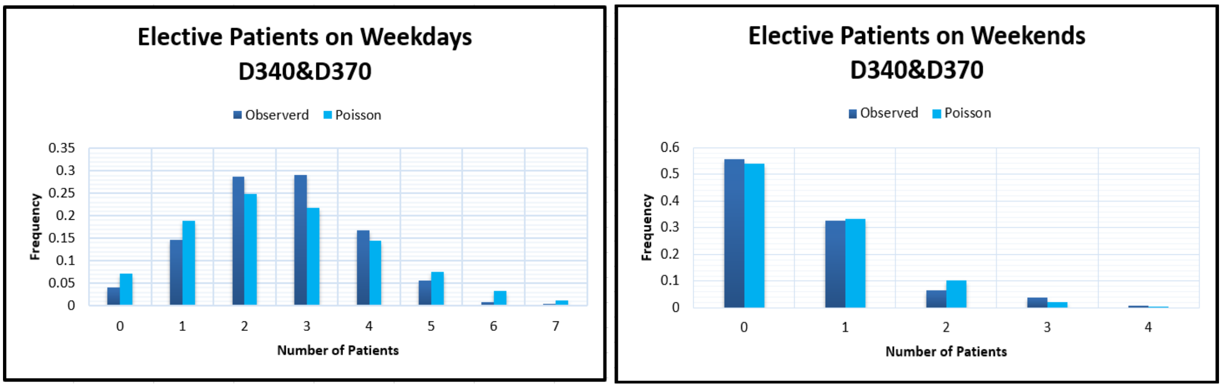


Figure C.1.: Elective arrivals split into week- and weekend days for the D340&D370 ward in 2017. *Data retrieved from data cube, n=251 and n=104 respectively.*

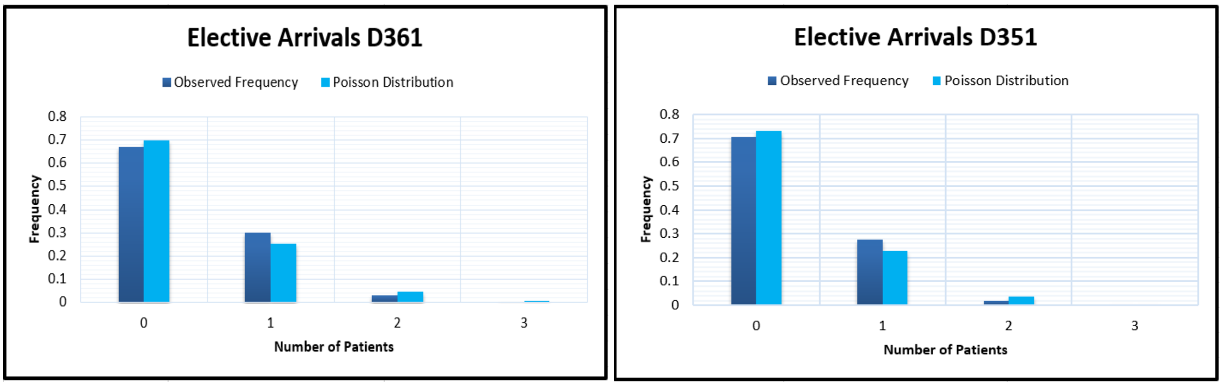


Figure C.2.: Elective arrival patterns for both MCUs in 2017. *Data retrieved from data cube, n=133 and n=153 respectively.*

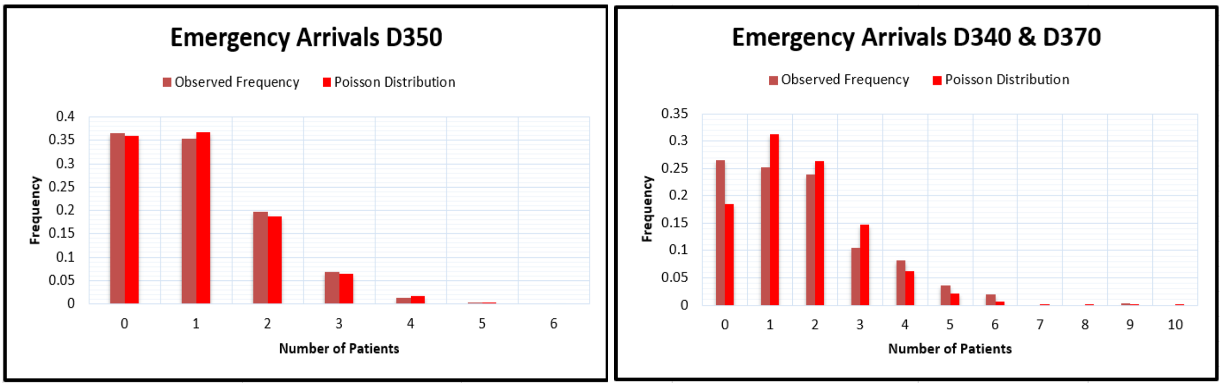


Figure C.3.: Emergency arrivals for both clinical wards in 2017. *Data retrieved from data cube, n=373 and n=616 respectively.*

# D. Arrivals per day

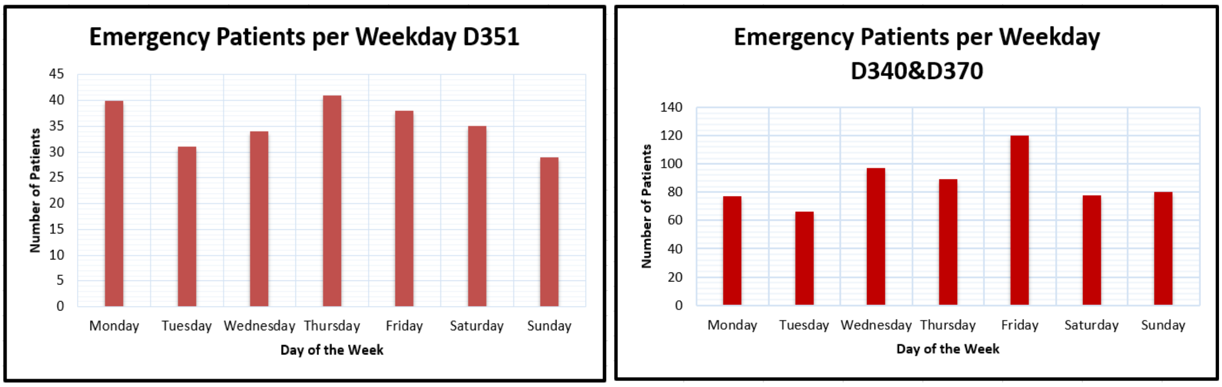


Figure D.1.: Emergency arrivals for the D351 MCU and D340&D370 clinical ward in 2017.*Data retrieved from data cube, n=248 and n=616 respectively.*

# E. Arrivals per hour

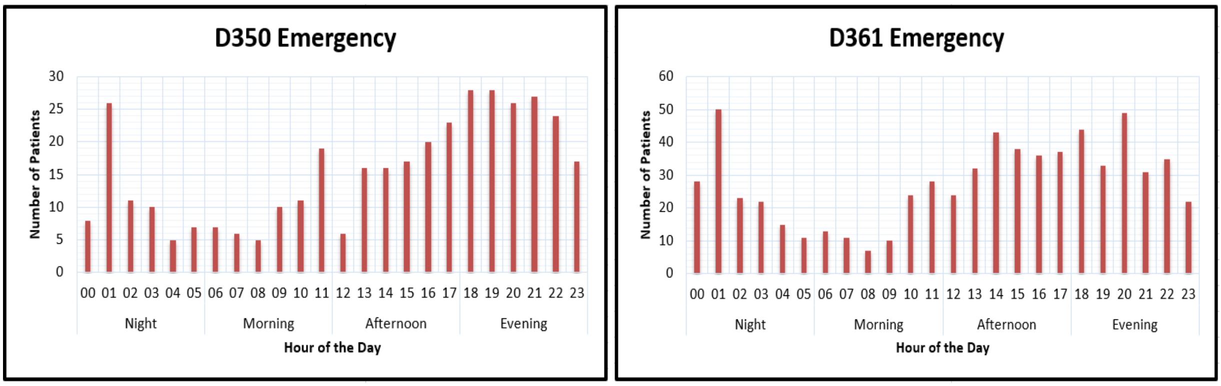


Figure E.1.: Hourly admissions of emergency patients for all wards in 2017. *Data retrieved from data cube, n=373 and n=666 respectively.*

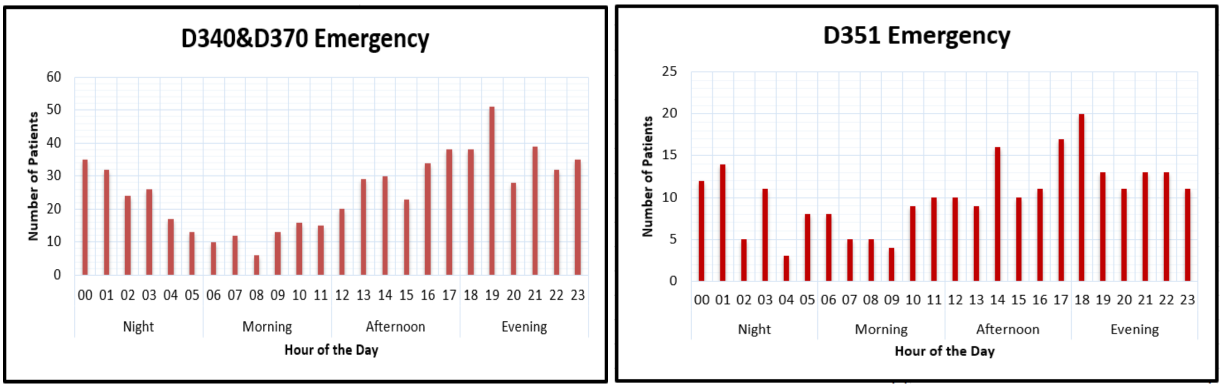


Figure E.2.: Hourly admissions of emergency patients for all wards in 2017. *Data retrieved from data cube, n=616 and n=248 respectively.*