Radboud universitair medisch centrum
Servicebedrijf, adviesgroep Procesverbetering & Implementatie

Postbus 9101, 6500 HB Nijmegen
F.C. Donderslaan 2

www.radboudumc.nl

On determining a minimum and maximum arrival rate to decrease overcrowding at a nursing ward

L.A.C.P. Baars

Date

June 13, 2019

Supervisory committee

Prof. Dr. Ir. E.W. Hans
Dr. Ir. A.G. Maan-Leeftink
Ir. B. van Acker
Ir. W. Winasti

Management summary

Context and motivation of this research

This research focuses on the bed capacity problem of the combined urology (URO) and gynaecology (GYN) nursing ward at the Radboud university medical centre (Radboudumc) in Nijmegen, The Netherlands. The Radboudumc is a large hospital, where more than 100,000 patients arrive annually. The focus of this research is on the planned patients. Unplanned patients, who mostly come in through the emergency department, are outside the scope of this research.

The mismatch in demand for, and capacity of available beds can be large. The Radboudumc recently implemented Real-Time Demand Capacity management (RTDC) method, which is expected to provide a partial solution to this problem. The RTDC method involves a meeting between all head-nurses at 9 AM to discuss arrivals to, and discharges from the hospital. In this meeting, the foreseen bed-capacity issues are discussed and a specific course of action to act upon these issues is formulated.

To further tackle the problem of overcrowding of the wards, the Radboudumc wants to know to what level they should reduce the variation in the arrival rate at the wards in order to reduce the expected overcrowding, when demand for beds at a ward is larger than capacity. This research aims to gain insights into how much variation in elective patient arrivals at the wards need to reach an expected level of overcrowding of no more than 2% In order to do so, a simulation model is developed in Microsoft Excel, in which different possible configurations for the variation in arrival rate and the number of discharges before 2PM at the ward can be explored.

Approach

After reading through the literature on arrival rate and length of stay (LOS) distributions at hospital wards, this research concludes that the found distributions cannot be applied to the URO/GYN ward. Therefore, a stochastic discrete event simulation model with a general arrival rate distribution and the historical length of stay (LOS) distribution was developed. The Normal and the Triangular distribution are chosen for the arrival rate because these distributions can accommodate a decrease in the distribution's variation. The LOS is modelled as the number of days a patient is likely to stay (from historical data). It further incorporates the possibility to explore the effect of discharging patients earlier through a decrease in the LOS by increasing the percentage of patients discharged before 2 PM.

The simulation model is developed in Microsoft Excel, since the organisation is familiar to this program and it is widely available. The developed simulation model uses Visual Basic for Applications (VBA) to generate each patient's LOS and the number of arrivals. Although the end-user of the model, the department Process Improvement and Implementation (PVI), is unfamiliar with VBA, the threshold of learning it is lower than learning to use a completely new application. The literature further confirms that a spreadsheet program is a good way to model the patient flow of a nursing ward.

Results

The simulation model shows that for the URO/GYN ward it is theoretically possible to reduce the overcrowding to 2%. In other words, the Radboudumc can achieve a 98% possibility to place an arriving patient at the right nursing ward. This can be achieved by reducing the variation of the arrival rate to 75% of the historical variation and discharge 62% of patients before 2 PM or decrease the arrival rate variation to 50% and discharge 52% of patients before 2 PM. The relationship between these variables appears to be linear. Less overcrowding is expected to result in less stress for the

nursing staff at the ward, better care for the patient and thus a higher patient satisfaction. Since the ward operates more efficiently, it is further expected that less patients need to be cancelled due to overcrowded wards.

Conclusions

This model gives insight into the effect of the arrival rate variation and the discharges before 2 PM on the overcrowding of a hospital ward. It provides the surgery schedulers of the Radboudumc with guidelines for the number of patients that they can schedule during a working day, and at the same time, shows the necessity of early discharges. PVI can now set a quantified goal and its role is to facilitate change by showing the need for change to the scheduling department and the ward, supporting the change by suggesting process improvement steps, and controlling on the improvement process such that progress is continuous and maintained.

Scientifically, this research suggests a new way of modelling patient flow that is not exact, but sufficient for practical purposes. It gives insight in the effect of the RTDC principle of early discharges and the combination of early discharges with less arrival rate variation in planned patients. Furthermore, this research presents two cases in which the arrival rate and LOS cannot be modelled according to any distribution commonly found in the literature on hospital ward arrival rates and LOSs. Lastly, this research confirms the notion in earlier studies that patient flow can be well-modelled in a spreadsheet program.

Limitations of this research are the following: 1) There is a clear distinction between discharging a patient before or after 2 PM while in practice the difference may be as little as one minute. 2) The arrivals at the ward cannot truly be considered as random arrivals, since the arrival rate standard deviation per week is smaller than the expected standard deviation per week if the arrivals would be random. This does however not appear to be of great influence on the model. 3) The model in Microsoft Excel appeared to get stuck for no apparent reason on a few occasions. Further research should focus on validating the relationship found between discharging more patients before 2 PM and decreasing the variation in arrival rate yields the result suggested by this model.

Outlook

Based on the results we have three suggestions for the Radboudumc concerning their planning process and the use of this model. First, since it is noted that when URO arrivals peak, GYN arrivals do too, we suggest to ensure communication between the URO and the GYN planning department such that arrivals are more evenly spread over the year and the specialisations do not plan a peak number of surgeries at the same time.

Second, we observe that the effect of discharging patients earlier, i.e. before 2 PM, has a larger effect on the overcrowding than decreasing the variation in the arrival rate does. Furthermore, the effect of reducing the variation in the arrival rate has diminishing returns. The optimal arrival rate variation for URO patients is 60% (both Triangular and Normal distributed arrival rate), and 60% (Triangular distributed arrival rate) or approximately 65% (Normal distributed arrival rate) for GYN patients.

Third, we suggest using this model with the Normal distribution for the arrival rate. The results from the Triangular distribution and the Normal distribution do not differ much, while the Normal distribution is both easier to use and gives smoother results which means it is easier to draw conclusions from the data.

The next step for PVI is to identify an improvement goal based on the results from this research and use the Plan, Do, Check, Act cycle to start the improvement process, in collaboration with all stakeholders.

Preface

Before you lies my master thesis, an advisory report on capacity management for the Radboudumc. This thesis is my final step in completing the master studies Industrial Engineering and Management at the University of Twente.

For somewhat over 10 months, three days a week, I have worked with great pleasure at the department PVI, the consulting group Process Improvement and Implementation (in Dutch: Adviesgroep Procesverbetering en Implementatie). It has not always been easy, breaking my head over scoping the research, structuring the thesis, and mathematical modelling issues.

When I needed someone to spar with, I could always rely on my supervisor Bart at the Radboudumc. He was a great help and always ready to listen to my thoughts. Windi also helped me much, especially in completing my literature review and the statistical part of this thesis. I am certain that, had I reached out more often besides our bimonthly Skype meetings, professor Hans would have made time to answer my questions. Nevertheless, he was a great supervisor giving me the feedback I needed to overthink my writing again.

I furthermore thank my family and friends. My family for always being interested in my work and progress although I wasn't always so happy to talk about it, and my friends for keeping me, as we say in Dutch, 'with one leg' in the student life while I was already living in a different city and working all day. Thank you all for your support.

Enjoy!

Laurens Baars

Nijmegen, June 13, 2019

Contents

Glossary

ED	-	Emergency Department.
Elective patient	-	Patient that is scheduled ((s)he has an appointment) to have surgery or a diagnostic appointment.
Emergency patient	-	Patient that comes into the surgery schedule through the Emergency Department or is not scheduled for a surgery or diagnostic appointment.
EPDS	-	Electronic patient database system.
ICU / MCU	-	Intensive Care Unit / Medium Care Unit.
Inpatient	-	Patient that is scheduled to remain in the hospital for at least one day and night.
LOS	-	Length of Stay of a patient, the number of days a patient will remain in the hospital.
Off-service patient	-	Patient receives care outside the ward assigned to treat their illness or condition.
OR	-	Operating room.
Outpatient	-	Patient that comes in and is discharged the same day.
Overcrowding	-	The situation where the demand for beds is higher than the available capacity.
PVI	-	The department Process Improvement and Implementation at the Radboudumc

1. Introduction

This research focuses on the bed capacity problem in the Radboud university medical centre (Radboudumc) in Nijmegen, The Netherlands, as the mismatch in demand for and capacity of available beds can be quite large. The newly implemented Real-Time Demand Capacity management (RTDC) method is expected to provide a partial solution to this problem. To further tackle the problem of overcrowding of the wards, the Radboudumc wants to know to what level they should reduce the variation in the arrival rate at the wards in order to reduce the expected overcrowding, when demand for beds at a nursing ward is larger than capacity. Using RTDC, the hospital expects to be able to mitigate the remaining overcrowding.

Section 1.1 places this research in context. Next, section 1.2 describes the challenge addressed in this research. Following, section 1.3 describes the goal and section 1.4 defines the scope in which this research is conducted. Section 1.5 discusses the scientific and practical relevance of this research. Finally, the research methodology and layout are depicted in section 1.6.

1.1. Research context

This section places the research in context, i.e. it gives a description of the Radboudumc and the general problem of overcrowding.

1.1.1. The Radboudumc

The Radboudumc is a university medical centre for patient care, scientific research and education, based in Nijmegen, The Netherlands. It was founded in 1956 as the Sint Radboudziekenhuis (St Radboud hospital). Later, the name changed to University Medical Centre St Radboud, before it got the current name Radboudumc in 2013 (https://www.radboudumc.nl/over-het-radboudumc/geschiedenis/geschiedenis, last accessed: 13-12-2018). It is now a large urban, tertiary care hospital with over 1,000 beds and almost 11,000 employees. Every year, over 100,000 new patients arrive, and 35,000 surgeries are performed (https://www.radboudumc.nl/over-het-radboudumc/spreekbeurt/over-het-radboudumc/grootte, last accessed: 13-12-2018).

This research is commissioned by the Radboudumc department PVI. PVI is an internal consultancy group which advices other hospital departments on their operational, tactical, and strategic processes. Approximately 45 consultants and administrative employees work in the consultancy group (https://www.radboudumc.nl/over-het-radboudumc/organisatie/organisatieonderdelen/adviesgroep-pvi/pvi-team, last accessed: 13-12-2018).

As will be further addressed in section 1.4, this research focuses on two specialisations of the department C5West, namely urology (URO) and gynaecology and obstetrics (GYN). Patients for these two specialisations are put together in one ward of 38 beds, but surgeries for both are planned separately. From September 2017 through August 2018, the URO specialisation treated 2158 patients and the GYN specialisation treated 1403 patients (data obtained from the electronic patient database system (EPDS)).

1.1.2. Overcrowding at hospital wards

Hospitals, including the Radboudumc, are under constant pressure to improve their operational methods and efficiency, as demand for hospital services increases, customers (patients) become more demanding, and the yearly budget increases are lowered (Bachouch, Guinet, & Hajri-Gabouj, 2012; Burdett & Kozan, 2016; Lega & De Pietro, 2005; Lovett, Illg, & Sweeney, 2016; Moldovan, 2018; Villa, Barbieri, & Lega, 2009). Hence, hospitals must find new ways to organise their processes in order to improve their efficiency and level of care.

The number of patients arriving at the nursing wards of the Radboudumc differs significantly from day to day and hour to hour. It happens almost every day that more than one ward has more demand for beds than available capacity. Overcrowding puts a large level of stress on the nursing staff that must find a suitable location for the patient on the spot.

The cause of variation in arrivals lies mainly with the planning of the operating rooms (ORs). The planning department does not consider the effect their planning has on the crowding of the wards. Their planning is not communicated to the wards, which means the wards do not know how many beds to open (i.e. how many staff to schedule) on any day. Currently, the number of beds is fixed to accommodate this uncertainty. Optimising the surgery schedule is considered a step too far, however, and having too many beds open (overcapacity) is expensive. Furthermore, a ward cannot simply place extra beds, since that would require extra nursing staff of which there is already a shortage.

1.2. The challenge

To manage the issue of overcrowding, the Radboudumc has implemented the Real-Time Demand Capacity (RTDC) method, described in Resar, Nolan, Kaczynski, & Jensen, (2011). This method, which will be more elaborately described in section 2.2.3, essentially brings the mismatch problems from the wards to the bed-meeting each day at 9 AM, where the head nurses sit together to identify capacity problems and define a solution for any foreseen capacity problems between 9 AM and 2 PM. The solution often involves discharging another patient earlier during the day or relocating patients to other wards or hospitals.

The RTDC method, however, does not provide full mitigation of overcrowding caused by the variation in arrivals. There is no formally defined amount of overcrowding that can be mitigated using RTDC, because Resar et al. (2011) have described RTDC very operationally. That is, only practices and results are presented and there is no (mathematical) model of the method. This research assumes that RTDC can mitigate 2% overcrowding. The current level of overcrowding is higher and hence the challenge to be tackled in this research is to gain insights into how much variation in elective patient arrivals at the wards need to be reduced in order to reduce the overcrowding to 2%.

1.3. The research goal

This research focuses on finding the amount of variation in arriving patients at wards which can still be managed by RTDC, given the case mix of the Radboudumc. As said earlier, the main problem lays with the operating room schedules. Although it is considered infeasible to develop a strict planning framework, limits can be set in terms of the number of surgeries per day or week within which the schedulers must plan surgeries in order to restrict overcrowding downstream. Any variation within these limits can then be corrected using RTDC.

Therefore, the main research question is defined as

> *"How much must the variation in demand for hospital beds, in terms of the number of arrivals, be reduced, such that 98% of arriving patients can be placed in a ward of primary or secondary choice?"*

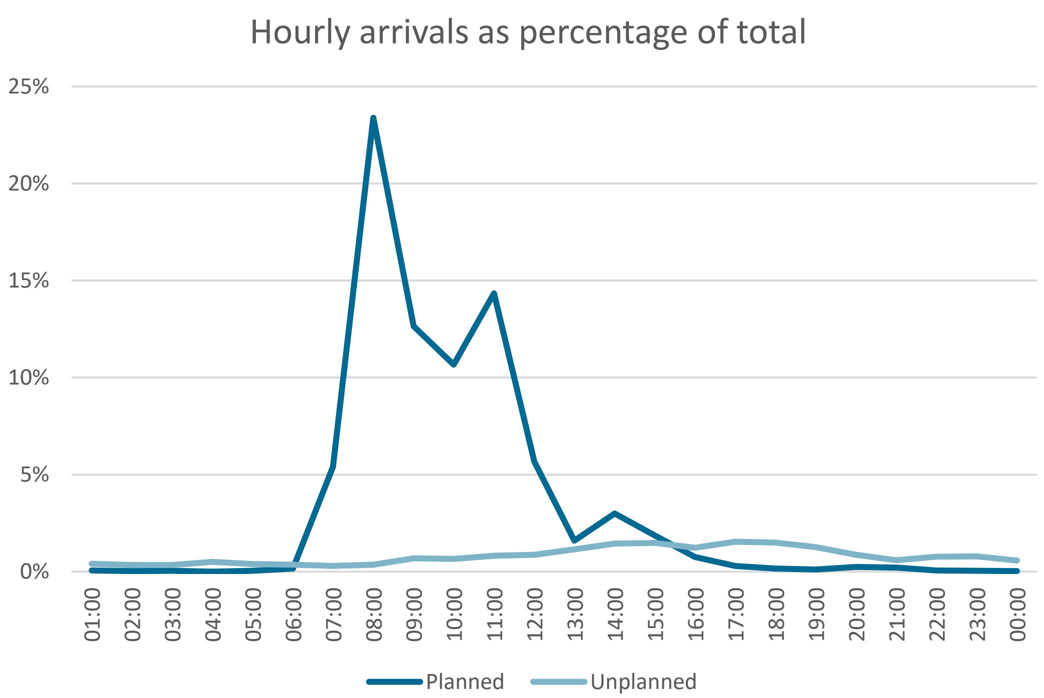
The goal of this research is to develop a model that facilitates exploration of the effect of different parameters on the level of overcrowding at different wards of the Radboudumc. These parameters are the variation in the number of arrivals for the two specialisations under study, and the accompanying percentage of discharges before 2 PM.

1.4. Scope of the research

This research focuses on two specialisations (URO and GYN), which both are managed in one ward. The model as described in the goal of this research will therefore focus on one (combined) ward, with the purpose of being generalisable if the inputs are adapted to other specialisations or wards. The ward of primary or secondary choice, as defined in the main research question, is in this research one ward where URO and GYN patients are placed, since we consider a fixed number of beds for URO patients and a fixed number for GYN patients. For example, URO patients are primarily placed in an URO bed, and if such a bed is unavailable, they are placed in a GYN bed. As RTDC is focused on freeing up beds by discharging patients before instead of after 2 PM, this is another variable considered in this research.

Another distinction to be made is between planned and unplanned patients. In the hospital, the difference between elective and emergency patients is not the same as the difference between planned and unplanned patients. Emergency patients arrive through the ED and are always unplanned, but not all unplanned patients are emergency patients. In consultation with several PVI consultants, it was chosen to distinguish between planned and unplanned patients, as unplanned patients do not have an appointment to be seen (for example by a physician). Unplanned patients will not be neglected, however, as they will be treated as a constant factor. In the model, a constant number of beds available is kept available for unplanned patients. Any future figure or graph will display only planned patients, unless otherwise indicated.

Two distinctive timeframes during the day are considered. The first is from midnight to 2 PM, and the second timeframe is from 2 PM until midnight. This separation follows the RTDC idea, where it is advocated to discharge patients before 2 PM to generate more capacity for arriving patients. Over 95% of admissions of URO and GYN patients fall in the first timeframe. See Graph 1. In this graph the planned and unplanned patients are displayed as a percentage of the *total* arriving patients. For the horizontal axis holds: the arrivals displayed at for example 7 AM are the arrivals between 6 AM and 7 AM.



Graph 1. Hourly arrivals as percentage of total arrivals (n=4408, t=365 days, source: EPDS)

1.5. Relevance

This research has both practical and theoretical implications. Practical implications are, firstly, the suggested scheduling limits for the URO and GYN specialisations. Based on the suggestions for improvement identified in this research, the Radboudumc should be able to decrease overcrowding at the URO/GYN ward to 2%. Decreasing the overcrowding increases the level of care for the patients and therefore increases patient satisfaction. Furthermore, if fewer patients are off-serviced to other wards, that frees up space for patients that should rightfully be placed in those wards. Hence, decreasing overcrowding at one ward has implications for related wards.

Second, this research provides a quantitative analysis of the effect of discharging more patients before 2 PM, as suggested by the RTDC methodology. Lastly, this research provides the Radboudumc with a decision-making tool not only for the URO/GYN ward, but for every ward at the hospital because adaptations are relatively easy.

This research further contributes to the literature by suggesting a simulation methodology that, as far as we know, has not been practiced before in a health context. Furthermore, this research quantifies, theoretically, the effect of discharging patients earlier and is therefore a step towards quantifying the separate components of RTDC.

1.6. Research approach

This section addresses the structure of the research by first explaining the used methodology and the accompanying research questions, and second by providing the reader with the research layout.

1.6.1. Methodology and research questions

In order to reach the research goal, several research questions are composed, which fit in the DMAIC improvement cycle. DMAIC is short for Define, Measure, Analyse, Improve, and Control. These concepts and their implication in this research are explained shortly:

Define - The problem is defined, and critical characteristics of the hospital operating procedures are described mainly through interviews with key personnel.

Measure - Data necessary for analysis are identified. In this research, this phase entails finding the necessary information for developing a simulation model through a literature review.

Analyse - The current situation is described quantitatively through data analysis. The data gathered is then used to develop a simulation model, and this model is validated with historical data.

Improve - (Theoretical) improvements are proposed based on the simulation model.

Control - A consolidation is proposed such that the improvements can be maintained in the future, after the actual implementation of the proposed improvement method.

The research questions are as follows:

Define

1. What are the current patient flow practices at the Radboudumc? [Chapter 2]
    a. What are the current patient flow characteristics?
    b. How are patients routed through the hospital?
    c. What are the key performance indicators concerning patient flow?
    d. What is Real-Time Demand Capacity management and what is the goal?

Measure

2. What methodologies can be used to best model a hospital unit in terms of patient flow? [Chapter 3]
    a. What is written in literature about managing and modelling patient flow at a tactical and an operational level?
    b. What are the advantages of using simulation, queuing theory, or Markov chains?
    c. What are often-found distributions for the arrival rate, and how can the time-dependent arrivals at hospital wards be modelled?
    d. How can the length of stay at hospital wards be modelled?

Analyse

3. How can the current situation be modelled, and arrival boundaries be identified? [Chapter 4]
    a. How can the arrival rate and length of stay distributions be modelled?
    b. How can the effect of the arrival rate variation and early discharges be modelled?

Improve

4. How much must the influx of patients at the wards in the Radboudumc be changed to mitigate demand surplus to the desired level? [Chapter 5]
    a. What is the effect of the variables on the patient flow?
    b. Which variable has, or combination of variables have, to be altered to reach the desired KPI values and distributions?

Control

5. How can the proposed solution be generalised to other units and hospitals? [Chapter 6]

1.6.2. Research layout

This research consists of eight chapters. Chapter 1 is the introduction, where the problem is defined in its context. Chapter 2 addresses the first research question. In this chapter, the underlying processes at the hospital are explained such that the reader gains a more in-dept insight of the situation this research is placed in. Furthermore, some general observations about the current patient flow of the two specialisations under study are made.

Chapter 3 contains the literature review. Several solution approaches toward modelling patient flow at a hospital unit are obtained from the literature. Important aspects here, are how to model time-dependent arrivals, and which distributions are often found for modelling arrivals and patient's length of stay at wards. These distributions serve as suggestions to test the data of the Radboudumc against.

Chapter 4 focuses on modelling the patient flow of the two specialisations under, and identifying the inflow boundaries by first testing the data from the hospital against possible distributions for the arrival rate and LOS. Once these are identified, the model can be completed in a way that fits this research best. Modelling RTDC and integrating this model in the model for the hospital unit identifies the inflow boundaries.

Chapter 5 presents the findings from running the model described in Chapter 4. It furthermore describes the implications for PVI, the primary user of this research and the model.

Chapter 6 describes how this research at the Radboudumc can be generalised to other hospitals, university medical centres or not. The aim of the fifth research question is to deliver a method to discover the extent of the current mismatch between demand for and capacity of hospital beds. This method needs to be easily adaptable to other hospital situations by changing input. If the model is written for one department within the Radboudumc and can be adapted to another department, it is likely to be usable by other hospitals as well.

Chapter 7 contains the conclusion, the discussion and limitations, and suggestions for further research. The output of this research is a model that provides insight into the effect of earlier discharges and a decrease in arrival rate variation, developed in an application familiar to the Radboudumc. It must be familiar in order to ensure that responsible employees are able to adapt it to new situations with new data. This model is designed to be used by PVI to find the effect of a decrease in arrival rate variation and more early discharges for any ward at the Radboudumc.

2. Patient flow

This chapter gives an overview of how patient flows are managed at the Radboudumc. The focus of this chapter, and the rest of this research, is the two specialisations URO and GYN. Data is drawn from the EPDS of the Radboudumc.

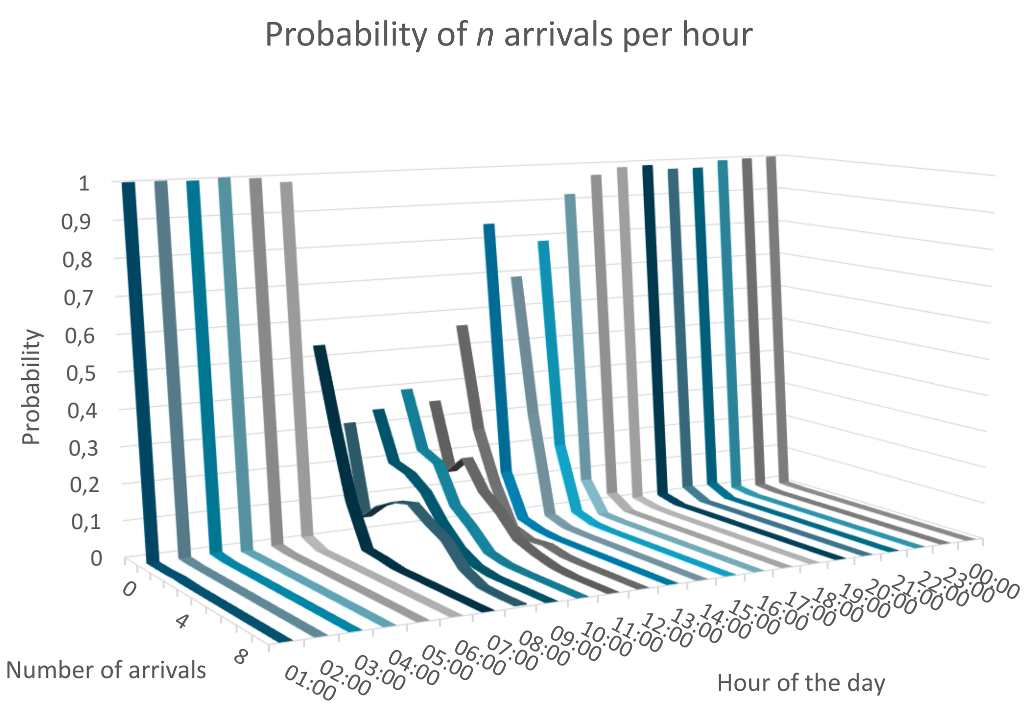
Section 2.1 describes the arrival rate, LOS, and available beds at the ward under study. Section 2.2 describes how patient flow is managed and includes a description of the RTDC management approach. Section 2.3 lists and quantifies key performance indications for the patient flow. Finally, section 2.4 concludes this chapter.

2.1. Current patient flow characteristics

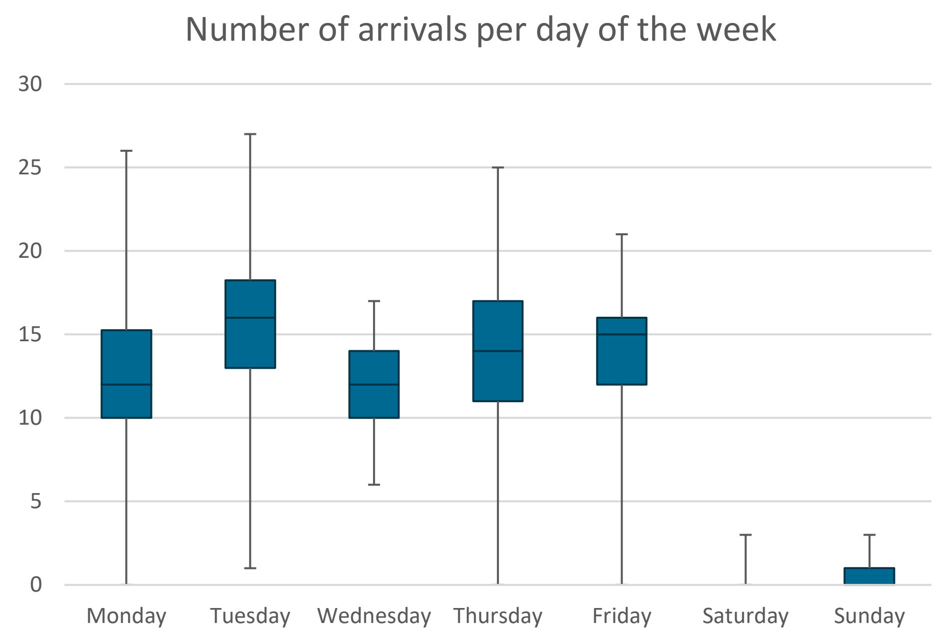
In this section, some general observations about the (planned) patient flow in terms of the number of arrivals, LOS, and discharge time will be made. A more thorough analysis is conducted in the Analysis phase, Chapter 4.

2.1.1. Patient arrival situation

The current patient arrival pattern at the URO/GYN ward is extremely variable. An example is a day where 7 patients arrive between 7 and 8 AM, 0 patients between 8 and 9 AM, 4 patient an hour later, and again 6 new patients between 9 and 10 AM. In the most extreme case, 10 patients arrive in one hour (data from 1-9-2017 until 31-8-2018). In Graph 2 the probability that *n* patients arrive in any time frame is displayed, and Graph 3 displays the variation in the number of arrivals per day of the week.



Graph 2. Probability of the number of arrivals per hour of the day (n=3561, t=365 days, source: EPDS)



Graph 3. Boxplot (max, 75th percentile, median, 25th percentile, min) of the number of arrivals per day of the week (n=3561, t=365 days, source: EPDS)

As becomes evident from these graphs, the variation in arrival rate is severe and it differs somewhat per day. Disregarding the weekend, where during the year only very few surgeries have taken place, the most stable day is Wednesday, where between 6 and 17 patients can be expected to arrive at the ward. This shows the necessity for change.

2.1.2. Available beds at the URO/GYN ward

The number of beds at the ward is considered fixed in this research. Apart from simplicity reasons, this also represents reality fairly well as the number of beds is kept rather constant in order to accommodate the variation in arrivals.

Since URO and GYN patients arrive at the same ward, there is no distinction between beds for any one specialisation. Rather, there are 38 beds in total, including those reserved for unplanned (often emergency) patients. This number of beds is the so-called emergency norm. To calculate overcrowding for the specialisations separately, the beds are divided according to the arrival rate ratio: $\frac{Avg.arrival\ rate\ specialisation}{Avg.total\ arrival\ rate} \times (Total\ number\ of\ beds - Total\ emergency\ norm)$. This is displayed in Table 1.

Total number of beds	Specialisation	Avg. arrival rate per day (planned patients)	Emergency norm (beds)	Beds allocated
38	URO	8.22	4.70	20.11
	GYN	5.23	0.40	12.79
Total			5.10	32.90

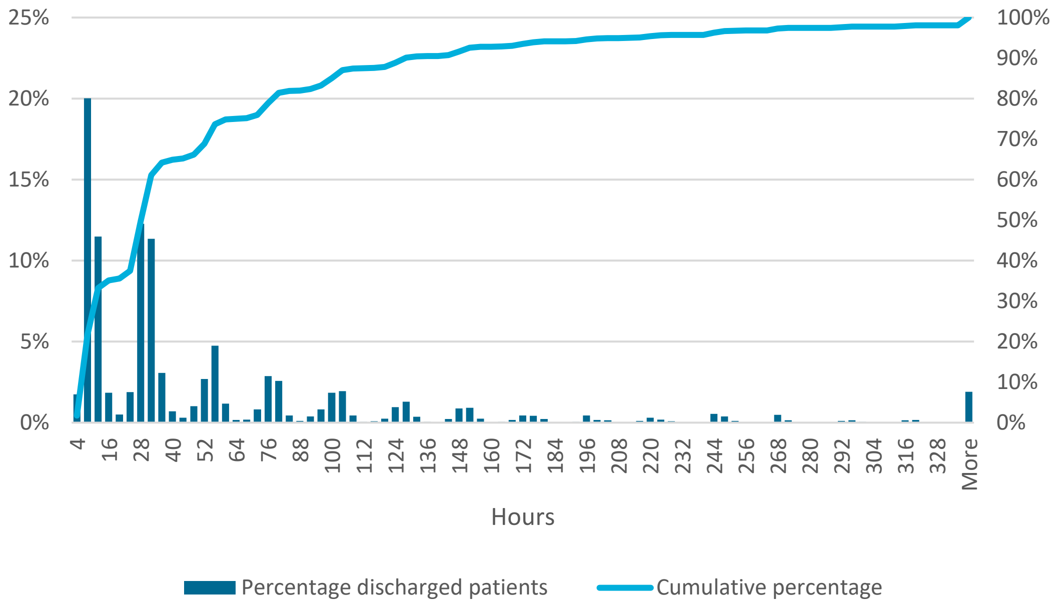
Table 1. Division of number of beds at URO/GYN ward

Note that this division in beds is purely hypothetical and does in no way affect the allocation of patients. Rather, it serves to point out which of the two specialisations has a higher utilisation and level of overcrowding.

2.1.3. Patients' LOS

A patient's LOS defines after how long he or she is discharged from the hospital. After discharge, his or her bed becomes available for the next patient. It can be considered the service time (from queuing theory) of the bed in the ward at which the patient is placed. There is a huge variation in the LOS, mainly due to the several patients that stay for over two weeks.

A large part of the patients, however, leaves the hospital within 24 hours. Interesting to see is that there are small peaks in the number of discharges every 24 hours, which would indicate that the nursing staff are rushing to discharge patients at a certain time. From the data, we see that this happens every 24 hours plus 4 to 8 hours. See Graph 4. Together with the observation that most patients come in before 2 PM, this indicates that the ward rushes to discharge most of the patients before the night. Indeed, approximately 56% of URO and 61% of GYN patients are discharged after 2 PM. Note that in this graph the whole week is considered.



Graph 4. Percentage of discharges per time unit of 4 hours (n=3561, t=365 days, source: EPDS)

2.2. Patient flow management

2.2.1. Patient routing through the hospital

In this section the main patient flows will be discussed, using Figure 1. It gives the reader a basic idea of the so-called care pathways a patient can take through the hospital.

There are four main ways in which a patient can arrive at a ward. The first is by means of an appointment for surgery. The patient is placed in a bed at the ward, awaiting admittance to the OR or Examination Room. Second, patients arriving at the ED and who do not need to be placed at the ICU or MCU are placed in beds at a ward. The third stream is via the ICU or MCU. These patients originate from the ED as well, but because they need to be placed at the ICU or MCU they arrive much (days) later at the wards and are therefore seen as a separate flow. Fourth and last, some patients are transferred from another ward or hospital to a ward at the Radboudumc. Any emergency patient that need to be transferred from another hospital to the Radboudumc due to specialisations available at the latter are considered to arrive via ambulance at the ED and are thus not a separate flow.

From the Examination Room patients might go directly to the OR if there is a serious complication. If it does not need to be treated the same day, it is assumed patients go home to come back to the hospital another day. From the OR patients can either go to the ward if there is no serious complication or to the ICU/MCU if the patient needs a higher level of care or monitoring. Only if a patient is stable, (s)he can be placed (back) at a ward. Going from the ward to the ICU/MCU is also possible if a patient's situation worsens to a critical level.

Commonly, patients coming in at the ED first go through a scanning procedure (for example, diagnostics or MRI). The place in the hospital where these processes take place is in Figure 1. called the Examination Room, to simplify the picture.

Leaving the hospital occurs either when a patient is discharged (i.e. sent home), (s)he deceases, or the patient is transferred to another hospital. Leaving a ward can also happen via a transfer to another ward in the same hospital.

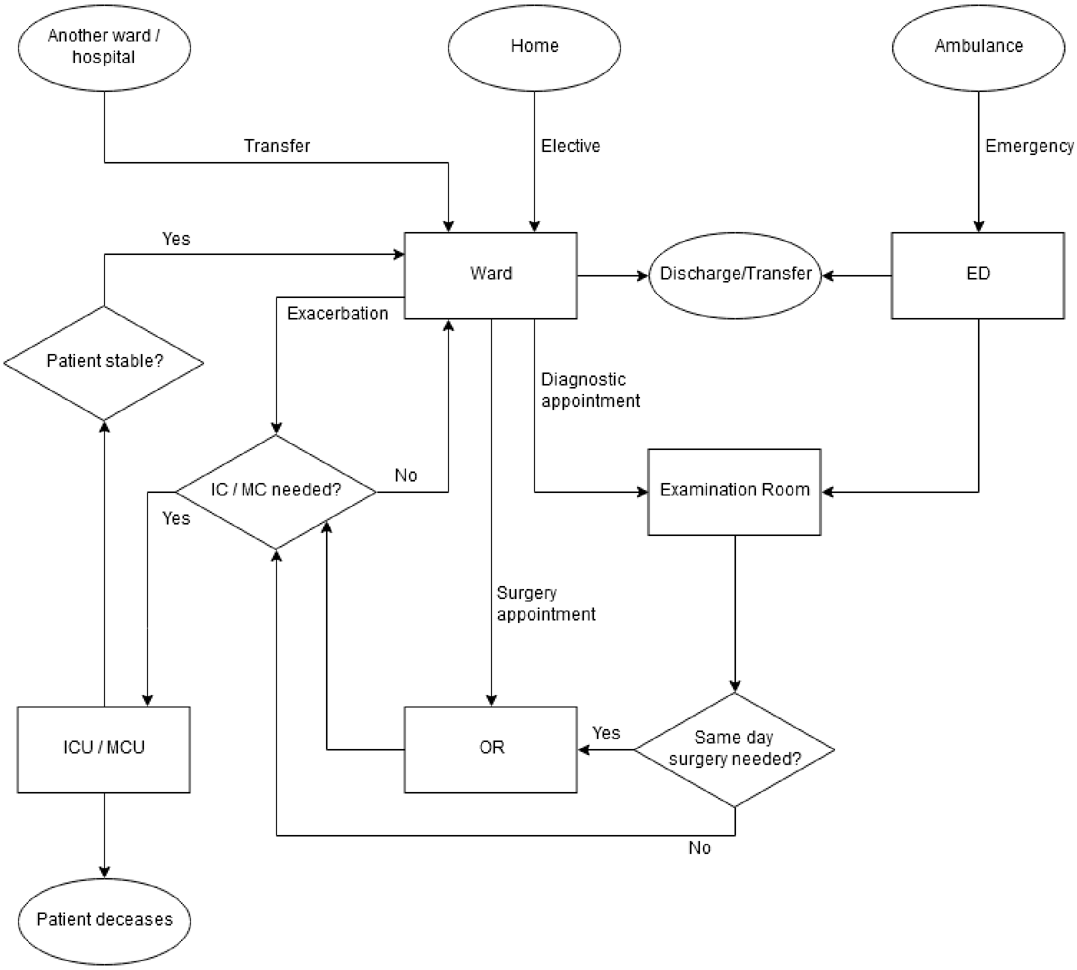


Figure 1. Patient flows at the Radboudumc

2.2.2. Planning OR and bed allocation

OR planning for a certain week starts as soon as there is a patient in need of surgery, but not earlier than two months before the surgery week that is being planned. Gradually, the surgery schedule is filled until about three weeks before the week of the surgeries. Three weeks before the surgery week, the schedule is conceptually finished and only minor changes are possible. The Thursday before the week of surgery, the planning is formalised, although often with minimal changes compared to the concept planning of three weeks earlier.

The Thursday before the surgery week, the wards receive the surgery planning as well. The allocation of beds is then a rather difficult task, given that the surgery planners have not considered the stress their planning puts on the bed occupancy and given that there are still some emergency patients coming in. Some beds are therefore reserved for emergency patients, while the rest is distributed on a first come-first serve basis. In the bed-meeting at 9 AM, the latest information regarding the surgery schedule is known (including, for example, cancellations), and solutions for the excess demand can be sought at other wards, and, in a rare case, at another hospital in the neighbourhood. In the next section, more is explained about this bed meeting and RTDC.

2.2.3. Real-Time Demand Capacity management

To decrease the time patients must wait for location in a nursing ward, and to ease the stress on the nursing staff, the Radboudumc adopted the RTDC method. The method was developed as a joint effort between the Institute for Healthcare Improvement (IHI) and several hospitals in the USA, in the IHI's Improving Hospital-wide Patient Flow Community. In the article of Resar et al. (2011), the method had proven itself useful and decreased, among other things, the mean LOS at the emergency department (ED) from well over 5 hours to under 4, a decrease of over 20%, at the University of Pittsburgh Medical Center (UMPC) Shadyside.

It was identified by Resar et al. (2011) that at UMPC Shadyside the most admissions occur before 2 PM, hence if those admissions were not managed properly overcrowding would occur after 3 PM. Therefore, RTDC aims at managing the morning admissions better, which eases the pressure in the afternoon. The method consists of four steps, to be taken before or during each bed meeting at 9 AM (at the Radboudumc):

1. Predicting capacity. Before the bed meeting, appointed representatives of each hospital ward should have a list of the scheduled discharges before 2 PM that day. In combination with the already available beds, this will define the capacity.

2. Predicting demand. Demand consists of scheduled (elective) patients, and unscheduled (emergency) patients requiring a bed. These scheduled admissions are quite easy to predict, based on the OR schedule for that day, and potential transfers. A prediction of emergency patients is obtained using historical data.

3. Developing a plan. If the demand should exceed capacity for a certain ward, a plan must be made. This plan should be detailed, and at least include a Who, What, Where, and by When. An example plan would be: "John to arrange with 3 W by 9:00 A.M. to transfer an off-service patient by 10:00 A.M." (Resar et al., 2011, p. 211).

4. Evaluating the plan. In each bed meeting the predictions and the plans of the day before are evaluated, in order to determine the impact of RTDC, and to learn from and increase prediction accuracy.

RTDC practices are consistent with recommendations by H*Works (the Health Care Advisory Board Company in the USA), as described by Coulombe & Rosow (2004).

RTDC is expected to decrease the LOS of patients in the hospital and to improve the off-service process by having clear what the demand for and capacity of beds is at 9 AM and acting upon any discrepancies in demand and capacity. A Flow Coordinator oversees the execution of plans made during the bed meeting and coordinates any unforeseen demand for beds during the day.

2.3. Key Performance Indicators

Together with the Radboudumc four Key Performance Indicators (KPIs) for patient flow are identified. These KPIs are used to evaluate the performance of the hospital and to identify points of improvement.

1. **The median discharge time**.

   The discharge time is a KPI since it is a focus point of RTDC. The median discharge time is taken rather than the average as the median gives less weight to outliers. The median of an ordered set of numbers is the middle number in an odd set of numbers, or the average between the two middle numbers in an even set. The median discharge time is currently 2:50 PM.

2. **The length of stay per patient**.

   The LOS is related to the discharge time, as earlier discharges result in a lower LOS. Next to the average LOS, the median LOS is obtained to correct for outliers that draw of the mean. The average is calculated as

   $$\frac{1}{n}\sum_{i=1}^{n} LOS_i \ ,$$

   where n is the total number of patients from which we obtain the average LOS. The calculation of the median is the same as described in KPI 1.

   The average LOS per patient is 2.35 days, or 56 hours 29 minutes. The median LOS, however, is 1.17 days, or 28 hours 4 minutes. Specified for the two specialisations, the LOS is 2.64 days or 63 hours 16 minutes (mean) and 1.19 days or 28 hours 40 minutes (median) for URO and 1.92 days or 46 hours 4 minutes (mean) and 1.09 days or 26 hours 14 minutes (median) for GYN patients.

3. **The off-service rate**.

   The off-service rate is equal to the level of overcrowding and therefore the metric under study.

   It is calculated as

   $$O_i = \begin{cases} P_i - C\ , & (\varphi + \psi > C) \bigwedge (\psi \leq C) \\ A_i\ , & (\varphi + \psi > C) \bigwedge (\varphi > 0) \\ 0\ , & (\varphi + \psi < C) \bigvee (\varphi \leq 0) \end{cases}$$

   where

   $\varphi = A_i - D_i$

   $\psi = P_{i-1}$

   $P_i = P_{i-1} + A_i - D_i$

   and

A_i is the number of arrivals in block i

D_i is the number of discharges in block i

P_i is the number of patients in the system in block i

C is the capacity of the ward

O_i is the overcrowding in block i in absolute number of patients.

Blocks (of time) are numbered continuously.

The off-service rate (or percentage overcrowding) is then calculated as

$$\sum\nolimits_{i=1}^{n} O_i \Big/ \sum\nolimits_{j=1}^{n} A_j \, .$$

This results in an off-service rate of 8.85% for the URO/GYN ward, if the day is divided in two blocks, following the RTDC principle: Block 1 is midnight until 2 PM and Block 2 is from 2 PM until midnight. Justification for the choice of blocks is found later, in section 4.3.1.

4. **The First-Time-Right rate**.

   The First-Time-Right rate is the percentage of all patients that is, upon arrival, placed in the ward designated to treating their condition. It is an important indicator of the quality of service for the Radboudumc. In this research, the First-Time-Right rate is the same as $1 - \text{off-service rate}$, as it is nearly impossible to check for all the patients that change wards halfway during their stay at the hospital.

2.4. Conclusion on the current patient flow

From this chapter we conclude that the number of arrivals per day is volatile (see again Graph 3). Since over 95% of patients come in before 2 PM, the most important time window is, following RTDC, "before 2 PM". RTDC is expected to reduce the current LOS because it's focus is on discharging patients earlier during the day, and in doing so, beds come available earlier and more patients can be placed at the right ward. In developing a model, we incorporate the arrival rate volatility (variation) and LOS. The literature review in the next chapter focusses on how these variables can be modelled.

3. Literature review

In the literature review, the research questions posed in section 1.6.1 are answered. Each section in Chapter 3 answers one of the four sub-questions. Articles for this literature review were obtained as described in Appendix I.

Next to providing the reader with an overview of previous studies concerning the management and modelling of patient flows in a hospital, this section aims to find common distributions of the arrival rate and length of stay of patients in hospital wards. These distributions will be used to see if they can be applied to the data of the Radboudumc.

Section 3.1 answers the first sub-question, "what is written in literature about managing and modelling patient flow at a tactical and an operational level?". Section 3.2 focuses on the second sub-question, "what are the advantages of using simulation, queuing theory, or Markov chains?". Subsequently, section 3.3 answers the third sub-question "what are often-found distributions for the arrival rate, and how can the time-dependent arrivals at hospital wards be modelled?". Section 3.4 answers the last sub-question "how can the length of stay at hospital wards be modelled?". Finally, section 3.5 aligns the literature review with the situation at the Radboudumc.

3.1. Managing and modelling patient flow at an operational and tactical level

As this research is conducted on the border between the operational and tactical level, this section includes literature research on managing and modelling patient flow on both levels.

3.1.1. Patient flow at the operational level

Active bed management (ABM) is a method to improve the bed assignment to new admissions and management of resources of the department of medicine (Howell, Bessman, Marshall, & Wright, 2010). In general, ABM involves employing a 24h hospitalist function that triages patients (admits patients to the ED or redirects them to an appropriate observation unit) and directs patients from the ED to the right inpatient ward (Soong, et al., 2016). With the implementation of ABM, Howell et al. (2010) observed fewer hospital diversion hours and a shorter ED-to-ICU time. Soong, et al. (2016) also report a significant increase in ED efficiency despite the slightly different implementation of ABM in the four hospitals described in their study. Similarly, De Anda (In Press) studied the effects of employing a flow nurse coordinator in the ED to improve patient throughput. The result was a 20% decrease in patient transport time from the ED to an inpatient ward. ABM, and the idea to employ a hospitalist or flow nurse coordinator to oversee the admission process are also the base concepts of RTDC.

Barnes, Hamrock, Toerper, Siddiqui, and Levin (2016) built on the RTDC idea of discharge prioritisation by predicting the patient discharge pattern using machine learning. The regression random forest method they used outperformed clinician's estimates of the number of discharges before 2 PM and midnight, although the method's predictions were more sensitive and more specific, meaning they predicted a higher number of discharges, with more false positives. Using new machine learning techniques can improve the discharge prediction and thus give a more realistic estimate of the number of available beds during the day. A pitfall is to discharge patients according to the prediction. RTDC aims to discharge patients as soon as possible, but not too soon.

Helm, AhmadBeygi, and Van Oyen (2011) developed a Markov decision process model to control hospital admissions. They defined three zones based on hospital occupancy, where under 85% occupancy expedited patients (patients with non-emergency complications that come in through the ED) are called in, and above 95% occupancy elective surgeries are cancelled, although these thresholds can be changed per hospital or even per unit, if necessary. The method appears to be more effective with the percentage of arriving emergency patients increasing. However, compared to the USA, there are very few patients coming in through the ED in The Netherlands, which would limit the effects of implementing this method.

Klein & Reinhardt (2012) used a spreadsheet program to model the patient flow and compared it to modelling it in the simulation program MedModel. The authors concluded with 95% certainty that the results of the simulation in both programs are within 4% of one another, if 1 or more patients per 20 minutes are discharged from the unit. Although the spreadsheet program (Microsoft Excel) took much longer to run (8 minutes to generate one data point versus 3 minutes in MedModel), the main advantages of using spreadsheets are that such a program is usually available at any hospital ,and it does not require the training of users (or at least to a much lesser degree compared to a simulation program like MedModel) because it is already the standard for manipulating, analysing, and storing data.

We conclude that hospital arrivals are significantly different in for example the United States of America and The Netherlands. The percentage of unplanned patients in the USA (as mentioned for example in Howell et al. [2010]) is much higher than in The Netherlands, which could explain the focus on ED patient flow in literature. Considering the similarity of the alternatives, RTDC appears to be a good way to manage overcrowding operationally. The method of Helm et al. (2011) is different, but can be used in addition to the other models. If ABM or RTDC is implemented, the thresholds of Helm et al.'s model can be set higher. If the cause of overcrowding is a structural one, however, RTDC nor any of the other operational methods will provide a measure for the necessary change. In the next section, therefore, tactical approaches to managing patient flow are discussed.

3.1.2. Patient flow at the tactical level

Tactical patient flow improvements can be achieved in several ways, one of them being the optimisation of the nursing schedule such as in Elkhuizen, Bor, Smeenk, Klazinga, and Bakker (2007). Since optimising nursing staff capacity is not the focus of the present research, this will not be explored any further.

Much like Helm et al. (2011), Nunes, de Carvalho, and Rodrigues (2009) developed a Markov decision model for patient admission control. They do not consider a day-to-day schedule, however, but rather schedule per period. Hence, their research is more focused on tactical or strategic scheduling. Furthermore, as also noted by Hulshof, Boucherie, Hans, and Hurink (2013), this solution still lacks practical applicability because the solution space in practical instances would be too big.

Hulshof et al. (2013) concluded in their literature review that solutions for tactical resource allocation and admission planning problems are either myopic (focusing on only one part of the care chain), focused on developing long-term cyclical plans, or do not provide a feasible solution for real-life sized problems (such as in Nunes et al. [2009]). Therefore, their paper presents a dynamic mixed integer linear programming (MILP) patient planning solution that incorporates multiple departments, resources, and care processes. This solution, however, is not easy to grasp for hospital employees as it is rather technical. Furthermore, an assumption in the model of Hulshof et al. (2013) is that patient arrivals are known and deterministic. This assumption makes sense if we consider that patients are planned, but in practice every surgery day is different, making the model complex. Similarly, Kumar, Costa, Fackrell, and Taylor (2018) developed a (simpler) stochastic MILP where a distinction is made between short-stay and long-stay patients.

Landa, Sonnessa, Tànfani, and Testi (2018) developed a discrete event simulation model to address tactical decision problems for hospitals that already employ a bed manager (with a similar function as in RTDC), such as the number of patients misallocated, the number of elective patients delayed, the average waiting time in the ED before admission, and the average number of patients waiting to be admitted. Pareto-optimal configurations are proposed for bed manager decisions, based on a simulation model that optimised one variable at a time (sometimes at the expense of another variable), but mainly their model points out the relation between the variables, i.e. quantifying how optimisation of one variable may lead to the worsening of another.

Larsson and Fredriksson (2019) have taken a more qualitative approach to tactical optimisation of the planning process, by identifying important facets for developing a model based on a review of the literature and case studies.

Concluding, patient flow can be managed on the tactical level by either managing the nursing schedule or by patient admission and allocation planning. There are several models that consider the expected bed occupation at the wards when scheduling surgeries. These models, although useful, are expected to be too complex for hospital staff to use. Hence, we identify the need for an easy to grasp method that can be widely implemented.

3.2. Advantages and disadvantages of using simulation, queuing theory, or Markov chains

Computer simulations, queuing theory, and Markov chains are widely adopted in modelling patient flow. Literature is divided on what is the best approach.

Wiler, Giffrey, and Olsen (2011) reviewed the literature on modelling ED crowding and compared formula-based methods, regression-based modelling, time-series analysis, queuing theory, and discrete event (or process) simulation (DES) for modelling the ED. In the authors' assessment, both queuing and DES models have a good ability to predict process improvement impacts, but the ease of model development is poor. Regression-based modelling and time-series analysis would be easy to develop and use but provide a poor ability to predict improvement impact.

McClean, Barton, Garg, and Fullerton (2011) argued that analytic models can be used in simple situations, although simulation may be necessary if the situation becomes a little more complex. Mathematical models are not very user-friendly and require sometimes very broad assumptions, while a graphical simulation can more easily be built iteratively, together with stakeholders (Everett, 2002). An example of such an assumption is homogeneous patients or patient groups, which is not always realistic (Davies & Davies, 1994). If these assumptions are relaxed, the mathematics of these models become extremely complex (Davies & Davies, 1994; Hu, Barnes, Marshall, & Wright, 2018). In a simulation, one could provide a patient with specific characteristics that influence the patient's route through, or time at the hospital.

Standfield, Comans, and Scuffham (2017) also listed several advantages of discrete event simulation (DES) compared to Markov cohort modelling: Markov cohort modelling incorporates implicit time delays that did not represent delays observed in practice, but which are inherent to the modelling process. Furthermore, DES is more patient-specific since Markov modelling uses transition probabilities and therefore generates a more general solution. Simulation modelling is computer-based, and it is therefore much easier to create many different paths, depending on the attribute (age, sex, disease history, etc.) for example Cooper, Brailsford, and Davies (2007). This requires much more data, however, and the pitfall is not to make the model more elaborate than necessary (Davies & Davies, 1994).

Simulation models, however, take considerable time to run (Cooper et al., 2007). Standfield et al. (2017) compared Markov cohort modelling and DES and found that the runtime was 4.4 seconds for the former, and 10 hours for the latter. Markov and semi-Markov models are in that sense much more convenient (Davies & Davies, 1994). On the other hand, Markov models are very hard or impossible to solve analytically if transition probabilities change over time (which could be represented by semi-Markov models) and require stricter assumptions about the stochastic behaviour of the underlying process. Spreadsheets might provide an outcome for these models.

They need to be modelled in sequential time periods, however, where patients are homogeneous in any state (Cooper et al., 2007).

Queuing theory is useful in modelling and analysing resource-constrained industrial settings because of the little data requirements and easy implementation, for example by means of a spreadsheet program (Hu et al., 2018). Though queuing theory therefore seems a natural approach to modelling health care processes, Hu et al. (2018) noted that a hospital, specifically an ED, is not a general service setting. Rather, it requires the prioritisation of patients to provide timely access to necessary health services, which makes modelling an ED more complex than a general service provider. It is noted, however, that the processes at an ED are quite different from those at a general hospital unit (Hu et al., 2018). Hence, for the simpler processes, queuing models might provide a realistic option.

Hu et al. (2018) suggested that combining queuing theory with simulation can provide the best of both in modelling health care practices. McClean et al. (2011) also used simulation and an analytic approach complementary. The analytic (mathematical) model is used to model the basic scenario because of its computational efficiency, while DES is used to model the complex elements. For simpler models, queuing models can be used to validate simulations (or vice versa) (Hu et al., 2018). The choice between an analytical or simulation model should depend on the need to model interactions between individuals (Barton, Bryan, & Robinson, 2004; Hu et al., 2018).

For an extensive review of the advantages and disadvantages of Markov and simulation modelling, the reader is referred to Standfield, Comans, and Scuffham (2014).

Concluding, both computer (discrete event) simulation and analytical approaches such as Markov and queuing models have their limitations. The best advice is to combine the two approaches, where analytical models are used to model a hospital ward until the need to model patient-specific attributes (such as medical history) exceed the advantage of a simple or simplified model, and when interactions play an important role. In that case, small parts should be modelled in a simulation program.

3.3. Arrival patterns and time-dependent arrivals at hospital wards

3.3.1. General arrival process

Belciug and Gorunescu (2016) modelled arrivals to a Geriatric unit according to a Poisson distribution and state that all associated assumptions are reasonable for a stable hospital system, where justification is based on their literature review. An obstetric ward can also be modelled with Poisson arrivals (and general or exponential service times), as shown in Takagi, Kanai, and Misue (2017). The key here, is the unplanned arrivals. Hospitalisation of geriatric patients, as well as obstetric patients, are not planned and can therefore be said to occur randomly. The same holds for the arrivals at an emergency cardiac unit, as described by De Bruin, van Rossum, Visser, and Koole (2007). They considered the mean arrival rate per day, and not an arrival rate per hour, as arrival intensity. This is because the average number of (unplanned) arrivals is quite stable, consistent with the findings of Belciug and Gorunescu (2015; 2016).

That time-independent arrivals do not hold for every emergency department is shown by McCarthy et al. (2008), who observed ED arrivals and considered 4 equal time intervals of 6 hours for each day. They applied a Poisson log-linear regression methodology in order to predict the number of arrivals during a certain time interval. Several Poisson assumptions are satisfied, however, such as the assumption of independent arrivals: the arrivals of one hour do not predict the number of arrivals the next hour. Comparing to the random arrivals of a geriatric unit (Belciug & Gorunescu, 2015; 2016) and obstetric ward (Takagi, Kanai, & Misue, 2017), ED arrivals are not necessarily independent of time since logically, accidents occur more frequently during the day when there are more people awake.

Bittencourt, Verter, and Yalovsky (2018) considered a general and a surgery ward unit, where patients from the emergency department are not treated. Interestingly, they found no evidence to reject the Poisson arrival hypothesis, which is in contrast with Green and Nguyen (2001), who argue that the use of an *M/M/s* model for general surgery (where they consider 50% unplanned, meaning urgent or emergent, surgeries) may overestimate delays because at least half of the surgeries are planned and therefore not random.

Because of the nature of emergency arrivals, many studies model them as a Poisson process. In this research, only planned patients are considered, however. As the *M/M/s* model assumes that arrivals occur according to a time-homogeneous Poisson process (Green, Soares, Giglio, & Green, 2006) and the arrival rate of planned patients at the Radboudumc between 7 and 8 AM is almost four times as large as the arrival rate between noon and 1 PM, it makes sense to explore how to model time dependent arrivals. Therefore, the modelling of time-dependent arrivals is explored in the next section.

3.3.2. Time-dependent arrivals

Knessl and Yang (2002) considered an *M(t)/M(t)/1* model, hence a queuing model with time-dependent Poisson arrivals, time-dependent exponential service rates, and one server. Without explicit proof, they stated that if the service rate is constant, the time-dependent $\rho(t) = \lambda(t)/\mu = (b - a\mu t)^{-2}$, where $\rho(t)$ is the probability that a new arrival must wait at time t, $\lambda(t)$ is the time dependent arrival rate, μ is the service rate, and *a* and *b* are constants. If the number of servers can be increased, this can serve as a potential way to model a general hospital ward. Tan, Knessl, and Yang (2013) considered *M(t)/M/1 – K* systems, where the capacity may vary from 1 to *K*. The assumption made in both papers, however, is that the arrival rate varies slowly (namely, λ = [*t/K*], where *K* is the maximum capacity of the system).

Kao & Chang (1988) suggested the use of a piecewise polynomial to model ED arrivals. They consider day-of-week and time-of-day effects on the arrival function, as will also be done in this paper. Using a piecewise polynomial reduces the mathematical complexity compared to an exponential polynomial by using low-order terms. Kao and Chang (1988) mentioned a few issues to consider when using their approach. Primarily, their approach does not guarantee a non-negative rate function estimated from data. Furthermore, the approach does not describe a method to find the degree of the polynomial within a pair of adjacent breakpoints. Finding the degree of the polynomial is therefore an exploratory process, and Kao and Chang (1988) suggest working only with fourth-degree polynomials or lower.

Kim and Whitt (2014a) found that the arrivals at a hospital ED were consistent with a nonhomogeneous Poisson process[1] (NHPP), but only when corrected for data rounding (an NHPP cannot deal with interarrival times of zero seconds), the use of inappropriate subintervals (which contain too much variation), and overdispersion caused by inappropriately combining data in an effort to increase the sample size). Especially for that second issue, it is important to choose time intervals in which the arrival rate is piecewise constant. If that is possible, an NHPP model with several linear arrival rate functions can be identified by means of a Kolmogorov-Smirnov (KS) test Hu et al., 2018). When the KS test fails, Kim and Whitt (2014a) provides guidelines on how to deal with the issues, such that modelling of time-dependent arrivals as an NHPP is possible. Kim and Whitt (2014b) described the difference between the conditional uniform KS and the Lewis KS test, to test

[1] A nonhomogeneous Poisson process is a Poisson process with a time-varying arrival rate with intensity function *(t)*, where λ is the arrival rate at time *t* (Leemis, 1991).

data against an NHPP. They find that the Lewis test has a greater power than the conditional uniform KS test.

Leemis (1991) stated that queuing systems with a time-dependent arrival rate are often modelled as a NHPP. For terminating simulations, the article demonstrates an estimation technique for the cumulative intensity function of an NHPP over an interval (0, *S*], without having to specify parameters or weighting functions. It is suggested that multiple realisations are used to create such a cumulative intensity function. Arkin and Leemis (2000) elaborated on Leemis (1991) by providing a technique that estimates the cumulative intensity function from (multiple) overlapping realisations. Increasing the number of realisations increased the precision of the estimator of the cumulative intensity function. Again, more realisations will result in a more precise estimate, as it reduces the width of the confidence interval.

In a container terminal environment, Dhingra, Kumawat, Roy, and De Koster (2018) suggested the use of a Markov-modulated Poisson process (MMPP) to model time-dependent arrival rates in a semi-open queuing network. In an MMPP, a Markov process defines the rate parameter of a Poisson process. A comprehensive explanation of MMPP models is given in Fisher and Meier-Hellstern (1993). The arrival rates of trucks at a container terminal in Rotterdam, The Netherlands, as described in Dhingra et al. (2018), seems to resemble the arrival pattern of patients at the Radboudumc. In order to use an MMPP for modelling a hospital ward, adaptation would be necessary since trucks can wait in line if the servers are occupied while patients need to be off-serviced to another ward. An advantage of this model is that a certain capacity can be reserved for, in our case, emergency patients.

Zychlinski, Mandelbaum, and Momcilovic (2018) used a fluid model to approach time-varying arrivals to multiple stations in tandem. Stations in tandem have together one queue, and a patient can only proceed to the next station if there is a server (bed) available. This is called blocking after service (BAS). As patients generally do not go from a bed to a waiting room to another bed but rather remain in the same bed until they can be transferred, this is a good way to model patients' flow through the hospital. Zychlinski et al. (2018) considered a $G_t/M/N/(N+H)$ model to represent hospital arrivals and service, where N represents the number of servers and H represents the number of places in the waiting room (queue). Their fluid model approach implies that patients are modelled as a continuous fluid flowing through the hospital, which main advantage is that system dynamics can be described by differential equations.

It seems that not only random, or unplanned, arrivals can be modelled by the Poisson distribution, since Bittencourt et al. (2018) found no evidence to reject the hypothesis of Poisson arrivals in a general ward with no emergency arrivals. For the time-dependent arrivals however, we find that there is again much literature on ED arrivals. Nevertheless, a piece-wise polynomial approach such as the NHPP, the MMPP, and the fluid model seem to be good approaches to model the time-dependent arrivals at the Radboudumc.

3.4. Modelling the length of stay of hospital wards

Herwartz, Klein, and Strumann (2016) considered the Poisson and negative binomial distributions for lengths of stay. The negative binomial distribution differs from the Poisson distribution in that it allows the mean and variance (squared standard deviation) to be different, i.e. the variance to mean ratio[2] (VMR) of a negative binomial distribution need not be equal to 1. Therefore, this distribution is a good fit if the LOS of the patients is overly-dispersed (Herwartz et al., 2016). In the normal care clinical ward, De Bruin et al. (2007) observed an average LOS of 7 days, with a coefficient of variation (the ratio of the standard deviation to the mean) of 1.07. This could indicate an exponential distribution, as the coefficient of variation for the exponential distribution is equal to 1. For a general and surgery ward, Bittencourt et al. (2018) found that, applying the Kolmogorov-Smirnov (KS) method, the Log-Logistic probability model was most plausible to model the service rate. However, they could not reject the hypothesis that the service times were exponentially distributed.

Belciug and Gorunescu (2016) considered a multi-compartment model for the LOS in a geriatric unit, where each compartment can represent a certain treatment such as acute care, short-, or long stay, and rehabilitation. Two assumptions are underlying this model: the first is an arrival rate that is Poisson distributed, and the second is a phase-type[3] (*PH*) service distribution. In such a phase-type distribution, the number of components equals the number of compartments in the system (Belciug & Gorunescu, 2015). Phase-type distributions have the advantage of being very versatile, meaning they can closely approximate almost any nonnegative distribution, but at the same time, they have a Markov structure which allows for easy analysis (Fackrell, 2009). The reader is referred to Fackrell (2009) for a comprehensive and more elaborate introduction to *PH* distributions.

Atienza, García-Heras, Muños-Pichardo, and Villa (2008) assessed the fit of a finite mixture distribution made up from Gamma, Weibull and Lognormal distributions to the length of stay of several diagnosis-related groups (DRGs). They conclude that the mixture of these distributions results in a distribution that has a lower discrepancy with the empirical data than any of the distributions individually. However, as the DRGs are significantly different it is impossible to find a general model that fits all. Furthermore, the mixed model is much more difficult to compute than any individual one, although Atienza et al. (2008) chose to outweigh a model that fits the empirical distribution better to one that is less complex. Indeed, Ickowicz and Sparks (2017) agreed that the model of Atienza et al. (2008) is complex and added that it is furthermore very sensitive to the initial input. Both Atienza et al. (2008) and Ickowicz and Sparks (2017) drew the same conclusion: the model from an individual family of distributions is computationally less complex, but two major drawbacks are: (1) the appropriate family must be selected, and (2) the fit with empirical data is worse.

A next step in the process, which is outside the scope of this research, is to find an objective method to predict a patient's LOS. Grand and Putter (2016) stated that often-used methods for calculating the expected LOS rely on the Markov assumption, i.e. the memoryless property of a stochastic process, and do not incorporate the effect of independent variables on the expected LOS. They proposed the use of pseudo-observations in combination with landmarking to construct regression models to calculate the expected LOS. Using generalised estimating equations, the pseudo-observations are used as an outcome to fit a generalised linear regression model. Unfortunately, Grand and Putter (2016) did not conduct a goodness-of-fit test and did not compare the results of

[2] The variance to mean ratio (VMR) is a dispersion measure of a probability distribution, i.e. a measure of variability, and calculated as the ratio of the variance, or squared standard deviation, to the mean: $VMR = {\sigma^2}/{\mu}$

[3] A phase-type (*PH*) distribution is essentially a generalised Erlang type-*k* distribution, which is a distribution of *k* sequential exponential distributions (*k* phases). A customer starts service and needs to finish all *k* service components before another customer can enter the system. The generalisation of the phase-type distribution entails that many other distributions than just the Erlang distribution can be used in the phases (Gross, Shortle, Thompson, & Harris, 2008).

their regression model to results that would be obtained with any of the conventional methods of estimating the LOS.

3.5. Literature review synthesis with the Radboudumc

The results of Howell et al. (2010), Soong, et al. (2016), and De Anda (In Press) show a positive result of the implementation of methods comparable to RTDC. Furthermore, the studies of Klein and Reinhardt (2012) agree with the idea to develop a model in an application familiar to the hospital. At the Radboudumc, historical patient data can be extracted from the electronic patient record in Excel-format, which can be used to develop an historical event-based model that can later be adapted with new data. This model must be focussed on providing information that structurally improves patient flow, when RTDC can manage any occurring problems live, at the operational level.

A piece-wise polynomial approach such as the NHPP, the MMPP, and the fluid model, as described in section 3.3.2, are possible approaches to tackle the problem addressed in this thesis. Even then, there is still the question of the ease of implementation. As with the distribution for the LOS, we opt for a simple approach as the model needs to be able to be adapted by people with limited statistical knowledge, in order to correct for new data or use the model for other wards in the hospital.

As mentioned in section 3.4, Atienza et al. (2008) and Ickowicz and Sparks (2017) chose to outweigh model accuracy over simplicity. Although accuracy is indeed important, we believe that, considering the knowledge on mathematical modelling or simulation programs of the expected users of the model which is to be developed, it should be simplicity over accuracy. A certain degree of compromise on the fit with empirical data is to be allowed in order to make the model understandable and manageable for the users of it. Therefore, considering the ease of use, the most promising distributions to model the length of stay are therefore the Poisson and negative binomial (Herwartz et al., 2016), and phase-time (Belciug & Gorunescu, 2015; 2016; Fackrell, 2009) distributions.

4. An adaptive patient flow model

This chapter dives into the mathematical analysis of the patient flows at the Radboudumc. A model is developed in Excel and validated for the current patient flow situation.

First, section 4.1 describes the assumptions in this research. Section 4.2 follows with a test for the arrival rate and LOS distribution, and then tests the in the literature suggested distributions against the historical data. Next, section 4.3 describes how the simulation model is developed, after which the model is validated based on historical data in section 4.4. This validation is only a validation of the base model, however, where the simulation base model is compared with the current situation in practice.

4.1. Assumptions

In order to build an analytical model for the hospital, we take the following assumptions:

1. *There are no batch arrivals*. If multiple patients are registered to have arrived at a ward at the same time, which is registered to the minute, it is assumed that these patients arrived equally distributed during that minute. This is a logical assumption as it is unlikely that two patients travel together to the hospital and therefore, they arrive independently of each other, not at the same time.
2. *Our data is correct*. An important assumption, since if the data is incorrect, no conclusions can be drawn from the model.
3. *The average arrival rate is equal on every day of the 5-day working week, with zero arrivals during the weekends*. This assumption is not valid in practice, as there are many more arrivals on the Tuesdays and Fridays than on Mondays, Wednesdays, or Thursdays. It is, however, common knowledge that in order to increase the utilisation of a ward the variation in arrivals must be reduced, which is primarily done by having an equal arrival rate each day. It is further assumed that the hospital can use the limits suggested in this thesis to schedule a different number of patients each day.
4. *Arriving patients before 2 PM can be placed in beds of discharged patients before 2 PM*. This assumption is explained and validated in section 4.3.1.

4.2. Current arrival and length of stay distributions

In order to later validate the model, the current patient flow distributions must be further analysed. In sections 2.1 and 2.3, some general observations have been made regarding the arrival process and LOS distribution. In this section, in the literature suggested distributions are tested against historical data.

4.2.1. Choice of goodness-of-fit test

To validate the hypothetical distributions against historical data, two goodness-of-fit tests can be used. The first is Pearson's Chi-squared test, useful for binned data, and the second is the Kolmogorov-Smirnov (K-S) test, which is used with continuous and binned data. Since our independent variable is arrival time and time is a continuous variable that could be binned in bins of, for example, one hour, both tests could theoretically be used.

The advantage of the Chi-squared test is that it is widely known and simple to implement. The K-S test, however, generally has more power (i.e. probability that the null-hypothesis is correctly rejected), except in cases where there is a small sample size, and it has a lower probability of a type-I error (a false positive, i.e. the null-hypothesis was rejected although it was true) (Wang, 2009). One major drawback of the Chi-squared test is that it requires at least 80% of the cells have an expected frequency of 5 or more. Furthermore, no cell may have an expected frequency of less than 1 (Mitchell, 1971). The K-S test does not have these requirements.

It is precisely for these requirements that the Chi-squared test cannot be used in this research. Since GYN has some planned nightly arrivals, we must consider the entire day instead of, for example, compare arrivals between 6 AM and 8 PM. Furthermore, the number of arrivals per hour of the working day do not exceed 2, on average. Hence, if we want to test for the probability of *n* arrivals per hour of the day, the Chi-squared test is not the appropriate one to use. Much more useful is then the K-S test, which compares the theoretical and historical distributions.

The KS test-statistic is given by

$$D_n = \sup_x \{|F_n(x) - F(x)|\} ,$$

where $F_n(x)$ is the cumulative empirical distribution function and $F(x)$ the cumulative distribution we are testing for, in this case the NHPP with parameter $\lambda(t)$. We want to test whether $\sqrt{n}D_n > K_\alpha$, where n is the number of observations, in this case the number of days with a certain timeframe. Since we consider only working days (Monday to Friday) for one year, this number is 260. K_α is found from $\Pr\{K \leq K_\alpha\} = 1 - \alpha$, where we take α to be 0.05, as is common. K is the random variable in the Kolmogorov distribution, given by

$$K = \sup_{t \in [0,1]} \{|B(t)|\}$$

Then

$$\Pr\{\sqrt{n}D_n > x\} = 1 - \Pr\{K \leq x\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2) \tag{1}$$

and K_a is equal to the value of x when $\Pr\{K \leq x\} = 1 - \alpha$.

As Vrbik (2018) points out, the accuracy of (1) is rather low. When $n = 1000$, the maximum error is almost 0.9% , while when $n = 10$, it is almost 7%. Vrbik (2018) proposes several alternatives to improving this accuracy and recommends replacing

$$x \rightarrow x + \frac{1}{6\sqrt{n}} + \frac{x-1}{4n} ,$$

which decreases the maximum error when $n = 10$ to 0.27%. Hence, we obtain the new equation

$$\Pr\{\sqrt{n}D_n > x\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\left[-2k^2\left(x + \frac{1}{6\sqrt{n}} + \frac{x-1}{4n}\right)^2\right] \tag{2}$$

4.2.2. Arrival distribution

As stated in section 3.3.2., Bittencourt et al. (2018) found no evidence to reject the null hypothesis that the arrivals at a ward with no emergency patients were Poisson distributed. Certain assumptions inherent to the Poisson distribution are not satisfied in a hospital context with planned patients, however. In particular, the Poisson distribution assumes that the rate at which an event (arrival) occurs is equal during the time considered. One interval of time can therefore not have a lower arrival rate than another interval of time. What we observed in Chapter 2, however, is that much fewer patients arrive during the afternoon hours than in the morning.

An observation of the data learns that the variance-to-mean ration (VMR) of the weekly arrivals before 2PM is larger than 1, that is, for the two specialisations under observation in this study. A characteristic of the Poisson distribution is its VMR of 1. We will therefore not consider the Poisson distribution as a possible arrival distribution.

Modelling a nonhomogeneous Poisson process

We consider a non-homogeneous Poisson process (NHPP) for the arrival rate and compare it to the empirical number of arrivals. An NHPP makes sense as the number of arrivals per hour is greatly dependent on the time of the day. Within a certain timeframe, however, arrivals may be distributed according to a Poisson process.

In order to model such an NHPP, formulae from course notes of an MIT module in electrical engineering and computer science is used (available from https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/course-notes/MIT6_262S11_chap02.pdf, last accessed: 18-02-2019). In these course notes it is stated that for an NHPP with right-continuous arrival rate $\lambda(t)$, the distribution of the number of arrivals $N(t,T)$ in the interval $(t,T]$ is given by

$$\Pr\{N(t,T) = n\} = \frac{[m(t,T)]^n \exp[-m(t,T)]}{n!}, \quad \text{where}$$

$$m(t,T) = \int_t^T \lambda(u)du$$

The NHPP can also be viewed as a homogeneous Poisson process in non-linear time. If one considers a homogeneous Poisson process of rate 1, $\{N^*(s); s \geq 0\}$, then the NHPP is given by $N(t) = N^*(\Lambda(t))$ for each t, where $\Lambda(t) = \int_0^t \lambda(u)du$.

We take for each $m(t,T)$ the historic average number of arrivals in that timeframe (data from one year). The numbers are given in Appendix II. In this table, t represents the one-hour time frame, i.e. $t = 1$ is the timeframe from midnight to 1 AM, $t = 2$ is the timeframe from 1 AM to 2 AM, etcetera.

Results and conclusion

We arrange our data per hour so that an arrival rate $\lambda(t)$ per hour is obtained. Each time interval t is one hour. We use this arrival rate to calculate a Poisson process for each hour, as per (2). Next, $\sqrt{n}D_n$ is calculated for each t. If we take α to be 0.05 and given that n is 260 (the number of working days in a year), we find that $K_\alpha = 0.51$, i.e. the test statistic $\sqrt{n}D_n$ must be larger than 0.51 in order to reject the null-hypothesis that the empirical distribution is an NHPP with parameter $\lambda(t)$, and is more likely to be any other distribution.

The results are displayed in Table 2.

Specialisation	Distribution	$\sqrt{n}D_n$
Urology	NHPP	0.912
Gynaecology	NHPP	1.235

Table 2. Calculation of √n D_n for the NHPP arrival rate for Urology and Gynaecology

Since the test statistic for both specialisations is much larger than the critical value 0.51, we must reject the null-hypothesis that the arrival rate for both the URO and GYN is conformed an NHPP with rate $\lambda(t)$.

4.2.3. Length of stay distribution

As with the arrival rate, LOS distributions will be compared using the K-S method, with adaptation from Vrbik (2018). As suggested in literature we consider the Poisson distribution, negative binomial distribution (which is a special type of Geometrical distribution), and a Phase-Type distribution. Concerning the latter, we take a distribution with three exponential phases, i.e. a three-phase hyperexponential distribution. The underlying assumption for this distribution is that the LOS of patients may be divided into short-stay, medium-stay, and long-stay. The average LOS for these groups is found by minimising the difference between the theoretical distribution and the empirical distribution. This problem is solved using the Excel Solver option.

The Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the number of events in a fixed time interval, given that the events are not dependent on each other, or on time:

$$\Pr\{X = k\} = \exp(-\lambda)\frac{\lambda^k}{k!},$$

Where k is the number of events in any interval (as each interval is equal), and λ is the event rate, equal to the mean number of events in each interval.

The negative binomial distribution

The negative binomial distribution is, as the Poisson distribution, a discrete probability distribution. It expresses the probability of s successes after r failures in a series of i.i.d. Bernoulli trials if there is a probability p of a success, and given by

$$\Pr\{X = s\} = \binom{r+s-1}{s} p^r(1-p)^s$$

Naturally, $0 \leq p \leq 1$.

The hyperexponential distribution

The hyperexponential distribution is a continuous probability distribution of several parallel exponential distributions. The probability density function is given by

$$f_X(x) = \sum_{i=1}^{n} f_{Y_i}(x)p_i ,$$

where $f_{Y_i}(x)$ is the probability density function of a random variable Y_i and p_i is the fraction of this density function in the global function $f_X(x)$. Naturally, $\sum_{i=1}^{n} p_i = 1$.

Results and conclusion

Specialisation	Distribution	$\sqrt{n}D_n$
Urology	Poisson	6.664
	Negative binomial	3.244
	Phase-type hyperexponential	3.155
Gynaecology	Poisson	4.960
	Negative binomial	1.979
	Phase-type hyperexponential	0.622

Table 3. Calculation of $\sqrt{n}D_n$ for the LOS fit of three different distributions for Urology and Gynaecology (n_URO = 2158, n_GYN = 1403, t = 365 days, source = EPDS)

Although the number of observations n is different than for the arrival rate, the critical value remains at 0.51. Then it is clear from Table 3, all three LOS distributions for both URO and GYN patients are significantly different from the historical distribution.

4.2.4. Comments on the current arrival rate and LOS

As becomes clear from the sections above, there is no specific distribution to describe the arrival rate or the LOS, although these distributions have been suggested in the literature. The number of arrivals for the URO and GYN specialisations cannot be modelled according to a NHPP with rate $\lambda(t)$, where t is a period of one hour. Furthermore, the LOS of URO and GYN patients cannot be described by either a Poisson, Negative binomial or three-phase hyperexponential distribution. A solution for the main problem will therefore be sought in a more general direction, as will be described in the next section.

4.3. Model development

As we found that an analytical approach to the situation at the Radboudumc was not fruitful, we opted for a stochastic, discrete event simulation (DES) model in Microsoft Office Excel. It is stochastic to allow for different scenarios and show the solution space, and simulates discrete events in order to reduce complexity and be able to model the simulation in Microsoft Excel. As mentioned earlier, Excel is a powerful visual tool and familiar to most in the organisation. The simulation, however, requires the use of Excel VBA (Visual Basic for Applications), with which the organisation is not familiar. Considering the visual nature of Excel and the easy syntax of VBA, we believe that it is nevertheless the proper program to use. Referring to section 3.2, we opt for a combination of a probability-based model and a DES model.

Section 4.3.1 describes how arrivals are modelled. Next, section 4.3.2 explains how a LOS is assigned to each arriving patient. Finally, the integration into the simulation model is described in section 4.3.3.

4.3.1. Generating arrivals

In section 4.2, no clear distribution for the number of arrivals was established, and any of the considered distributions is hard to use this in a simulation aimed at changing the variation of the influx of patients. Another approach to modelling arrivals is therefore needed.

The day was divided into two blocks, based on the RTDC concept: the first block is from midnight to 2 PM, and the second block is from 2 PM until the end of the day. This results in the following arrival rates:

Ward	Arrival rate before 2 PM	S.D. before 2 PM	Arrival rate after 2 PM	S.D. after 2 PM
Urology	7.801	3.945	0.421	0.600
Gynaecology	5.038	3.328	0.188	0.446

Table 4. Arrival rates URO and GYN before and after 2 PM

In order to be able to change the deviation in arrivals from the mean, a distribution that accommodates arbitrary changes in its variation is required to model the arrivals. Two examples of such distributions are the Normal and the Triangular distribution. Both these approaches are validated in two ways. First, the Flow Coordinator stated that "wards can handle the arrivals themselves before 2 PM", which means that it is the arrivals per block that matter, rather than the arrivals per hour. Second, the approach is validated visually, by comparing graphs of the number of arrivals per block. These graphs are found in Appendix III.

For the Normally distributed arrivals, the number of arriving patients in one of the two time blocks is obtained by generating a random number between -1 and 1 from the Normal distribution, then multiplying it with a specified degree of variation and the standard deviation of the block, and add that to the average arrival rate, i.e.

$$\begin{aligned}&\text{Arrivals in block } i \text{ of ward } j \\ &\qquad = Average\ arrival\ rate_{ij} + Random(-1{,}1) \times \text{degree of variation} \times S.D._{ij}\end{aligned}$$

We take the variation to be 1.96 at max, such that we consider 95% of all possible arrival rates. $Random(-1{,}1)$ is represented by a Normally distributed random variable with mean 0 and standard deviation 0.5. This way of modelling leaves some probability of a random value larger than 1 or smaller than -1. Such excesses are controlled by setting a minimum and maximum arrival rate.

The Triangular cumulative distribution function is given by

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, & a \le x \le c \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-c)}, & c \le x \le b \\ 1, & x > b \end{cases},$$

where a, b, and c are the minimum, maximum and peak, respectively.

The optimal values a, b, and c are found using the Excel Solver function. Of course, we cannot have negative arrivals, hence each time a negative number of arriving patients is generated, we assume there are zero patients entering the system.

Although the Triangular distribution provides a little better visual fit with the historical data (see Appendix III), in the further analysis the Normal distribution is included because it provides for easier manipulation. The K-S test gives a result for both the URO and GYN specialisation that rejects the null-hypothesis that the data follows a Triangular distribution, however (K-S test value of 3.15 and 2.20, respectively). These distributions are nevertheless used, as they provide for easy manipulation.

Variation in the arrival rate is obtained by varying the degree of variation (Normal distribution) and the minimum and maximum values (Triangular distribution). Note that the peak of the Triangular distribution is not equal to the mean arrival rate and hence if we vary the minimum and maximum of the triangle, the peak needs to be adjusted as well in order to keep the average arrival rate constant. This way of modelling moves the Triangular distribution towards the Normal distribution when the variation decreases.

4.3.2. Assigning a LOS to each patient

In order to generate a list of random patients with a pre-given LOS, we use the probability of a certain LOS in days. We generate a random number and obtain the LOS in days based on the random number's position in the cumulative distribution function for the LOS. For this, it would make sense to use the empirical cumulative distribution probabilities.

We generate 10,000 patients in this way. When A arrivals are generated, we loop over these A patients and use the next patient in the LOS list and his/her assigned LOS to determine when (s)he patient leaves the ward. As Excel provides an easy way to list and number arrivals and departures, we do the following: suppose we have an arriving patient in row r, and his/her LOS is 8 days. Then we list the departure in row $r + 8$.

Next, it is determined in which block the patient is discharged. We specify a percentage of patients that is discharged before 2 PM in advance. Then in order to determine the block in which the patient is discharged, we generate a random number between 0 and 1. If the number is smaller than the specified percentage, the patient is discharged in the block 'Before 2 PM', otherwise he/she is discharged in the block 'After 2 PM'. This way of modelling the discharges can be justified by remembering we put all the patients with a LOS between $x - 1$ and x days in the block of x days. Hence, making the LOS of each patient smaller than x days represents reality better than when we take the full x days. For the ward (specialisations URO and GYN combined), the average percentage of patients discharged before 2 PM is 43%, and after 2 PM is 57%.

In pseudo-code, where CD_X is the cumulative distribution function of the LOS and X is the LOS in days:

```
For 1 to 10000 do

      Generate random number R
      Find X for which holds CD_(X-1) <= R < CD_X
      Return X-1

      Generate random number S
      If S < Discharge fraction before 2 PM
            Discharge in block "before 2 PM"
      Else
            Discharge in block "after 2 PM"
      End If

Loop
```

This way of modelling results in a simulated LOS that closely represents practice, as shown in Appendix IV.

4.3.3. Integration into static simulation

We make a static simulation by defining all variables on a "Dashboard" sheet. We write a macro that generates the LOS of 10,000 patients in one sheet for Urology (called "URO LOS"), and 10,000 patients in another sheet for Gynaecology (called "GYN LOS"). Another macro subsequently generates a random number of arrivals for each block and uses the generated LOSs to determine when the patient is discharged, as is described in section 4.3.2. The output of this second macro is written to two sheets, again one for Urology ("URO") and one for Gynaecology ("GYN"). In these sheets we generate 3 years plus one month of arrivals and discharges. This one month is the so-called warm-up period. As we assume every patient is discharged no later than 29 days after he/she has arrived, after a month we should have a stable system. The generation of three years is an arbitrary decision based on the time it takes to run the simulation (approximately 3.7 seconds for one run)

The data from the two sheets "URO" and "GYN", where the arrivals and discharges are listed, are combined in a sheet called "Total". In this sheet we observe how the two wards in combination perform, as they are the best alternative for each other, if patients cannot be placed at the preferred ward. We want the number of off-serviced patients from the "Total" sheet to be no more than 2%. Some screenshots of the model are provided in Appendix V.

This simulation has four variables, namely the variation in the number of arrivals for the two specialisations considered, and the percentage of discharges before 2 PM for the same specialisations. The effects of these variables on the percentage of off-serviced patients are explored in Chapter 5.

4.4. Model validation

The developed model is validated using several metrics that can be read from the simulation, and from the historical data. The average arrival rate (per day and per week), the standard deviation (per day and per week), the average number of patients at the ward at any given time, the bed occupation, and the number of off-serviced patients are considered. It must be noted that this validation only considers whether the way of modelling accurately describes the historical data while the model is constructed based on the same data.

4.4.1. Comparison of output

We take the arrival process as described by empirical data. That is, from historical data we obtain the probability that n patients arrive in a given block. The assumption that each day is equal is of particular importance here. The results from the simulation are summarised in Table 5 below.

Metric		Historical			Simulation		
		URO	GYN	Ward	URO	GYN	Ward
Arrival rate/day	< 2 PM	7.801	5.038	12.839	7.789	5.038	12.827
	> 2 PM	0.421	0.188	0.609	0.423	0.189	0.611
S.D. /day	< 2 PM	3.945	3.328	4.476	3.934	3.321	5.147
	> 2 PM	0.600	0.446	0.760	0.599	0.447	0.720
S.D./week	< 2 PM	7.676	5.646	11.032	8.774	7.421	11.493
	> 2 PM	1.293	0.850	1.539	1.196	0.809	1.584
Average # of Patients in the system		15.652	7.471	23.123	15.025	6.854	21.879
Bed occupation		78%	59%	70%	75%	54%	79%
Overcrowding		19.8%	10.6%	8.9%	15.5%	7.7%	6.4%

Table 5. Comparing historical situation and simulation output, n = 500 runs

The results show that the input parameters (mean arrival rate and standard deviation per day) are almost equal, which indicates the input values give a good representation of the arrival rate per day. More interesting is the standard deviation per week, which is larger in our simulation. An explanation for this difference is that the assumption that each day is the same is incorrect. If the number of arrivals is random and the likelihood of n patients arriving in a block is Normally distributed, the pooled standard deviation of weekly arrivals is given by $S.D._{day} * \sqrt{number\ of\ days}$, which results in a value 8.821 for the URO, and 7.442 for the GYN specialisation (using the historical standard deviation). This is rather close to the values obtained in our simulation model. An explanation for this deviation between historical and simulated data is that arrivals in fact are not that random.

Schedulers allow for a large deviation in the number of arrivals per day, but per week they are more similar.

For the standard deviation per week at the ward, however, we see that there is almost no difference between the historical and simulated data. This may indicate that the schedulers for the URO and GYN specialisation do not communicate their planning with each other and confirms what employees at PVI think happens: if one of the schedules peaks in terms of the number of arrivals per week, the other one does too.

Overcrowding at the ward is in this research stated as 8.9%. That is, overcrowding is calculated to be 8.9% if we consider the day to have two time blocks as described in section 4.3.1. If we look per hour, however, and consider 24 time blocks, overcrowding is 22%, since patients are often discharged from 11 AM or midday onward, while arriving patients come in around 7 or 8 AM. This mismatch is not visible when considering two blocks per day. There are multiple ways, however, to correct this mismatch. For example, arriving patients are assigned to a bed and these patients then undergo surgery for several hours. During this time, another arriving patient may be put in the ward and a place is double booked. It also happens that a patient is temporarily assigned an emergency bed, a bed that is reserved for an emergency patient. This is often for only a few hours, until the next patient is discharged. In this way the nursing staff ensures that the assumption made in this research is valid, and it is precisely what is meant by “wards can handle the arrivals themselves before 2 PM”.

Furthermore, the results show that the average number of patients in the system is lower than observed from historical data. A possible explanation for this deviation is the few patients that remain at the ward longer than 28 days. Assume a patient remains at the ward for 3 months, then for two months the average at the ward is 1 patient higher than in the simulation. From the historical data, there are 6 urology and 7 gynaecology patients that have a LOS longer than 28 days. This LOS can be anywhere between 29 and 113 days for the observed year. In practice, this bed is therefore occupied during the entire LOS of such a patient, while in the simulation it is assumed patients have all left the ward after 28 days.

In order to test which approach towards modelling arrivals is superior, Table 6 below shows results for a Normally distributed arrival process and a Triangular distributed arrival process. Modelling arrivals according to a distribution of which the variation can be changed is critical to answering the main research question, as it is assumed that the variation in arrivals causes the overcrowding of nursing wards.

Metric		Triangular Distribution			Normal Distribution		
		URO	GYN	Ward	URO	GYN	Ward
Arrival rate/day	< 2 PM	7.885	4.917	12.788	7.822	5.126	12.948
	> 2 PM	0.422	0.491	0.912	0.482	0.244	0.726
S.D. /day	< 2 PM	4.002	3.345	5.203	3.868	3.156	4.991
	> 2 PM	0.599	0.933	1.067	0.568	0.433	0.705
S.D./week	< 2 PM	8.909	7.477	11.629	8.660	7.055	11.160
	> 2 PM	1.188	1.705	2.390	1.169	0.765	1.551
Average # of Patients in the system		15.230	6.975	22.220	15.170	7.026	22.196
Bed occupation		76%	55%	80%	75%	55%	80%
Overcrowding		16.4%	8.3%	7.1%	15.6%	7.5%	6.7%

Table 6. Simulated results with Normal distributed arrival process and Triangular distributed arrival process, n = 500 runs

5. Simulation and data analysis

In Chapter 4, a simulation model was developed that displays a simplified reality. This fifth chapter contains the simulation results and lists several options for lowering the variation in the number of arrivals, such that the overcrowding is not higher than 2%.

Section 5.1 describes how the number of runs in each experiment is determined. Section 5.2 depicts the different simulation configurations and the results, with a brief analysis of the numbers. Section 5.3 concludes with the implications for PVI, the user.

5.1. Determining the number of runs for each experiment

Each experiment is a simulation with a single configuration of variables. Each experiment can contain multiple runs, in order to increase the reliability of the experiment results. The simulation simulates three years of arrivals and discharges in one run. The results from such a run, however, vary to a certain degree, mainly because of the random variable assumption in the arrival rate. One run takes approximately 3.7 seconds to run (Windows 10, Intel i5-8250, 8 GB RAM). Since the number of experiments, and hence the runtime, increases exponentially, this section explores the trade-off between running-time and reliability.

In order to find a suitable number of runs, we want the confidence interval (CI) to be as small as possible. Therefore, we simulate 200 runs and obtain the 95% CI for each variable using Student's t-value. The upper- and lower bound of the CI are calculated by

$$\bar{x} \pm t_{df} \frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the mean of the variable over the runs considered, t_{df} is the t-value for df degrees of freedom, s is the sample standard deviation, and n is the number of runs. We then obtain the spread of this CI, in percentage. The spread is calculated as

$$\frac{CI_{+} - CI_{-}}{\bar{x}},$$

where CI_{+} and CI_{-} are the upper- and lower bound of the CI, respectively, and $\bar{x}$ is again the mean of the variable over the runs considered.

A summary of the results for the Triangular and Normal distribution are provided in, respectively, table 15 and table 16 in Appendix VI. These tables show the width of the 95% CI interval, as a percentage of the mean value for that metric. For example, if the mean of a metric is 10 and the 95% CI width is 2%, it means that the 95% CI is $[9.90 , 10.10]$.

Considering the above, it is accepted that 100 runs are sufficient to get an accurate result from the simulation. More runs do not decrease the spread sufficiently to justify the increase in runtime. Namely, 100 runs per experiment have an approximate runtime of 6 minutes and 10 seconds. An example of the number configurations for the different variables and the following total runtime is given in Table 7. There are four variables in this research, namely the level of variation in the number of arrivals for the (1) URO and (2) GYN specialisation, and the percentage of discharges before 2 PM for the (3) URO and (4) GYN specialisation. Note that, with 100 runs and 16,384 columns on an Excel sheet, there cannot be more than 163 experiments at once.

Var 1	Var 2	Var 3	Var 4	Total experiments	Expected runtime (hrs)
2	2	2	2	16	1.64
3	3	2	2	36	3.70
3	3	3	3	81	8.33
3	3	4	4	144	14.80
3	3	5	5	225	23.13
4	4	5	5	400	41.11

Table 7. Possible number of configurations for the four variables and the subsequent runtime in hours

5.2. Experiment results and analysis

This section describes the results of different experiments. As we run experiments with two different distributions for the arrival rate, we conduct each experiment twice. We subsequently show whether there is any significant difference between the two approaches. If not, it indicates that for any ward the Normal distribution can be used to model the arrivals. If there is a significant difference, it shows that for any generalisation careful consideration must take place to find the right distribution.

5.2.1. Description of experiments

As only a configuration with less than 163 experiments can be run at once, we are limited to at most two variables with 3 values and two variables with 4 values. Several experiment configurations are run, in order to determine the effect of a decrease in the variability of the arrivals, an increase in the percentage of discharges before 2 PM, and the interaction between the two. The different configurations are listed in the Table 8.

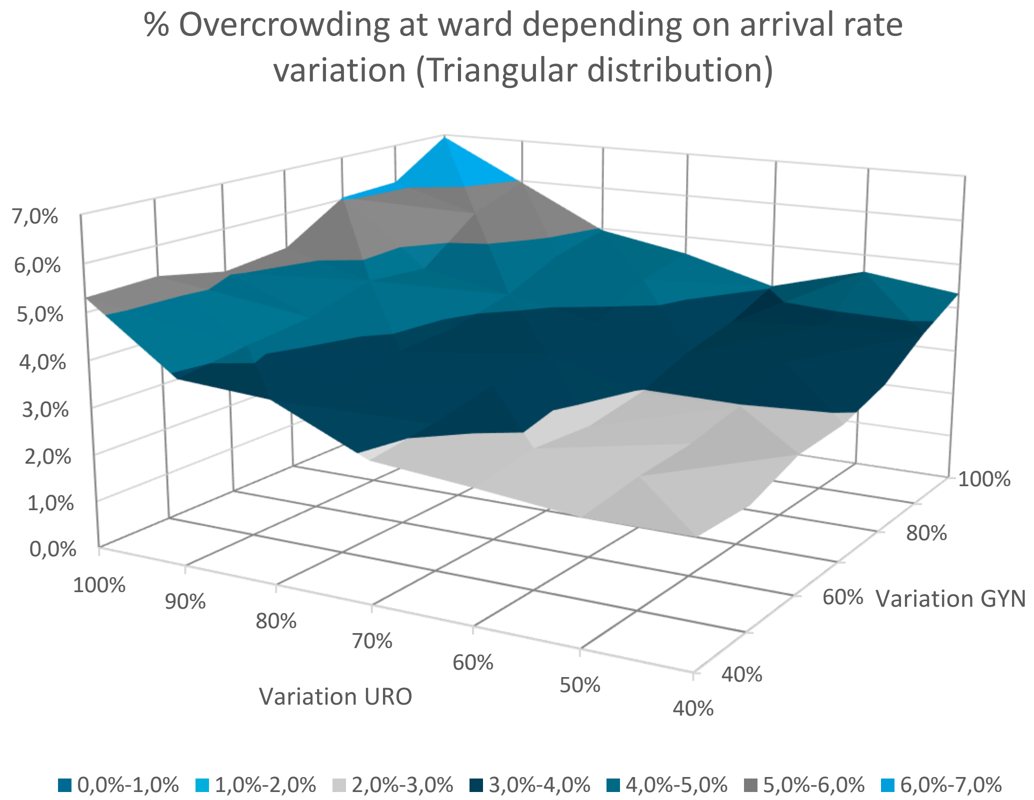
		URO			GYN		
Config.	Variable	From	To	Step	From	To	Step
1	Variation	40%	100%	10%	40%	100%	10%
1	Disch. time	100%	100%	0%	100%	100%	0%
2	Variation	100%	100%	0%	100%	100%	0%
2	Disch. time	46%	76%	5%	39%	63%	5%
3	Variation	50%	100%	25%	50%	100%	25%
3	Disch. time	46%	76%	10%	39%	69%	10%

Table 8. Experiment configurations

5.2.2. Results

This section displays the results of different experiments.

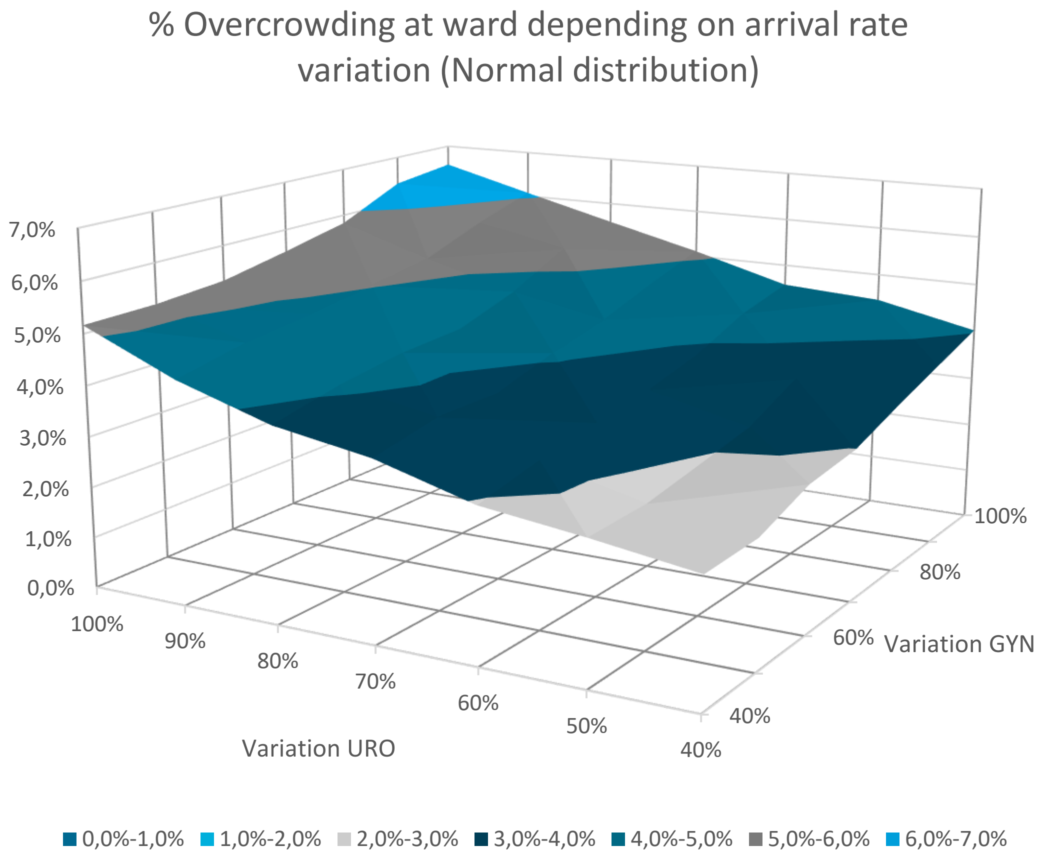
As can be seen, there is not much difference between the results of the Triangular arrival distribution and the Normal arrival distribution (Graph 5 and Table 9, and Graph 6 and Table 10), although the plane in Graph 6 is much smoother. The decrease in overcrowding is rather similar for the two distributions, and the decrease is consistent and thus independent from the level of variation of the other specialisation. For URO, the decrease of overcrowding is 0.43 and 0.42 percentage point (p.p.) for the Triangular and Normal arrival rate distribution, respectively, and for GYN these values are 0.27 and 0.28 p.p., respectively, for each 1/10th decrease in the arrival rate variation.



Graph 5. Relationship between the variation as a percentage of the historical variation and the overcrowding at the URO/GYN ward in percentage (Triangular arrival rate distribution)

URO\GYN	100%	90%	80%	70%	60%	50%	40%
100%	6,9%	6,1%	6,0%	5,2%	5,1%	5,3%	5,3%
90%	6,0%	5,5%	4,6%	4,7%	4,3%	4,1%	3,9%
80%	5,0%	4,7%	4,2%	3,9%	3,8%	3,6%	3,8%
70%	4,6%	4,4%	3,8%	3,5%	3,3%	3,1%	2,9%
60%	4,0%	3,6%	3,4%	3,0%	2,8%	2,9%	2,7%
50%	4,6%	3,8%	3,4%	3,0%	2,8%	2,7%	2,5%
40%	4,3%	3,9%	3,2%	2,9%	2,9%	2,5%	2,5%

Table 9. Table of the relationship between the variation as a percentage of the historical variation and the overcrowding at the URO/GYN ward in percentage (Triangular arrival rate distribution)



Graph 6. Relationship between the variation as a percentage of the historical variation and the overcrowding at the URO/GYN ward in percentage (Normal arrival rate distribution)

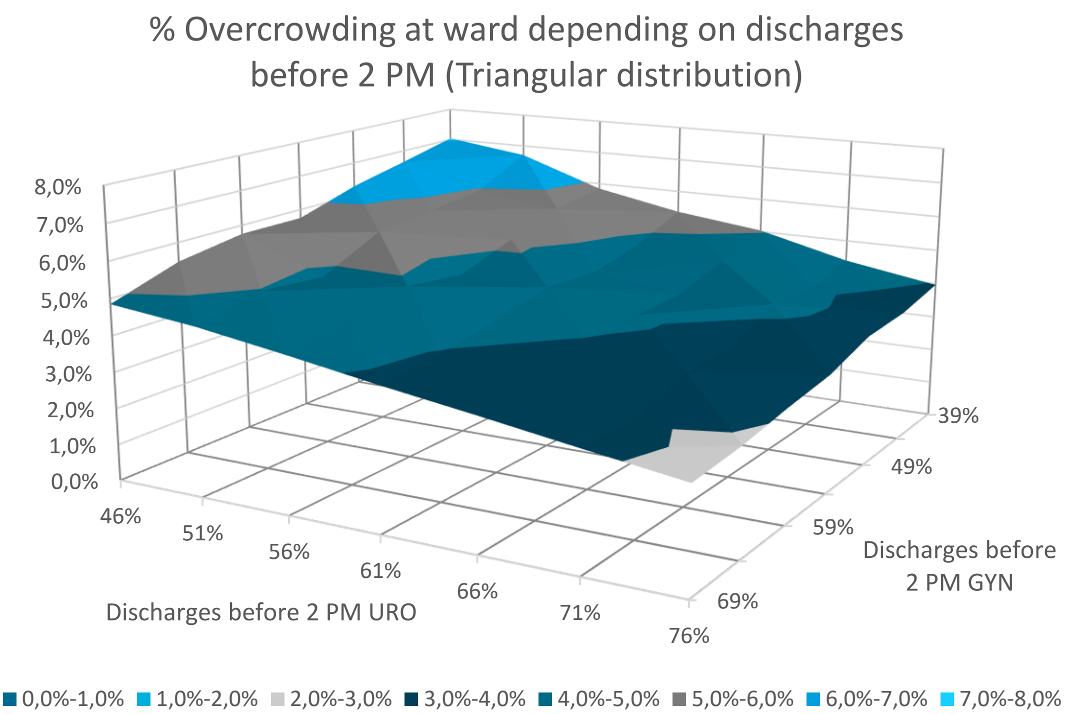
URO\GYN	100%	90%	80%	70%	60%	50%	40%
100%	6,6%	6,4%	5,8%	5,5%	5,3%	5,2%	5,2%
90%	6,1%	5,7%	5,2%	5,0%	4,8%	4,7%	4,4%
80%	5,6%	5,3%	4,7%	4,3%	4,3%	4,1%	3,8%
70%	5,1%	4,8%	4,4%	4,1%	3,7%	3,7%	3,5%
60%	4,6%	4,3%	3,9%	3,6%	3,4%	3,2%	3,0%
50%	4,5%	4,0%	3,6%	3,1%	2,9%	2,7%	2,7%
40%	4,0%	3,7%	3,4%	3,0%	2,9%	2,5%	2,4%

Table 10. Table of the relationship between the variation as a percentage of the historical variation and the overcrowding at the URO/GYN ward in percentage (Normal arrival rate distribution)

The effect of decreasing the variation in arrival rate diminishes however, as is shown in Graphs 19 through 22, which are found in Appendix VII. Note that the values on the vertical axis are the *decrease* in overcrowding in p.p. The relationship between the variation in arrival rate and overcrowding at the nursing ward therefore has the shape of a negative exponential curve. One can indeed expect the returns from decreasing the variation to diminish. Even if there is no variation in the arrival rate, there is still variation in the LOS of patients. Hence, regarding the decrease in variation, there is a lower bound of overcrowding that can only be lowered by decreasing the variation in LOS (or increasing the available capacity, i.e. the number of beds, but that option is outside the scope of this research). The ratio of the effect on the overcrowding of the variation in arrival rate versus the effect of the variation in LOS shifts towards the latter, which means that the effect of the variation in LOS becomes relatively larger. Hence, one can expect a negative exponential curve of the effect of a decrease in the arrival rate variation, which moves towards a lower bound that is larger than zero.

Regarding the effect of a change in the percentage discharged patients before 2 PM (Graph 7 and Table 11, and Graph 8 and Table 12), we observe that with a Normal arrival distribution there overcrowding is, in general, lower than if we use a Triangular arrival distribution. This difference is small, however. From the data of the simulation with the Triangular arrival rate distribution we observe that the effect of an extra 5 p.p. discharges before 2 PM for URO (GYN) patients, results in an average 0.50 (0.36) p.p. decrease in the overcrowding when 39% (46%) of GYN (URO) patients are discharged before 2 PM. When 69% (76%) of GYN (URO) patients are discharged before 2 PM, the effect of another 5 p.p. discharges before 2 PM for URO is only 0.35 (0.20) p.p. From the data obtained with a Normal arrival rate, these values are 0.49 (0.35) p.p. per 5 p.p. increase in the discharges before 2 PM for URO (GYN) patients when the discharges for GYN (URO) are 39% (46%), and 0.33 (0.19) p.p. when the discharges are 69% (76%). The effects obtained by the two arrival rate distributions are rather similar.

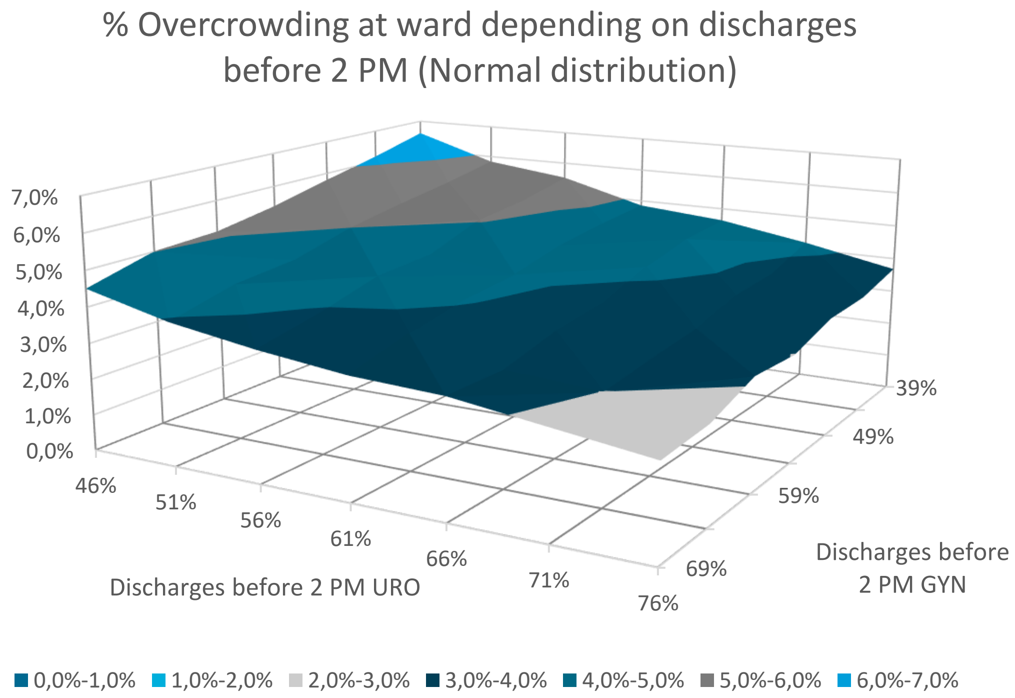
The decrease in the effect on the overcrowding of reducing the arrival rate variation is not observed when increasing the percentage discharges before 2 PM, as can be seen in Graphs 23 through 26 in Appendix VIII. The explanation is that the variation in LOS is not of influence on the percentage of discharges before 2 PM. Rather, more discharges before 2 PM free up space for arrivals that morning. Since the absolute amount of 5 p.p. more discharges is equal for any percentage, even so many arrivals can be placed at the ward instead of being off-serviced. We cannot conclude any trend from the Graphs 23 through 26, however. Further in Appendix VIII we included Graph 27 and 28 that more clearly show the trend, per specialisation, of discharging more patients before 2 PM (Normal arrival rate distribution only for simplicity). From these graphs, it seems this trend is linear or slightly negative exponential. We cannot draw a clear conclusion on this from these graphs, however.



Graph 7. Relationship between the percentage discharges before 2 PM of both specialisations and the overcrowding at the URO/GYN ward in percentage; historical variation and Triangular arrival rate distribution

URO\GYN	39%	44%	49%	54%	59%	64%	69%
46%	7,0%	6,6%	6,3%	5,8%	5,8%	5,5%	4,9%
51%	6,7%	6,0%	5,8%	5,6%	4,8%	5,0%	4,6%
56%	5,8%	5,6%	5,2%	4,6%	4,9%	4,5%	4,2%
61%	5,4%	4,8%	4,5%	4,6%	4,3%	3,9%	3,9%
66%	5,0%	4,6%	4,2%	4,2%	3,9%	3,5%	3,5%
71%	4,4%	4,1%	4,2%	3,6%	3,6%	3,1%	3,1%
76%	4,0%	3,7%	3,6%	3,2%	3,1%	2,9%	2,8%

Table 11. Table of the relationship between the percentage discharges before 2 PM of both specialisations and the overcrowding at the URO/GYN ward in percentage; historical variation and Triangular arrival rate distribution

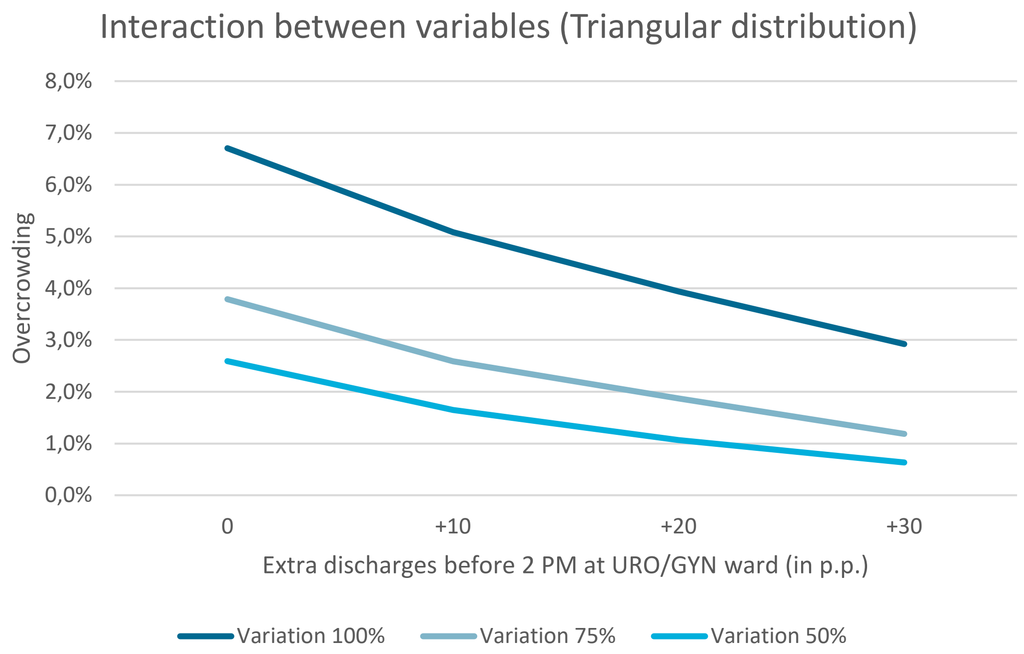


Graph 8. Relationship between the percentage discharges before 2 PM of both specialisations and the overcrowding at the URO/GYN ward in percentage; historical variation and Normal arrival rate distribution

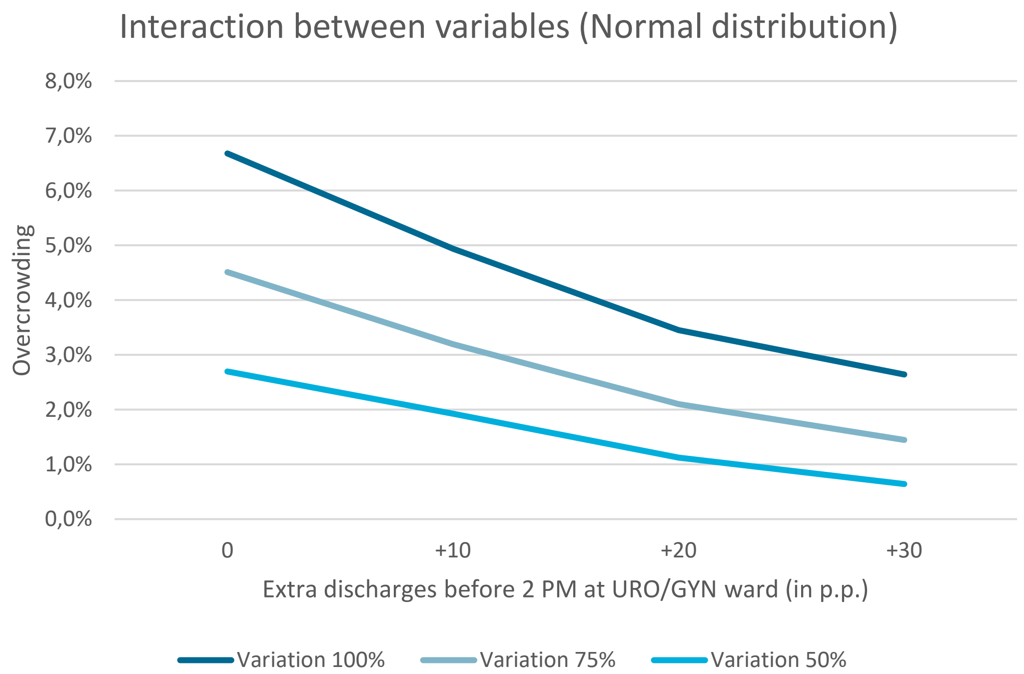
URO\GYN	39%	44%	49%	54%	59%	64%	69%
46%	6,6%	6,2%	5,8%	5,4%	5,1%	5,0%	4,5%
51%	5,8%	5,5%	5,3%	5,0%	4,6%	4,4%	3,9%
56%	5,5%	5,2%	5,0%	4,5%	4,4%	3,9%	3,6%
61%	4,8%	4,7%	4,6%	4,2%	4,1%	3,7%	3,3%
66%	4,6%	4,5%	4,2%	3,7%	3,5%	3,2%	3,2%
71%	4,2%	3,8%	3,7%	3,3%	3,2%	3,0%	2,9%
76%	3,7%	3,4%	3,4%	3,0%	3,1%	2,6%	2,5%

Table 12. Table of the relationship between the percentage discharges before 2 PM of both specialisations and the overcrowding at the URO/GYN ward in percentage; historical variation and Normal arrival rate distribution

Realistically, overcrowding cannot be reduced to 2% with a reduction in the variation of planned arrivals alone. It is not realistic to expect the variation can be reduced to 40% of historical variation. Similarly, it is not realistic to assume more than an additional 30 p.p. of patients can be discharged before 2 PM. Hence, also by discharging patients earlier we are not able to decrease overcrowding to 2%.



Graph 9. Effect of the combination of a decrease in variation and increase in discharges before 2 PM at the URO/GYN ward (Triangular distributed arrival rate)

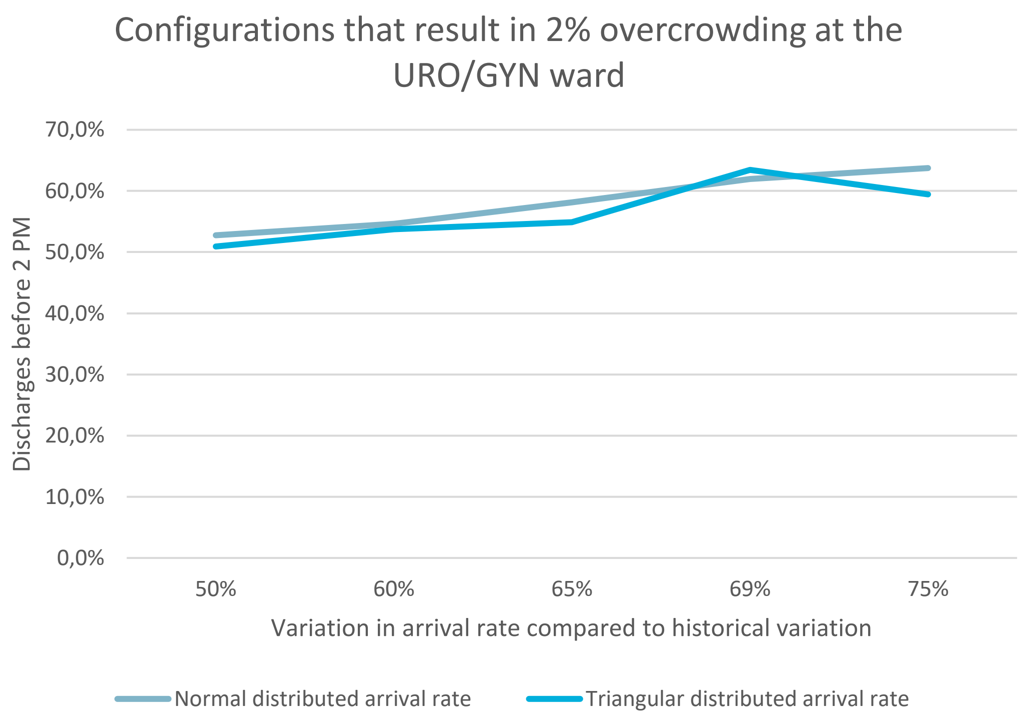


Graph 10. Effect of the combination of a decrease in variation and increase in discharges before 2 PM at the URO/GYN ward (Normally distributed arrival rate)

Graph 9 and Graph 10 show the interaction between variables. There is not much difference between the graphs, except for the line of 75% variation. We observe that the line obtained with a Normally distributed arrival rate reports an average 0.5 p.p. higher overcrowding. The difference decreases with more discharges before 2 PM. These graphs show that a combination of a reduction in variation and more discharges before 2 PM result in 2% or less overcrowding.

From Graphs 9 and 10, it seems that there is a decreasing effect of discharging patients before 2 PM. As has been explained before and shown in Appendix VIII, we observed a linear or slightly negative exponential trend. Therefore, Graph 9 and Graph 10 can be explained in two ways. First, we recognise the possibility of the occasional deviation from the trend, as is often observed in models dealing with random numbers. This, however, is unlikely since the decreasing effect is observed in all six lines in the two graphs. The second option is that there is indeed a slight decrease in effect of 5 p.p. more discharges before 2 PM, although this cannot be concluded from the Graphs 23 through 26 in Appendix VIII. An explanation is that extra discharges free up space for arriving patients, but on occasion, when there are more discharges than arrivals or when there is sufficient capacity for all arriving patients, discharging an extra patient does not contribute to decreasing the overcrowding.

The effect of a combination of a smaller arrival rate variation and more discharge before 2 PM is shown in Graph 11, where it becomes evident that the effect of more discharges before 2 PM have a larger effect on overcrowding than a reduction in arrival rate variation (currently, 43.1% of the patients at the ward are discharged before 2 PM). Furthermore, the normal arrival rate approach gives more consistent results, i.e. a smoother graph, than the Triangular arrival rate approach. There are no configurations of less than +30 p.p. discharges and an arrival rate variation of more than 75% of the historical variation that result in 2% discharges or less. Note that the variation on the horizontal axis is not linear, due to it being the weighted average of the configurations considered.



Graph 11. Interactions between reduction in variation and discharges before 2 PM at the ward that result in 2% overcrowding

Due to the simplicity and generality of the model, the first two KPIs cannot be compared to the base situation. Since from the model it is only known that a patient arrives or is discharged either before or after 2 PM, no specific median discharge time or average LOS can be calculated. One can reason, however, that these values are, with earlier discharges, lower than in the base situation. The off-service rate (in our model equal to $1 - \text{First-Time-Right rate}$) is 2%.

5.3. Implications for the model user

The results presented in section 5.2 display the effect of a decrease in the arrival rate variation and the discharges before 2 PM of the URO and GYN specialisations on the overcrowding at the combined nursing ward. The interaction between these variables is shown, to give direction on how overcrowding can be reduced to 2%. The model developed in this research hence serves to explore the effect of a reduction in arrival rate variation and more discharges before 2 PM on the overcrowding of a nursing ward. Furthermore, it can be used to show the effect of a less varying schedule to planning departments in order to convince them of the necessity to reduce the arrival rate variation of planned patients. This section describes the next steps for PVI.

5.3.1. Practical insights

The results from the model provide several interesting insights. The first is that the standard deviation per week is smaller than the expected standard deviation per week, as described in section 4.4.1. This implies that arrivals are in fact not random. Nevertheless, we observe that the overcrowding in the simulated model with the historical arrival rate and LOS is lower than in historically, see Table 5. According to one employee at PVI, this is because when there is a peak in URO patient arrivals, also GYN arrivals peak. The same holds for cases of few arrivals. Causes for these combined peaks and downs are, for example, holidays. Hence, the probability of arrivals cancelling each other out, i.e. many URO and few GYN arrivals or vice versa such that there are less excesses in the combined arrival rate, is lower historically. Therefore, the first suggestion is to ensure communication between the URO and the GYN planning department such that arrivals are more evenly spread over the year and the specialisations do not plan a peak number of surgeries at the same time.

Second, we observe that the effect of discharging patients earlier, i.e. before 2 PM, has a larger effect on the overcrowding than decreasing the variation in the arrival rate does. Furthermore, the effect of reducing the variation in the arrival rate has diminishing returns. Therefore, there is an optimum in the reduction in variation versus the increased number of discharges before 2 PM. This optimum does not consider what is feasible however, but it can be used to define the first point of improvement. The least improvement effort is found at this optimum, after all. Taking averages of the effect of the variables on the overcrowding, we can identify for which level of variation in the arrival rate the effect of discharging 5 p.p. more patients achieves a larger effect than reducing the arrival rate variation. This optimum for URO patients is 60% variation (both Triangular and Normal distributed arrival rate), and 60% (Triangular distributed arrival rate) or approximately 65% variation (Normal distributed arrival rate) for GYN patients.

Third, we suggest using this model with the Normal distribution for the arrival rate. The results from the Triangular distribution and the Normal distribution do not differ much, while for the Normal distribution it is easier to obtain necessary data, it is both easier to use, and gives smoother results which means it is easier to draw conclusions from the data.

5.3.2. The way forward

Now the results have been presented, an improvement goal can be set. This goal can be formulated from Graph 11, but the feasibility must be assessed by PVI. That is, PVI must assess how much variation can be reduced and they can then find the necessary extra discharges before 2 PM. This is the first step in the improvement cycle. This improvement cycle is called the PDCA (short for “Plan, Do, Check, Act”) cycle and the steps involve the following:

- **Plan**: The first step is identifying the problem, setting a goal, and identifying the necessary resources, criteria and constraints. It is at most a one-pager that gives direction to the improvement project and sets key performance indicators.
- **Do**: The second step involves defining concrete solutions. These solutions must be SMART (Specific, Measurable, Achievable, Relevant, and Time-bound). The solutions presented must be clear approaches towards achieving the goal that was set in the Plan-step. Tasks included in the solution approach must be assigned to specific functions, to ensure that the solution approach is also achievable in terms of human resources.
- **Check**: In the third step the solutions are reviewed. The solution approach that has the most promise is selected.
- **Act**: The fourth step is the implementation of the selected solution. Important is that the progress is monitored and (small) victories are celebrated.

It is important to involve all stakeholders in this improvement project since mutiny is likely to occur if employees do not identify with the suggested improvement or they feel that their jobs are affected without them having a say in the matter.

If, for some reason, it is not feasible to wait for a fully validated plan, or it is preferred to continue with implementing a solution as soon as possible, the PCDA can also be read as: Plan a solution, Implement the solution, Check and reflect on the results, and Act to maintain the situation.

PDCA is a cycle, since after the Act-step one can identify a new problem and goal and the cycle starts again. The first problem can, for example, be the high variation in the arrival rate at the URO/GYN ward and the goal is to reduce that variation for URO patients by 40% (since URO and GYN patients are planned separately). The next cycle can then focus on GYN patients, and a third cycle may focus on discharging patients earlier.

6. Generalisation to other wards or hospitals

This chapter briefly explores the hypothetical generalisation of the model to other wards or hospitals. Section 6.1 lists the requirements for adapting the simulation model to be used for other wards. Next, section 6.2 states the necessary changes to the model in order to incorporate more specialisations or wards. Lastly, section 6.3 identifies how the model can be adapted to be used in other hospitals.

6.1. Adaptation of the model to other wards

In order to use this simulation model for other specialisations and wards, the following inputs need to be obtained for each separate unit under study, where a unit is one of the two specialisations in this simulation, but can also be the entire URO/GYN ward as part of the C5West department:

- the arrival rate per day of patients before 2 PM and after 2 PM;
- the standard deviation of the arrival rate per day before 2 PM and after 2 PM;
- the LOS distribution of patients;
- the percentage discharges before 2 PM;
- the allocated capacity in the number of beds.

Changing the above-listed inputs in the model will make it available for analysis of other wards. The inputs need not be of a specialisation per se but can also be of a ward if one wants to conduct a larger analysis, for example of a department of 2 wards. The Triangular distribution is not considered here, since we suggested in section 5.3.1 to use the Normal distribution.

6.2. Inclusion of more specialisations

In order to analyse a ward with three specialisations, or a department with 3 or more wards, the model must change in three ways. First, the Dashboard and Simulation Output sheets need to be adapted in order to accommodate the extra inputs. Second, extra sheets must be opened, to include the LOS of these patients, and the three years of arriving and departing patients.

Third and lastly, the largest changes must be made in the VBA code. As the Dashboard sheet has been changed and the VBA code names specific locations in the sheet (such as cell “E28”), these references to cells must be changed. Furthermore, although the code is written in a way that a large part of the code is used multiple times, Excel must be told there are more than two specialisations or wards and that it thus needs to run the code more than three times.

A drawback of adding extra specialisations or wards to the simulation is the increase in runtime, although it is expected that this is only a linear increase. Nevertheless, as the simulation of multiple experiments already takes considerable time, adding one extra specialisation or ward to the simulation increases the runtime by hours.

6.3. Adaptation of the model to other hospitals

In principle, we expect that the model can be adapted to be used by other hospitals to explore the effect of a reduction in the planned arrival rate variation and discharge time on a ward’s overcrowding.

However, the model is written on the data extract of the EPDS of the Radboudumc and therefore the data analysis documents (which are used to analyse and structure the data from the EPDS) need to be rewritten for the other hospital.

7. Conclusion, discussion, limitations, and suggestions for further research

This chapter presents the conclusion of the research in section 7.1 and discusses it further in section 7.2. In section 7.3 limitations and suggestions for further research are listed.

7.1. Conclusion and directives to the Radboudumc

This research has answered the question to what level the variation in the number of arriving patients must be decreased to achieve a 98% probability that an arriving patient can be placed at the right nursing ward. We have done so by presenting a simple discrete event simulation model in Microsoft Excel that can easily be adapted to accommodate any other ward at the Radboudumc. Currently, approximately 8.9% of the arriving patients to this ward are off-serviced to other wards, while arrival rates vary from 0 to as many as 28 arrivals per day.

From the results of this model we concluded that the variation in arrivals at the URO/GYN nursing ward must be decreased by 25% if 64% of the patients at the ward can be discharged before 2 PM, or the variation must be decreased by 50% if 53% of patients can be discharged before 2 PM. The relation between the variables is approximately linear. More specifically, the minimum and maximum arrival rate at the ward are approximately between 3 and 24 patients per day (75% variation), and 6.5 and 20 patients per day (50% variation). These values are obtained following the Normal arrival rate distribution by multiplying the standard deviation of the number of arrivals of the specialisations with the suggested decrease in variation and adding (subtracting) two times this decreased standard deviation to (from) the mean number of arrivals to obtain the maximum (minimum) number of arrivals of that specialisation. Added, they result in the minimum and maximum arrivals at the ward. Note that this calculation is based on a decrease in the arrival rate variation of the specialisation, while the decrease in arrival rate variation at the ward is smaller. This is, however, the same way the model presented in this research works.

This model thus gives insight in the effect of the arrival rate variation and the discharges before 2 PM on the overcrowding of a hospital ward. It provides the surgery schedulers of the Radboudumc with guidelines for the number of patients that they can schedule during a working day, and at the same time, shows the necessity of early discharges. It is furthermore expected that less appointments have to be cancelled due to overcrowded wards. PVI can now set a quantified goal and its role is to facilitate change by showing the need for change to the scheduling department and the ward, supporting the change by suggesting process improvement steps, and controlling on the improvement process such that progress is continuous and maintained. These improvement steps are described in section 5.3.

Scientifically, this research suggests a new way of modelling patient flow that is not exact, but sufficient for practical purposes. It gives insight into the effect of the RTDC principle of early discharges and the combination of early discharges with less arrival rate variation in planned patients. Furthermore, this research presents two cases in which the arrival rate cannot be modelled according to any distribution commonly found in the literature on hospital ward arrival rates. Lastly, this research confirms the notion in earlier studies that patient flow can be well-modelled in a spreadsheet program.

The model is available upon request at e.w.hans@utwente.nl.

7.2. Discussion and limitations

This research focussed on developing a model to gain insights into the effect of a decrease in the arrival rate variation at the URO/GYN ward at the Radboudumc. Rather than developing a complicated mathematical or (discrete) simulation model where each arrival is generated separately, we developed a simple discrete event simulation model that incorporated the RTDC principles and that can be easily used by PVI to simulate other wards. To obtain a simple model, we had to allow for several assumptions. First, it is assumed that it does not matter at what time a patient is discharged, provided it is before 2 PM. Although the assumption is justified by the Flow Coordinator, one cannot expect there to be much difference between discharging a patient, in the most extreme case, at 13:59 PM instead of 14:01 PM. Since for the model this assumption draws a strict line between discharge times, it may explain the large effect of discharging patients before 2 PM, depicted by the model.

Blocking, as calculated by the Erlang-B (also known as the Erlang Loss) formula, can also be considered as off-servicing a patient. The ward of primary choice blocks a patient and (s)he has to be placed in another ward. The Erlang-B formula considers Poisson arrivals and exponential service times and is given in Appendix IX. That this is not the right approach is shown by the result for an average of 13.5 arrivals per day and a capacity of 32 beds, equal to the URO/GYN ward. and it gives a blocking probability of 7.16×10^{-6}, which is clearly far from the results obtained from the model in this research.

As noted, the percentage overcrowding in the base situation was already lower than historically (considering all assumptions, see Table 5 and Table 6). Therefore, the 2% as obtained from the simulation model may be higher in practice. One can expect, however, that with certain assumptions a more general model is obtained. It is expected that the assumptions listed in section 4.1 account for the difference between the model and the historical situation. Since the model using the Triangular or Normal distribution for the arrival rate instead of the historical distribution did not yield notable differences, we accept that the ward can be modelled with one of the two suggested distributions.

Concerning the arrival rate distribution, there is not much difference in the results obtained from the Normal or Triangular distribution, except that the Triangular distribution gives more varying results. For example, the Triangular distributed arrival rate suggests higher overcrowding when considering the effect of early discharges or arrival rate variation reduction separately but suggests lower overcrowding when considering the interaction between the variables than the results obtained with a Normal distributed arrival rate do. For consistency reasons, we therefore suggest using the Normal distribution. Furthermore, we note that the results from the Normal distributed arrival rate match the results from the simulation with the historical arrival rate distribution (right side of Table 5) more closely than the results from the simulation with Triangular distributed arrival rate do.

The assumption that arrivals are random appears to be invalid in practice. Since the standard deviation of arrivals per week is smaller than the expected standard deviation per week, calculated from the standard deviation per day. This was mentioned before in section 4.4.1. Furthermore, it was observed by PVI that when there is a peak in scheduled URO patients, there is also a peak in scheduled GYN patients. This is of course not represented by the model, as the arrivals are assumed to occur randomly. It is expected that the overcrowding is lower if arrivals are indeed random, hence the model presumably underestimates the level of overcrowding.

The historical arrivals considered in this research represent the actual arrivals to the hospital. However, there are some patients whose appointments are cancelled before they arrive at the hospital and hence those patients are not registered by the EPDS. These cancelled patients are not taken into account when considering the planning variation, which is therefore a little higher than assumed in this research. The arrival rate in the model can however easily be adapted to accommodate these patients, if it is known how many there are. Furthermore, there are some patients that are planned to have surgery on a Saturday or Sunday. There are few of these patients, and they are not incorporated in the model.

Furthermore, the model assumes every patient is discharged after at most 29 days. As earlier mentioned, it happens that a patient stays for over 3 months, and all this time (s)he occupies a bed. One outlier therefore has great impact on the bed occupation and can explain the difference in the simulated and historical overcrowding.

Lastly, it must be noted that the program delivered unusable results on several occasions. The cause is unknown and in a new run without adaptations to the model or the VBA code it delivered usable results, but at several times the simulation seems to have gotten stuck on the result of one run. That is, it seems that the model did not recalculate the sheet with results fast enough before it started to note down the results of the run in the “Simulation Output” sheet (automatic sheet updates are switched off to decrease the runtime). This resulted in the repetitive registration of the results of one run, up to 4,500 times (45 experiments, each with 100 runs). Therefore, caution must be taken when concluding the simulation results. The suggested and already implemented solution for this problem is to let the application ‘sleep’ for a short period of time, but it is unknown how much time of ‘sleep’ is required (adding 1 second to each run increases the total runtime of 36 experiments by an hour), and whether this delivers the necessary results.

7.3. Suggestions for further research

The first step in further research is to implement the recommendations from this research in practice, record the effects, and in doing so, validate the predictive power of the model presented in this research.

Next, when the model’s predictive value is determined, it can be adapted to other wards and be seen whether it delivers similar results. Since it is such a general model, however, we expect that this will not be a problem and that when the effects are validated for the URO/GYN ward it will be adaptable to accommodate other wards.

If there is no difference in the ease of discharging an URO or a GYN patient earlier, the model can be adapted in order to eliminate one of the variables, namely by changing the two variables for discharge time into one variable. This greatly reduces the number of experiments, and hence the runningtime for the simulation. Since there is currently a difference in the discharge time for both specialisations, this research considered the variables separately, however.

Finally, one could think of developing the model in a program that does not show glitches as Excel did on several occasions. Developing the model in another program furthermore decreases the runtime, provided the right model is used. Microsoft Excel is not built for simulations of this kind; hence the runtime is long and apparently it is prone to errors. This way of modelling was necessary, however, for PVI to understand its ins and outs and in order to work with it and adapt it to other wards.

8. Bibliography

Arkin, B., & Leemis, L. (2000). Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process from Overlapping Realizations. *Management Science, 46*(7), 989-998. doi:10.1287/mnsc.46.7.989.12037

Atienza, N., García-Heras, J., Muños-Pichardo, J., & Villa, R. (2008). An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine, 27*, 1403-1420. doi:10.1002/sim.3029

Bachouch, R., Guinet, A., & Hajri-Gabouj, S. (2012). An integer linear model for hospital bed planning. *International Journal of Production Economics, 140*, 833-843. doi:10.1016/j.ijpe.2012.07.023

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2016). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association, 23*, e2-e10. doi:10.1093/jamia/ocv106

Barton, P., Bryan, S., & Robinson, S. (2004). Modelling in the economic evaluation of health care: selecting the appropriate approach. *Journal of Health Services Research & Policy, 9*(2), 110-118. doi:10.1258/135581904322987535

Belciug, S., & Gorunescu, F. (2015). Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionalry computation. *Journal of Biomedical Informatics, 53*, 261-269. doi:10.1016/j.jbi.2014.11.010

Belciug, S., & Gorunescu, F. (2016). A hybrid genetic algorith-queuing multi-compartment model for optimizing inpatient bed occupancy and associated costs. *Artificial Inteligence in Medicine, 68*, 59-69. doi:10.1016/j.artmed.2016.03.001

Bittencourt, O., Verter, V., & Yalovsky, M. (2018). Hospital capacity management based on the queueing theory. *International Journal of Productivity and Performance Management, 67*(2), 224-238. doi:10.1108/IJPPM-12-2015-0193

Cooper, K., Brailsford, S., & Davies, R. (2007). Choice of Modelling Technique for Evaluating Health Care Interventions. *The Journal of the Operational Research Society, 58*(2), 168-176. doi:10.1057/palgrave.jors.2602230

Coulombe, K., & Rosow, E. (2004). The perfect match for throughput management: Clinical bed management and a patient-focused bed management dashboard. *Nurse Leader, 2*(5), 39-41. doi:10.1016/j.mnl.2004.07.008

Davies, R., & Davies, H. (1994). Modelling Patient Flows and Resource Provision in Health Systems. *Omega, 22*(2), 123-131. doi:10.1016/0305-0483(94)90073-6

De Anda, R. (In Press). Stop the Bottleneck: Improving Patient Throughput in the Emergency Department. *Journal of Emergency Nursing*. doi:10.1016/j.jen.2018.05.002

de Bruin, A., van Rossum, A., Visser, M., & Koole, G. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science, 10*, 125-137. doi:10.1007/s10729-007-9009-8

Dhingra, V., Kumawat, G., Roy, D., & de Koster, R. (2018). Solving semi-open queuing networks with time-varying arrivals: An application in container terminal landside operations. *European Journal of Operational Research, 276*, 855-876. doi:10.1016/j.ejor.2017.12.020

Elkhuizen, S., Bor, G., Smeenk, M., Klazinga, N., & Bakker, P. (2007). Capacity management of nursing staff as a vehicle for organizational improvement. *BMC Health Services Research, 7*, 197-206. doi:10.1186/1472-6963-7-196

Everett, J. (2002). A Decision Support Simulation Model for the Management of an Elective Surgery Waiting System. *Health Care Management Science, 267*, 89-95. doi:10.1023/A:1014468613635

Fackrell, M. (2009). Modelling healtcare systems with phase-type distributions. *Health Care Management Science, 12*, 11-26. doi:10.1007/s10729-008-9070-y

Fisher, W., & Meier-Hellstern, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation, 18*, 149-171. doi:10.1016/0166-5316(93)90035-S

Grand, M., & Putter, H. (2016). Regression models for expected lenth of stay. *Statistics in Medicine, 35*(7), 1178-1192. doi:10.1002/sim.6771

Green, L., & Nguyen, V. (2001). Strategies for Cutting Hospital Beds: The Impact on Patient Service. *Healt Services Research, 36*(2), 421-442. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089232/

Green, L., Soares, J., Giglio, J., & Green, R. (2006). Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Academic Emergency Medicine, 13*, 61-68. doi:10.1197/j.aem.2005.07.034

Gross, D., Shortle, J., Thompson, J., & Harris, C. (2008). *Fundamentals of Queueing Theory* (4th ed.). New Delhi: John Wiley & Sons, Inc.

Helm, J., AhmadBeygi, S., & Van Oyen, M. (2011). Design and Analysis of Hospital Admission Control for Operational Effectiveness. *Production and Operations Management, 13*, 359-374. doi:10.1111/J.1937-5956.2011.01231.x

Herwartz, H., Klein, N., & Strumann, C. (2016). Hospitalist bed management effecting throughput from the emergency department to the intensive care unit. *Journal of Applied Econometrics, 31*, 1159-1182. doi:10.1002/jae.2454

Howell, E., Bessman, E., Marshall, R., & Wright, S. (2010). Hospitalist bed management effecting throughput from the emergency department to the intensive care unit. *Journal of Critical Care, 25*, 184-189. doi:10.1016/j.jcrc.2009.08.004

Hu, X., Barnes, S., Marshall, R., & Wright, S. (2018). Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies. *International Transactions in Operational Research, 25*, 7-49. doi:10.1111/itor.12400

Hulshof, P., Boucherie, R., Hans, E., & Hurink, J. (2013). Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science, 16*, 152-166. doi:10.1007/s10729-012-9219-6

Ickowicz, A., & Sparks, R. (2017). Modelling hospital length of stay using convolutive mixtrue distributions. *Statistics in Medicine, 36*, 112-135. doi:10.1002/sim.7135

Kao, E., & Chang, S.-L. (1988). Modeling Time-Dependent Arrivals to Service Systems: A Case in Using a Piecewise-Polynomial Rate Function in a Nonhomogeneous Poisson Process. *Management Science, 34*(11), 1367-1379. doi:10.1287/mnsc.34.11.1367

Kim, S.-H., & Whitt, W. (2014a). Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? *Manufacturing & Service Operations Management, 16*(3), 464-480. doi:10.1287/msom.2014.0490

Kim, S.-H., & Whitt, W. (2014b). Choosing Arrival Process Models for Service Systems, Tests of a Nonhonmogeneous Poisson Process. *Naval Research Logistics, 61*(1), 61-90. doi:10.1002/nav.21568

Klein, M., & Reinhardt, G. (2012). Emergency Department Patient Flow Simulations Using Spreadsheets. *Simulation in Healthcare, 7*(1), 40-47. doi:10.1097/SIH.0b013e3182301005

Knessl, C., & Yang, Y. (2002). An Exact Solution for an M(t)/M(t)/1 Queue with Time-Dependent Arrivals and Service. *Qeueuing Systems, 40*, 233-245. doi:10.1016/j.spa.2013.02.002

Kumar, A., Costa, A., Fackrell, M., & Taylor, P. (2018). A sequential stochastic mixed integer programming model for tactical master surgery scheduling. *European Journal of Operational Research, 270*, 734-746. doi:10.1016/j.ejor.2018.04.007

Landa, P., Sonnessa, M., Tànfani, E., & Testi, A. (2018). Multiobjective bed management considering emergency and elective patient flows. *International Transactions in Operational Research, 25*, 91-110. doi:10.1111/itor.12360

Larsson, A., & Fredriksson, A. (2019). Tactical capacity planning in hospital departments. *International Journal of Health Care Quality Assurance, 32*(1), 233-245. doi:10.1108/IJHCQA-11-2017-0218

Leemis, L. (1991). Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process. *Management Science, 37*(7), 886-900. doi:10.1287/mnsc.37.7.886

Lega, F., & De Pietro, C. (2005). Converging patterns in hospital organization: beyond the professional bureaucracy. *Health Policy, 74*, 261-281. doi:10.1016/j.healthpol.2005.01.010

Lovett, P., Illg, M., & Sweeney, B. (2016). A Successful Model for a Comprehensive Patient Flow Management Center at an Academic Health System. *American Journal of Medical Quality, 31*(3), 246-255. doi:10.1177/1062860614564618

McCarthy, M., Zeger, S., Ding, R., Aronsky, D., Hoot, N., & Kelen, G. (2008). The Challenge of Predicting Demand for Emergency Department Services. *Academic Emergency Medicine, 15*, 337-346. doi:10.1111/j.1553-2712.2008.00083.x

McClean, S., Barton, M., Garg, L., & Fullerton, K. (2011). A modeling framework that combines Markov models and discrete-event simulation for stroke patient care. *ACM Transactions on Modeling and Computer Simulation, 21*(4), Article 25. doi:10.1145/2000494.20000498

Mitchell, B. (1971). A Comparison of Chi-Square and Kolmogorov-Smirnov Tests. *Area, 3*(4), 237-241. Retrieved from http://www.jstor.org/stable/20000590.

Moldovan, F. (2018). New Approaches and Trends in Health Care. *Procedia Manufacturing, 22*, 947-951. doi:10.1016/j.promfg.2018.03.135

Nunes, L., de Carvalho, S., & Rodrigues, R. (2009). Markov decision process applied to the control of hospital elective admissions. *Artificial Intelligence in Medicine, 47*, 159-171. doi:10.1016/j.artmed.2009.07.003

Resar, R., Nolan, K., Kaczynski, D., & Jensen, K. (2011). Using Real-Time Demand Capacity Management to Improve Hospitalwide Patient Flow. *The Joint Commission Journal on Quality and Patient Safety, 37*(5), 217-227. doi:10.1016/S1553-7250(11)37029-8

Soong, C., Wright, S., Kisuule, F., Masters, H., Pfeiffer, M., Kantsiper, M., & Howell, E. (2016). The Role of Hospitalists in Managing Patient Flow: Lessions from Four Hospitals. *Current Emergency and Hospital Medicine Reports, 4*, 159-163. doi:10.1007/s40138-016-0110-5

Standfield, L., Comans, T., & Scuffham, P. (2014). Markov Modeling and Discrete Event Simulation in Health Care: A Systematic Comparison. *International Journal of Technology Assessment in Health Care, 30*(2), 165-172. doi:10.1017/S0266462314000117

Standfield, L., Comans, T., & Scuffham, P. (2017). An empirical comparison of Markov cohort modeling and discrete event simulation in a capacity-constrained health care setting. *European Journal of Health Economics, 18*, 33-47. doi:10.1007/s10198-015-1756-z

Takagi, H., Kanai, Y., & Misue, K. (2017). Queueing network model for obstetric patient flow in a hospital. *Health Care Management Science, 20*, 433-451. doi:10.1007/s10729-016-9363-5

Tan, X., Knessl, C., & Yang, Y. (2013). On finite capacity queues with time dependent arrival rates. *Stochastic Processes and their Applications, 123*, 2175-2227. doi:10.1016/j.spa.2013.02.002

Villa, S., Barbieri, M., & Lega, F. (2009). Restructuring patient flow logistics around patient care needs: implications and practicalities from three critical cases. *Health Care Management Science, 12*, 155-165. doi:10.1007/s10729-008-9091-6

Vrbik, J. (2018). Small-Sample Corrections to Kolmogorov-Smirnov Test Statistic.

Wang, H.-M. (2009). Comparison of the Goodness-of-Fit Tests: the Pearson Chi-square and Kolmogorov-Smirnov Tests. *Metering Management Journal, 6*(1), 57-64. Retrieved 12 13, 2018, from https://pdfs.semanticscholar.org/0d3a/736708f50897690875b6e578e62657df437b.pdf

Wiler, J., Giffrey, R., & Olsen, T. (2011). Review of Modeling Approaches for Emergency Department Patient Flow and Crowding Research. *Academic Emergency Medicine, 18*, 1371-1379. doi:10.1111/j.1553-2712.2011.01135.x

Zychlinski, N., Mandelbaum, A., & Momcilovic, P. (2018). Time-varying tandem queues with blocking: modeling, analysis, and operational insights via fluid models with reflection. *Queueing Systems, 89*, 15-47. doi:10.1007/s11134-018-9578-x

Appendix I

Literature was obtained from two databases to which the Radboudumc has access:

- PubMed, https://www.ncbi.nlm.nih.gov/pubmed/
- Web of Science, https://apps.webofknowledge.com/

The search was limited to articles published in 1988 or later, i.e. 30 years or younger. Articles were selected based on their title (the number of articles selected is indicated by "selected *X* articles for further reading"). The number of articles was then limited by their abstract, which should at least contain a reference to the objective of the respective research questions. That means that in the abstracts of articles selected for section 3.1, the first sub-question, it should become clear that the article is about practices modelling and managing/improving patient flow from an operational and/or tactical perspective, and thus not a strategic perspective. For the second sub-question, discussed in section 3.2, it should be clear from the abstract that a comparison is made between queuing modelling, Markov modelling, and (discrete event) simulation, or between two of those three. For section 3.3, the abstract should contain reference to arrival patterns, or a way to describe, model, and compute time dependent arrivals. Finally, for the fourth sub-question, which is discussed in section 3.4, articles must describe a method to finding the LOS distribution, or list a (series of) distributions for the LOS of a hospital ward.

If in a selected article a reference was found to another article, not found via the above-mentioned search strategy, this article was also included if the author deemed it useful through selection based on the reference context, title, and abstract.

Search terms for the different research questions were the following:

Section 3.1: What is written in literature about managing and modelling patient flow at a tactical and an operational?

- PubMed
  - "Patient flow" AND "Operational": 86 hits
    - Added "Management": 39 hits. Selected 5 articles for further reading.
    - Added "Model": 34 hits. Selected 4 articles for further reading.
  - "Patient flow" AND Tactical: 0 hits.
  - Hospital AND Management AND Tactical: 0 hits.
- Web of Science
  - "Patient flow" AND "Operational *": 100 hits
    - Searched within results for "management": 42 hits. Selected 6 articles for further reading.
    - Searched within results for "model": 67 hits. Refined categories to 'Operations research management science', 'Health care sciences services', 'Management', 'Health policy services', and 'Medical informatics': 32 hits. Selected 3 articles for further reading.
  - "Patient flow" AND Tactical: 4 hits. Selected 2 articles for further reading.
  - Hospital AND Management AND Tactical: 76 hits. Selected 4 articles for further reading.

Section 3.2: What are the advantages of using simulation, queuing theory, or Markov chains?

- PubMed
  - "Health care" AND (Queuing OR Queueing): 97 hits
    - Selected 5 articles for further reading.

- Added “Patient flow”: 11 hits. Selected 3 articles for further reading.
- “Health care” AND Markov: 2233 hits
  - Added “comparison” and limited articles to last 5 years: 80 hits. Selected 2 articles for further reading.
  - Added “Patient flow”: 3 hits. Selected 1 article for further reading.
- “Health care” AND Simulation: 6028 hits
  - Added “Patient flow”: 64 hits. Selected 4 articles for further reading.
  - Added “Comparison” AND ((Queuing OR Queueing) OR Markov): 46 hits. No (new) articles selected for further reading.

- Web of Science
  - “Health care” AND (Queuing OR Queueing): 289 hits
    - Searched within results for “Patient flow”: 40 hits. Selected 1 article for further reading.
    - Searched within results for “Markov”: 24 hits. Selected 1 article for further reading.
    - Searched within results for “Simulation”: 88 hits. Selected 3 articles for further reading.
  - “Health care” AND Markov: 1095 hits
    - Searched within results for “Modelling” (1054 hits) and “Advantage”: 31 hits. No (new) articles selected for further reading.
    - Searched within results for “Patient flow”: 11 hits. Selected 1 article for further reading.
  - “Health care” AND Simulation: 3459 hits
    - Searched within results for “Discrete event” (289 hits) and “Review”: 47 hits. Selected 4 articles for further reading.

Section 3.3: What are often-found distributions for the arrival rate, and how can the time-dependent arrivals at hospital wards be modelled?

- PubMed
  - “Arrival rate” AND “Health care”: 10 hits. No (new) articles selected for further reading.
  - “Arrival rate” AND “time”: 55 hits. No (new) articles selected for further reading.
  - "Arrival" AND ("Queueing" OR "Queuing") AND "time": 43 hits. Selected 1 article for further reading.
  - “Arrival rate” AND “hospital”: 33 hits. Selected 2 articles for further reading.
  - "Arrival" AND “hospital” AND Distribution AND ("time varying" OR "time dependent"): 5 hits. Selected 1 article for further reading.
- Web of Science
  - “Arrival rate” AND “Health care”: 7 hits. Selected 1 article for further reading.
  - “Arrival rate” AND “time”: 1031 hits. Searched within results for “distribution”: 304 hits. Selected 10 articles for further reading
  - “Arrival rate” AND “hospital”: 32 hits. Selected 2 articles for further reading.
  - "arrival rate" AND "hospital" AND ("queuing" OR "queueing"): 16 hits. Selected 2 articles for further reading.

Section 3.4: How can the length of stay at hospital wards be modelled?

- PubMed
  - "Length of stay" AND "Health care": 17342 hits.
    - Added (“Queuing” OR “Queueing”): 17 hits. Selected 2 articles for further reading.

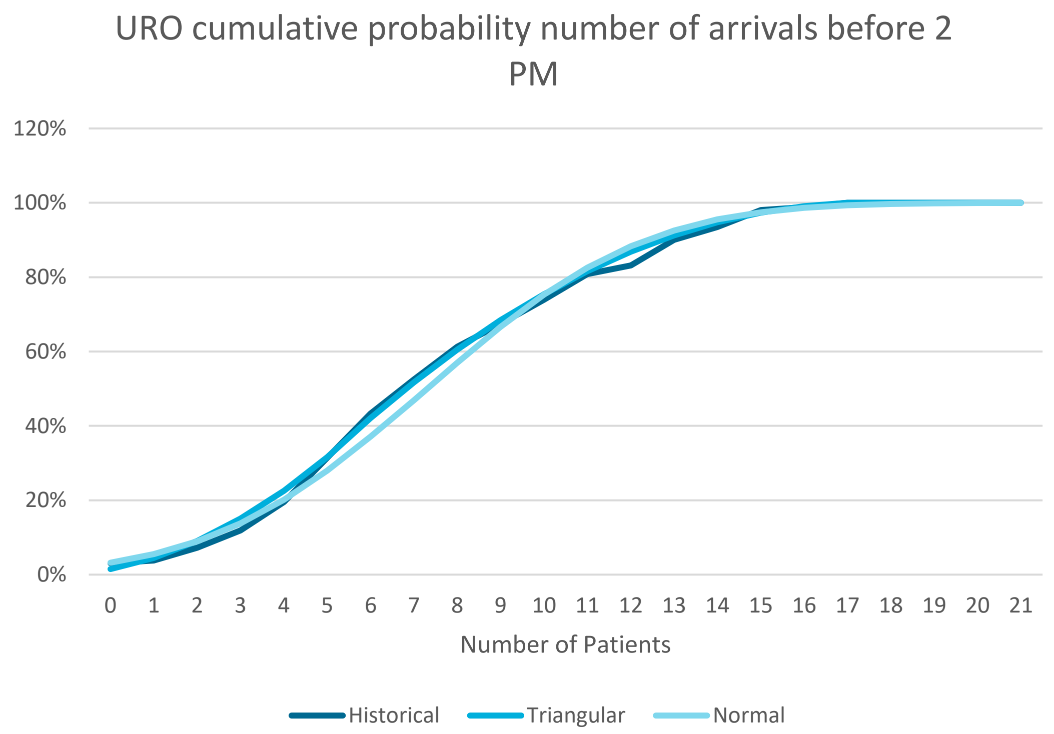
    - "Length of stay" AND “distribution” AND ward: 154 hits. Selected 3 articles for further reading.
    - “Length of stay” AND “general ward”: 115 hits. Selected no new articles for further reading.
- Web of Science
    - "Length of stay" AND "Health care": 4142 hits.
        - Added “Queuing” OR “Queueing”: 18 hits. Selected 2 articles for further reading.
    - "Length of stay" AND “distribution” AND ward: 38 hits. Selected 3 articles for further reading.
    - “Length of stay” AND “general ward”: 69 hits. Selected no new articles for further reading. Selected 1 article for further reading.

Appendix II

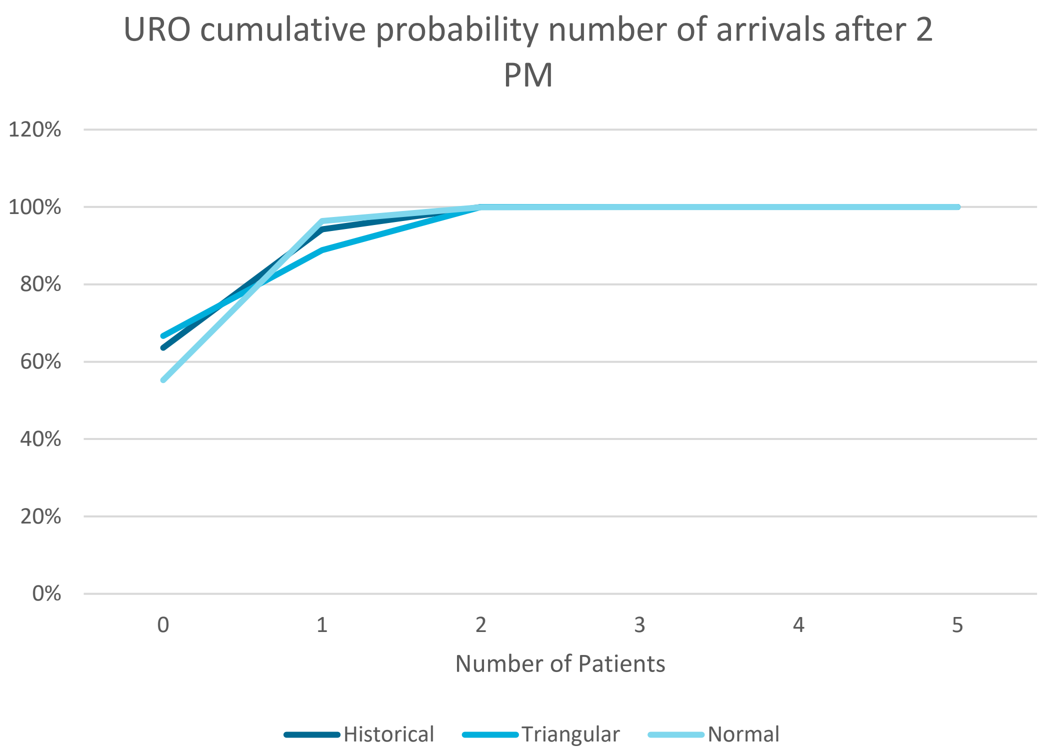
t	Urology total Arrivals/year	$m_{URO}(t,T)$	Gynaecology total Arrivals/year	$m_{GYN}(t,T)$
1	0	0	3	0,012
2	0	0	0	0
3	0	0	2	0,008
4	0	0	0	0
5	0	0	2	0,008
6	0	0	7	0,027
7	86	0,331	151	0,581
8	490	1,885	536	2,062
9	342	1,312	206	0,792
10	291	1,119	175	0,673
11	514	1,965	114	0,438
12	174	0,662	74	0,285
13	37	0,142	32	0,123
14	115	0,438	14	0,054
15	67	0,254	11	0,042
16	21	0,077	12	0,046
17	6	0,023	6	0,023
18	5	0,019	2	0,008
19	3	0,012	1	0,004
20	3	0,004	8	0,031
21	4	0,012	3	0,012
22	0	0	2	0,008
23	0	0	2	0,008
24	0	0	1	0,004

Appendix III

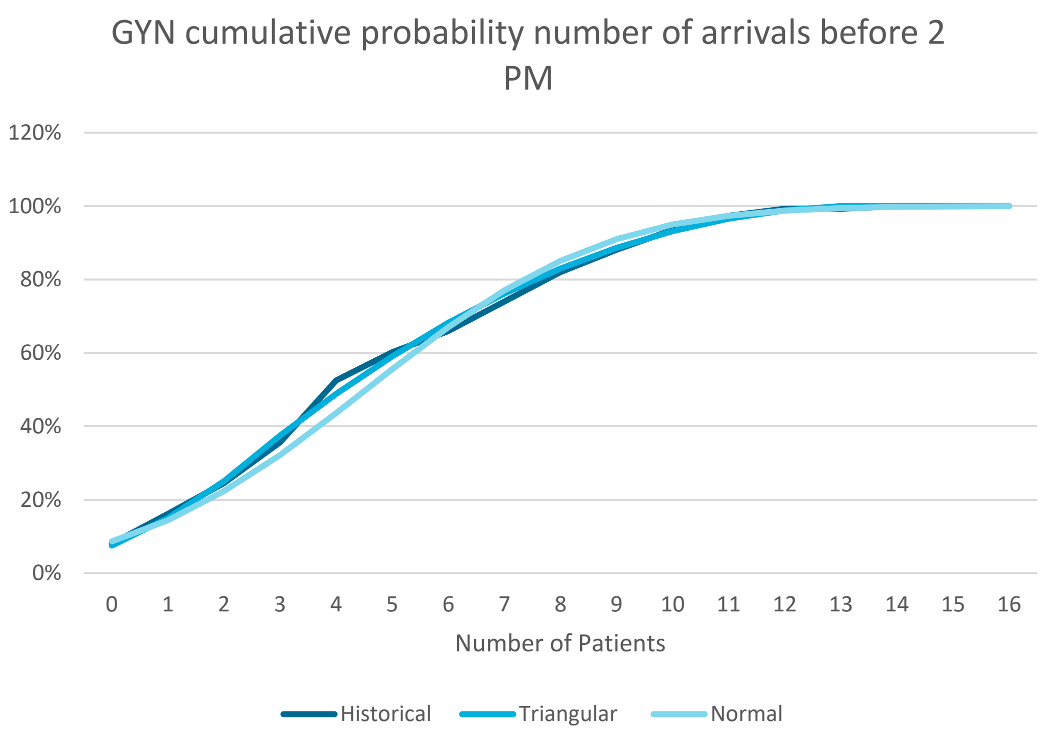
Below are distributions provided for the probability of n arriving patients. All distributions except the historical distributions are theoretical, i.e. they represent the expected number of arriving patients in the simulation if that arrival rate distribution is used. These distributions are obtained with the mean and standard deviation from the historical data for the normal distribution, and by calculating the best fit between the historical data and theoretical Triangular distribution using the Excel Solver function.



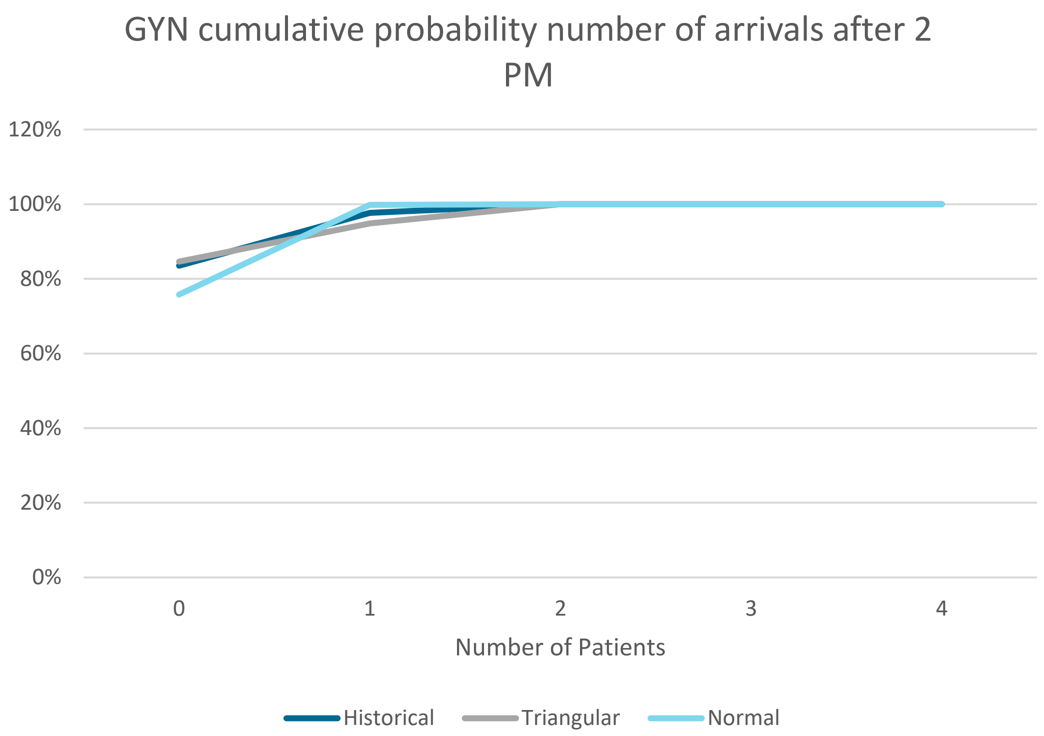
Graph 12. Cumulative distribution functions of the number of arriving URO patients before 2 PM (n_historical = 2158, t_historical = 365 days, source: EPDS)



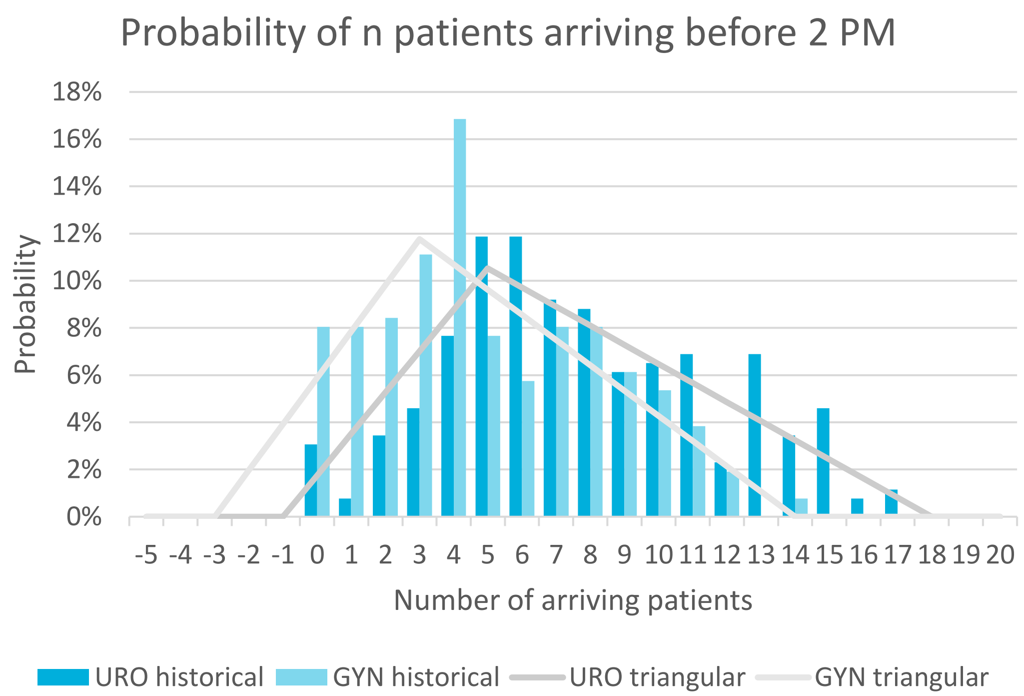
Graph 13. Cumulative distribution functions of the number of arriving URO patients after 2 PM (n_historical = 2158, t_historical = 365 days, source: EPDS)



Graph 14. Cumulative distribution functions of the number of arriving GYN patients before 2 PM (n_historical = 1403, t_historical = 365 days, source: EPDS)



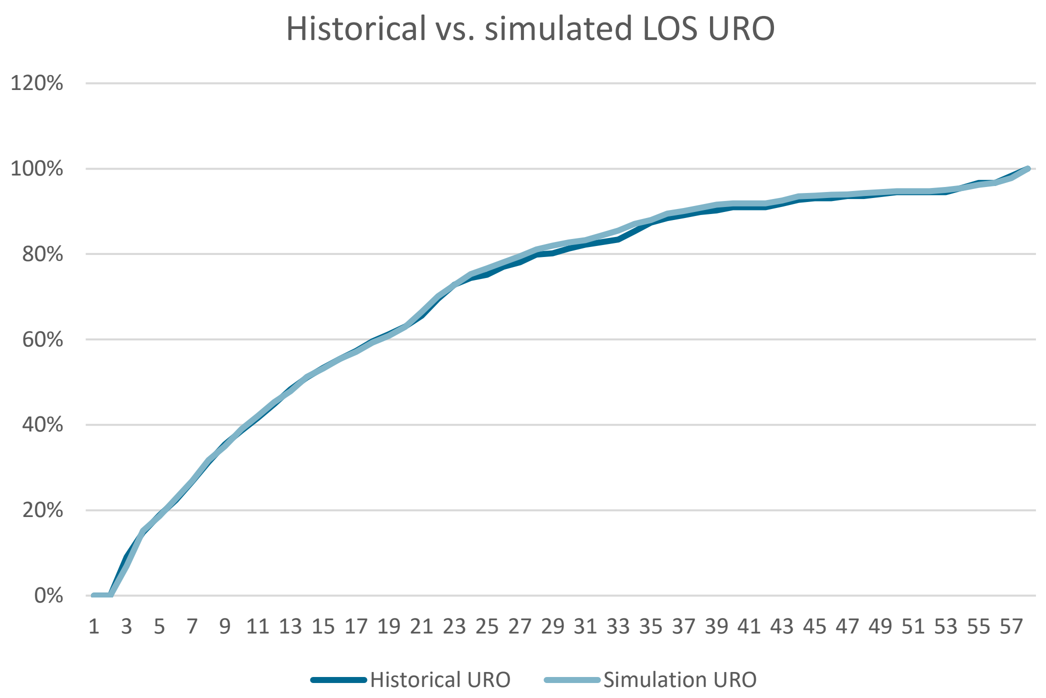
Graph 15. Cumulative distribution functions of the number of arriving GYN patients after 2 PM (n_historical = 1403, t_historical = 365 days, source: EPDS)



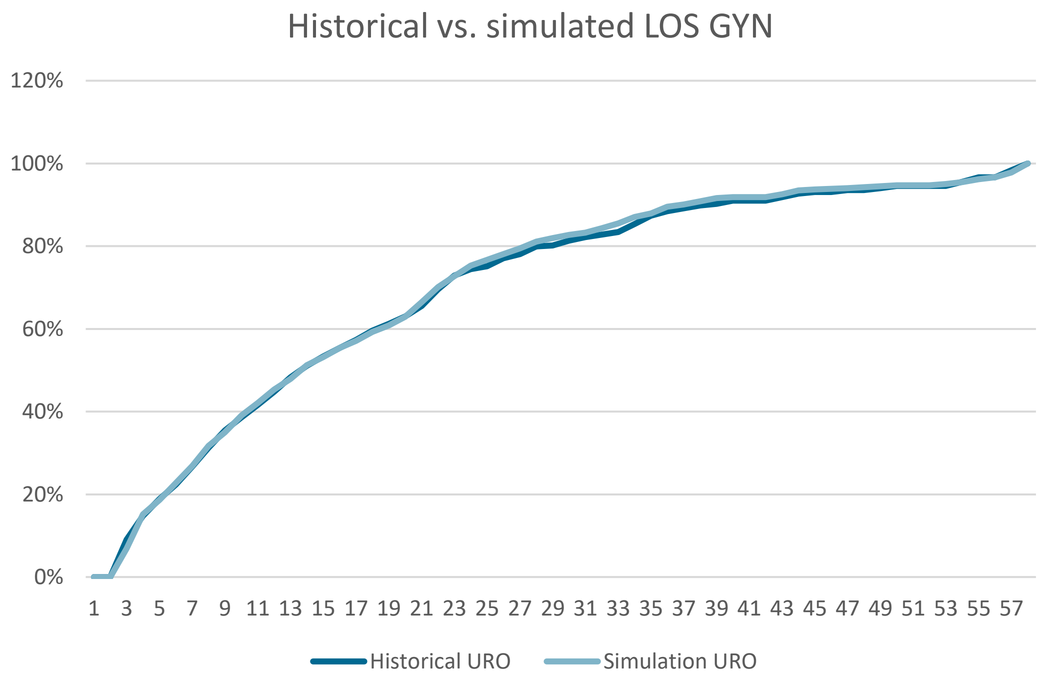
Graph 16. Probability density function for *n* arrivals before 2 PM for the URO and GYN specialisation

Note in Graph 16 that all negative arrival rate probabilities are changed to 0 arrivals in the simulation model.

Appendix IV



Graph 17. Cumulative distribution of historical LOS and simulated LOS of URO patients in number of days (n_historical = 2158, t_historical = 365 days, source = EPDS; n_simulated = 10,000)



Graph 18. Cumulative distribution of historical LOS and simulated LOS of GYN patients in number of days (n_historical = 2158, t_historical = 365 days, source = EPDS; n_simulated = 10,000)

Seasonality in LOS

Over the course of this research, the problem arose that the simulation did not represent reality at all. Therefore, all inputs were checked, and after long we concluded that the problem was that, in order to increase the reliability, we had taken data from 1 year and 10 months as input for the LOS. That resulted in a rather different distribution than if we had taken 1 year of data. First it was suggested this difference was attributable to seasonality, but it appeared that in March 2018 the Shot Stay Unit (SSU) was included in the URO/GYN ward. This implies more patients with a LOS of one day, as is observed in the data.

Table 13 shows the difference in LOS depending on the period under consideration.

LOS (days)	URO			GYN		
	1-9-2017 31-8-2018	1-12-2016 31-8-2018	1-9-2017 7-12-2018	1-9-2017 31-8-2018	1-12-2016 31-8-2018	1-9-2017 7-12-2018
1	31%	27%	30%	48%	52%	48%
2	65%	51%	65%	68%	69%	67%
3	74%	63%	74%	79%	78%	79%
4	81%	73%	81%	87%	85%	87%
5	85%	79%	85%	92%	91%	92%
6	88%	84%	88%	95%	94%	95%
7	91%	87%	91%	97%	96%	97%
8	92%	89%	92%	98%	97%	98%
9	93%	91%	94%	98%	98%	98%
10	94%	92%	95%	98%	98%	98%
11	96%	94%	96%	99%	98%	99%
12	97%	95%	97%	99%	99%	99%
13	97%	96%	97%	99%	99%	99%
14	98%	97%	98%	99%	99%	99%
15	98%	97%	98%	99%	99%	99%
16	98%	97%	98%	99%	99%	99%
17	98%	98%	99%	99%	99%	99%
18	99%	98%	99%	99%	99%	99%
19	99%	99%	99%	99%	99%	99%
20	99%	99%	99%	99%	100%	99%
21	99%	99%	99%	99%	100%	100%
22	99%	99%	99%	99%	100%	100%
23	99%	99%	99%	99%	100%	100%
24	99%	99%	100%	99%	100%	100%
25	100%	99%	100%	99%	100%	100%
26	100%	99%	100%	100%	100%	100%
27	100%	99%	100%	100%	100%	100%
28	100%	100%	100%	100%	100%	100%
More	100%	100%	100%	100%	100%	100%

Table 13. Cumulative distribution of patient LOS depending on the period under consideration

The difference is largest between the first and second column of both specialisations. Simulation results are notably different, especially for the URO specialisation (see Table 14).

URO	1-9-2017 31-8-2018	1-12-2016 31-8-2018
Bed occupation	75%	96%
Average # patients	15,046	19,226
Overcrowding %	15,8%	34,0%
Discharges < 2PM	46,4%	46,4%
GYN		
Bed occupation	54%	53%
Average # patients	6,896	6,809
Overcrowding %	8,0%	7,2%
Discharges < 2PM	39,5%	39,5%
Combined ward		
Bed occupation	79%	94%
Average # patients	21,94	26,03
Overcrowding %	6,7%	15,7%

Table 14. Simulation results from simulating with LOSs obtained from different time periods

In this research, data from one year (1-9-2017 through 31-8-2018) is taken.

Appendix V

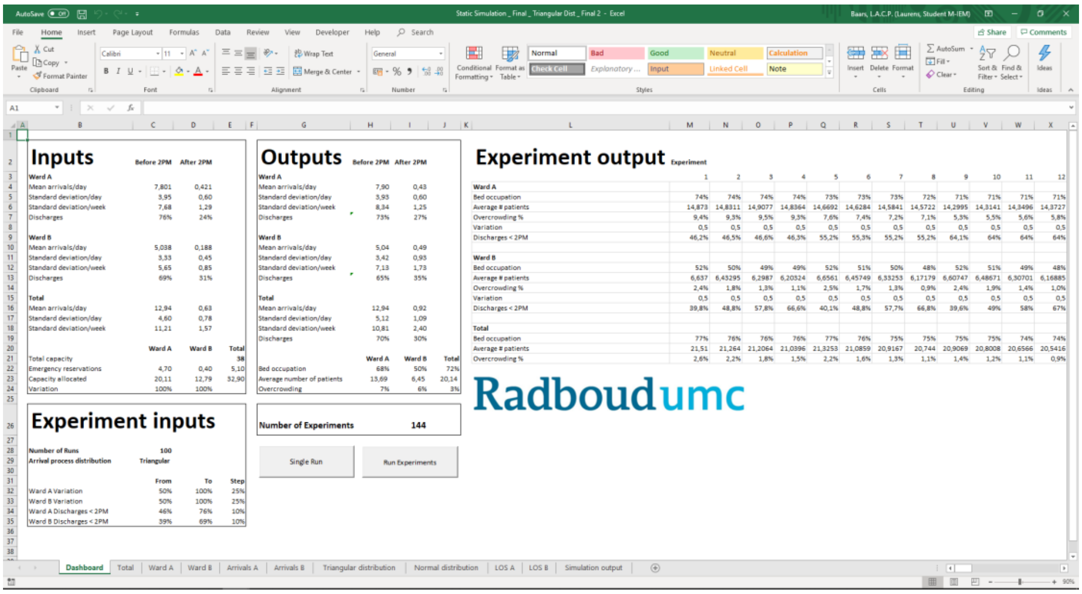


Figure 2. Dashboard of the simulation model

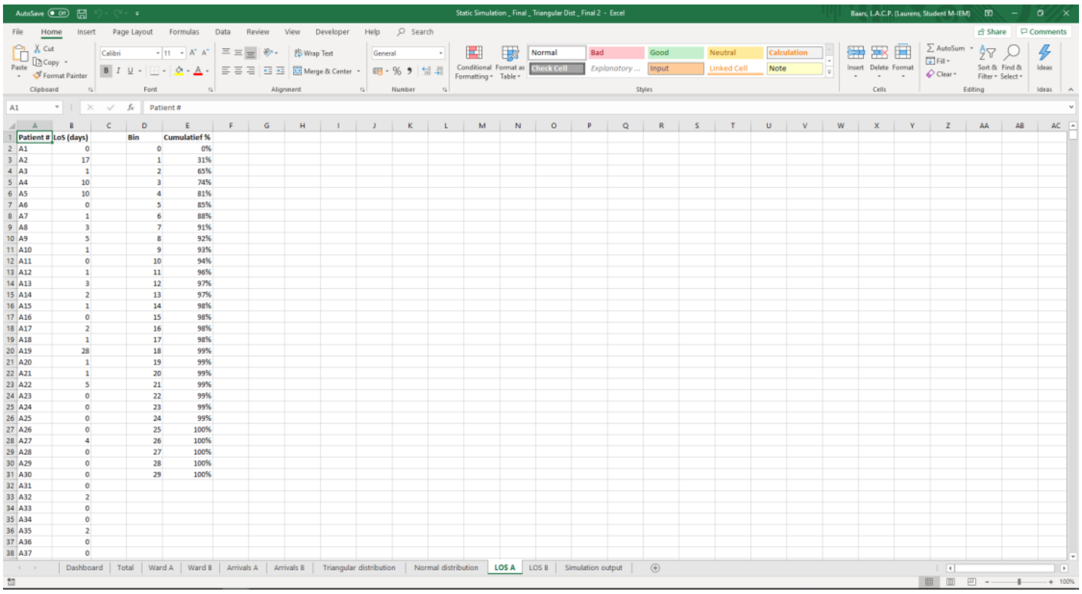
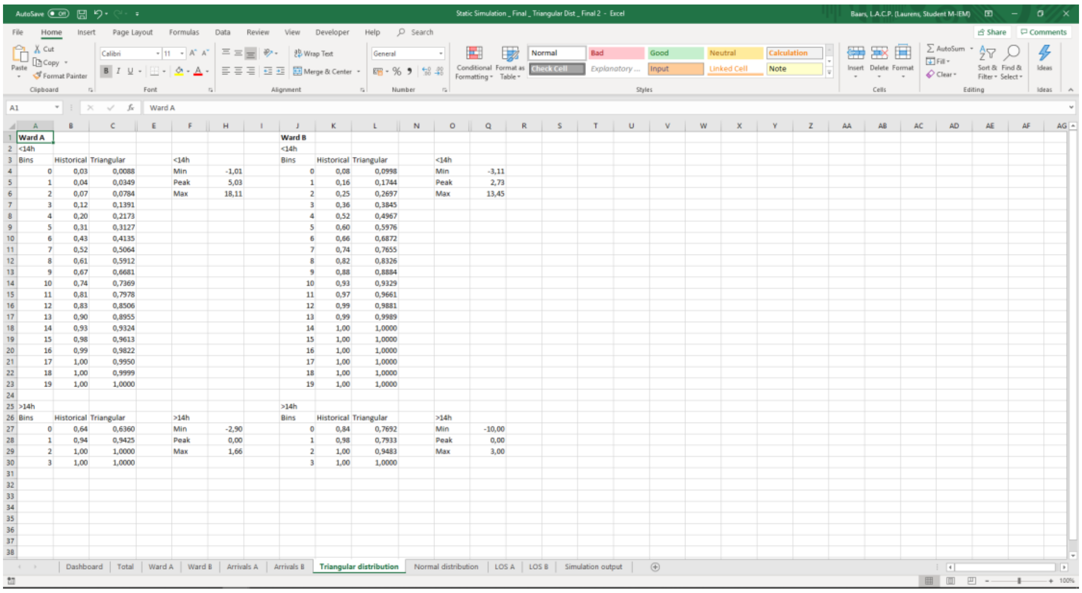


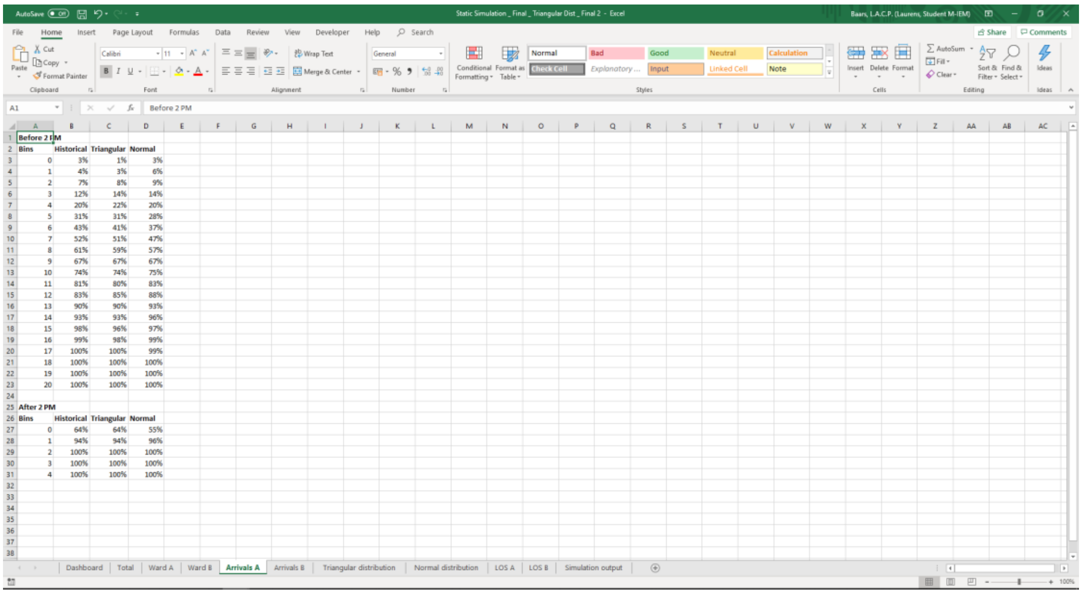
Figure 3. Generated LOSs of URO patients in days



Bins	Historical	Triangular	<14h		Bins	Historical	Triangular	<14h	
Ward A					Ward B				
<14h					<14h				
0	0,03	0,0088	Min	-1,01	0	0,08	0,0998	Min	-3,11
1	0,04	0,0349	Peak	5,03	1	0,16	0,1744	Peak	2,73
2	0,07	0,0784	Max	18,11	2	0,25	0,2697	Max	13,45
3	0,12	0,1391			3	0,36	0,3845		
4	0,20	0,2173			4	0,52	0,4967		
5	0,31	0,3127			5	0,60	0,5976		
6	0,43	0,4135			6	0,66	0,6872		
7	0,52	0,5064			7	0,74	0,7655		
8	0,61	0,5912			8	0,82	0,8326		
9	0,67	0,6681			9	0,88	0,8884		
10	0,74	0,7369			10	0,93	0,9329		
11	0,81	0,7978			11	0,97	0,9661		
12	0,83	0,8506			12	0,99	0,9881		
13	0,90	0,8955			13	0,99	0,9989		
14	0,93	0,9324			14	1,00	1,0000		
15	0,98	0,9613			15	1,00	1,0000		
16	0,99	0,9822			16	1,00	1,0000		
17	1,00	0,9950			17	1,00	1,0000		
18	1,00	0,9999			18	1,00	1,0000		
19	1,00	1,0000			19	1,00	1,0000		

Bins	Historical	Triangular	>14h		Bins	Historical	Triangular	>14h	
0	0,64	0,6360	Min	-2,90	0	0,84	0,7692	Min	-10,00
1	0,94	0,9425	Peak	0,00	1	0,98	0,7933	Peak	0,00
2	1,00	1,0000	Max	1,66	2	1,00	0,9483	Max	3,00
3	1,00	1,0000			3	1,00	1,0000		

Figure 4. Calculation of cumulative arrival rate distribution (Triangular distribution)



Before 2 PM

Bins	Historical	Triangular	Normal
0	3%	1%	3%
1	4%	3%	6%
2	7%	8%	9%
3	12%	14%	14%
4	20%	22%	20%
5	31%	31%	28%
6	43%	41%	37%
7	52%	51%	47%
8	61%	59%	57%
9	67%	67%	67%
10	74%	74%	75%
11	81%	80%	83%
12	83%	85%	88%
13	90%	90%	93%
14	93%	93%	96%
15	98%	96%	97%
16	99%	98%	99%
17	100%	100%	99%
18	100%	100%	100%
19	100%	100%	100%
20	100%	100%	100%

After 2 PM

Bins	Historical	Triangular	Normal
0	64%	64%	55%
1	94%	94%	96%
2	100%	100%	100%
3	100%	100%	100%
4	100%	100%	100%

Figure 5. Arrival rate distributions for URO patients

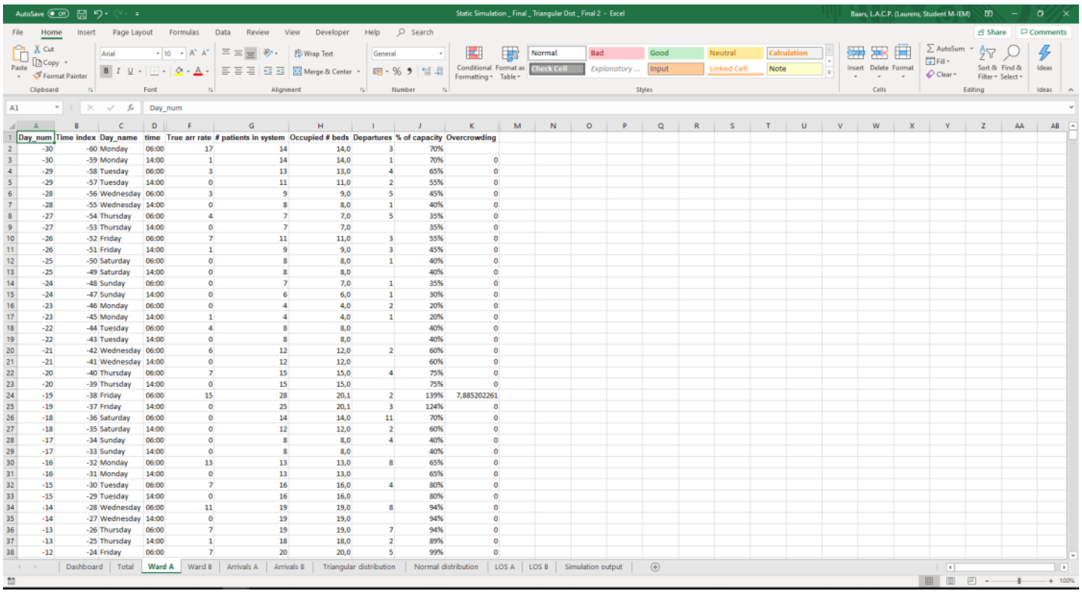


Figure 6. Simulated URO patients

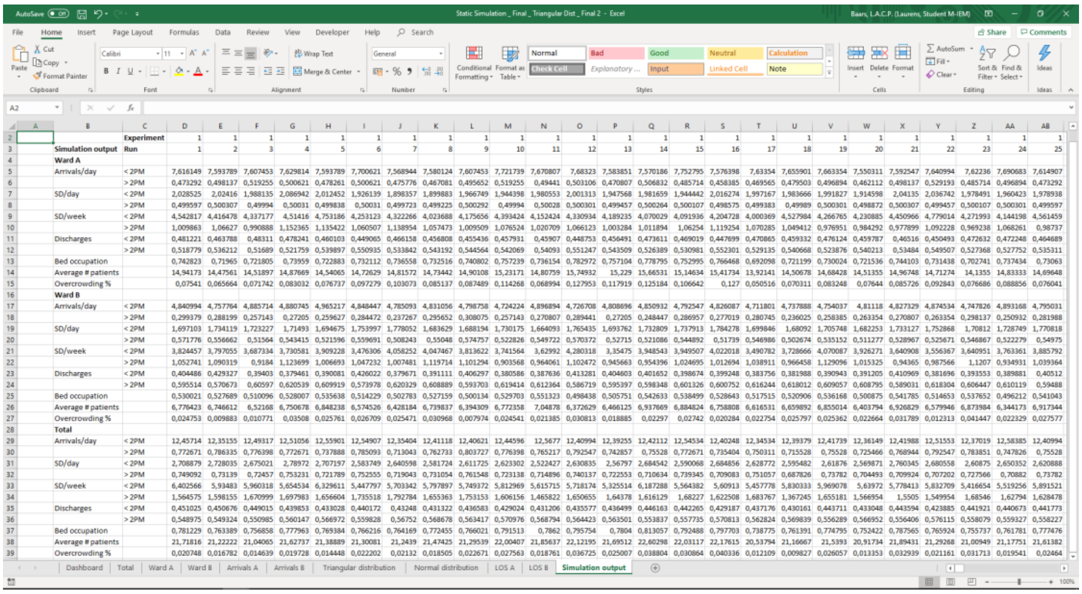


Figure 7. Simulation output per run

Appendix VI

Variable	Runs				
	10	50	100	150	200
Arrivals URO <2PM	2.7%	1.1%	0.7%	0.6%	0.5%
Arrivals GYN <2PM	2.7%	1.5%	1.0%	0.8%	0.7%
Arrivals URO >2PM	4.5%	2.6%	2.0%	1.6%	1.4%
Arrivals GYN >2PM	10.9%	4.1%	2.6%	2.2%	1.9%
St.dev. URO <2PM	4.3%	1.4%	0.9%	0.7%	0.6%
St.dev. GYN <2PM	4.1%	1.3%	0.9%	0.7%	0.6%
St.dev. URO >2PM	2.8%	1.4%	0.9%	0.8%	0.7%
St.dev. GYN >2PM	5.3%	1.8%	1.1%	1.0%	0.9%
Avg. # patients URO	5.4%	2.0%	1.5%	1.3%	1.1%
Avg. # patients GYN	5.5%	3.4%	2.4%	1.9%	1.7%
Avg. # patients Ward	3.2%	1.7%	1.2%	1.1%	0.9%
Overcrowding URO	22.0%	7.3%	5.7%	4.7%	4.3%
Overcrowding GYN	22.2%	14.5%	10.3%	8.0%	6.7%
Overcrowding Ward	32.3%	10.5%	7.9%	7.0%	6.1%
Discharges URO <2PM	4.2%	1.3%	1.0%	0.8%	0.7%
Discharges GYN <2PM	4.0%	1.6%	1.2%	1.0%	0.8%
Discharges Ward <2PM	3.1%	1.0%	0.8%	0.6%	0.5%

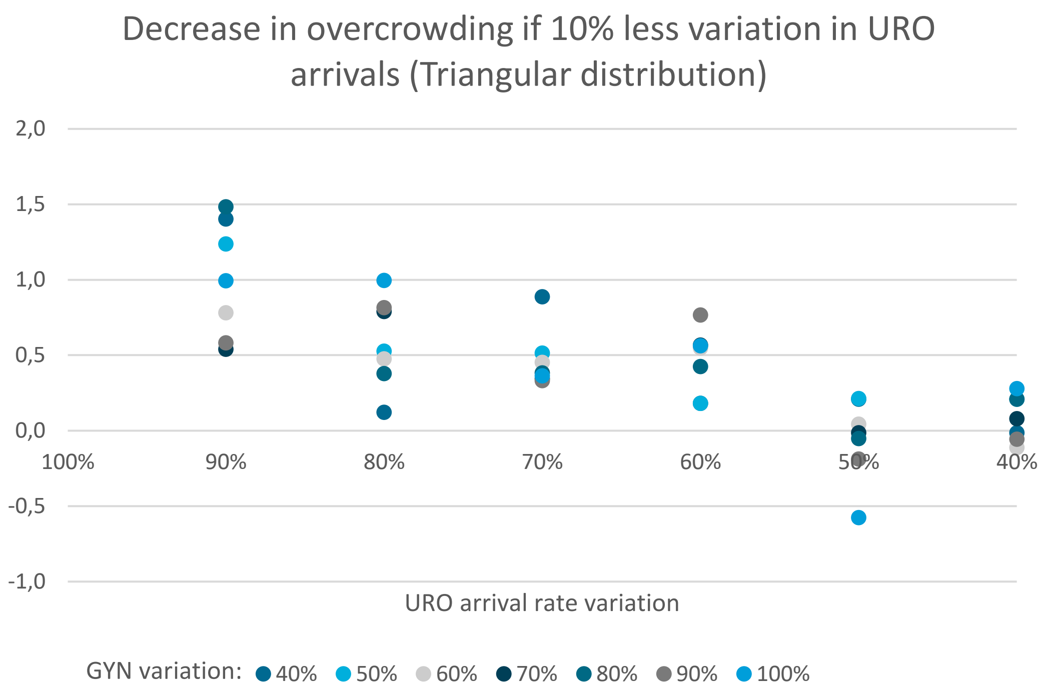
Table 15. 95% CI spread as percentage of the mean value, using the Triangular distribution for the arrival rate

Variable	Runs				
	10	50	100	150	200
Arrivals URO <2PM	1.8%	0.8%	0.6%	0.6%	0.5%
Arrivals GYN <2PM	3.0%	1.1%	0.8%	0.7%	0.6%
Arrivals URO >2PM	5.7%	2.3%	1.6%	1.2%	1.1%
Arrivals GYN >2PM	2.4%	1.2%	0.8%	0.8%	0.6%
St.dev. URO <2PM	7.7%	3.0%	2.1%	1.7%	1.5%
St.dev. GYN <2PM	3.6%	1.8%	1.2%	1.0%	0.9%
St.dev. URO >2PM	7.0%	3.4%	2.1%	1.8%	1.6%
St.dev. GYN >2PM	2.8%	1.0%	0.8%	0.7%	0.6%
Avg. # patients URO	5.5%	2.1%	1.4%	1.2%	1.2%
Avg. # patients GYN	9.6%	2.7%	2.0%	1.7%	1.4%
Avg. # patients Ward	5.2%	1.6%	1.1%	1.0%	0.9%
Overcrowding URO	20.2%	8.5%	6.0%	5.3%	4.9%
Overcrowding GYN	43.1%	11.7%	9.0%	7.8%	6.4%
Overcrowding Ward	25.9%	11.0%	8.5%	7.3%	7.0%
Discharges URO <2PM	4.3%	1.4%	1.0%	0.8%	0.7%
Discharges GYN <2PM	3.6%	1.8%	1.2%	1.0%	0.9%
Discharges Ward <2PM	7.2%	3.0%	2.3%	1.9%	1.5%

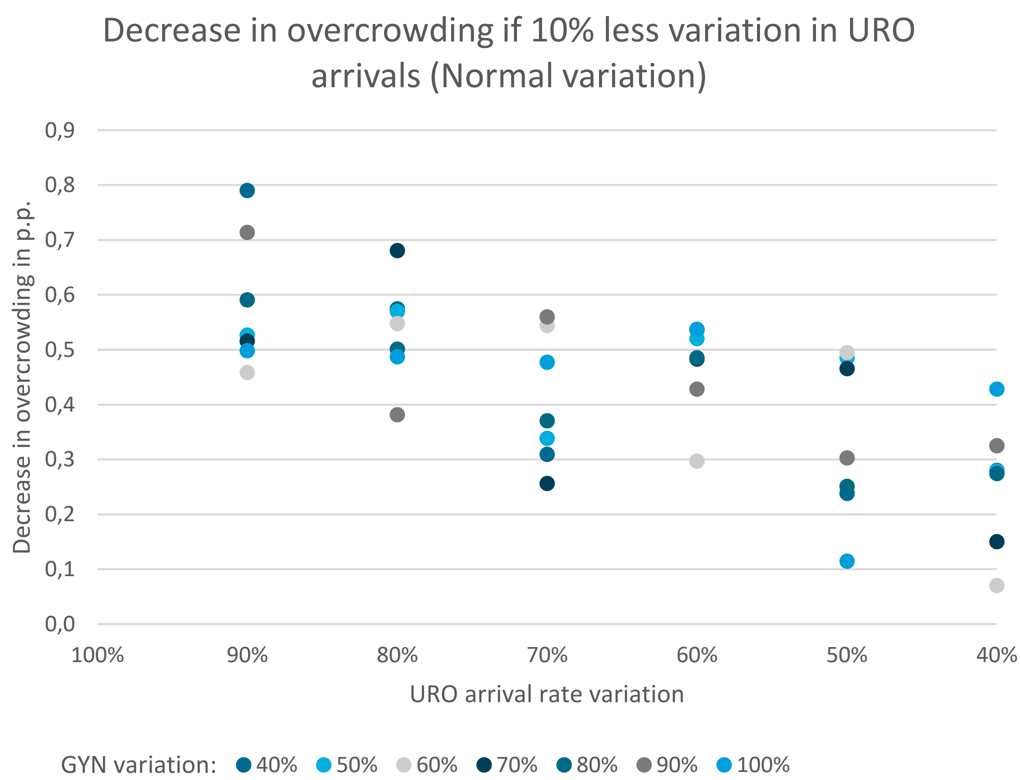
Table 16. 95% CI spread as percentage of the mean value, using the Normal distribution for the arrival rate

Note that, because of the randomness in the simulation, these numbers do not 100% accurately present the spread in the CI, i.e. if another 500 would be performed, the spread would likely be somewhat different. It gives however an indication of the decrease in the spread by adding more runs. These numbers also give an estimate of the actual spread and show that the two distributions give similar results for the spread.

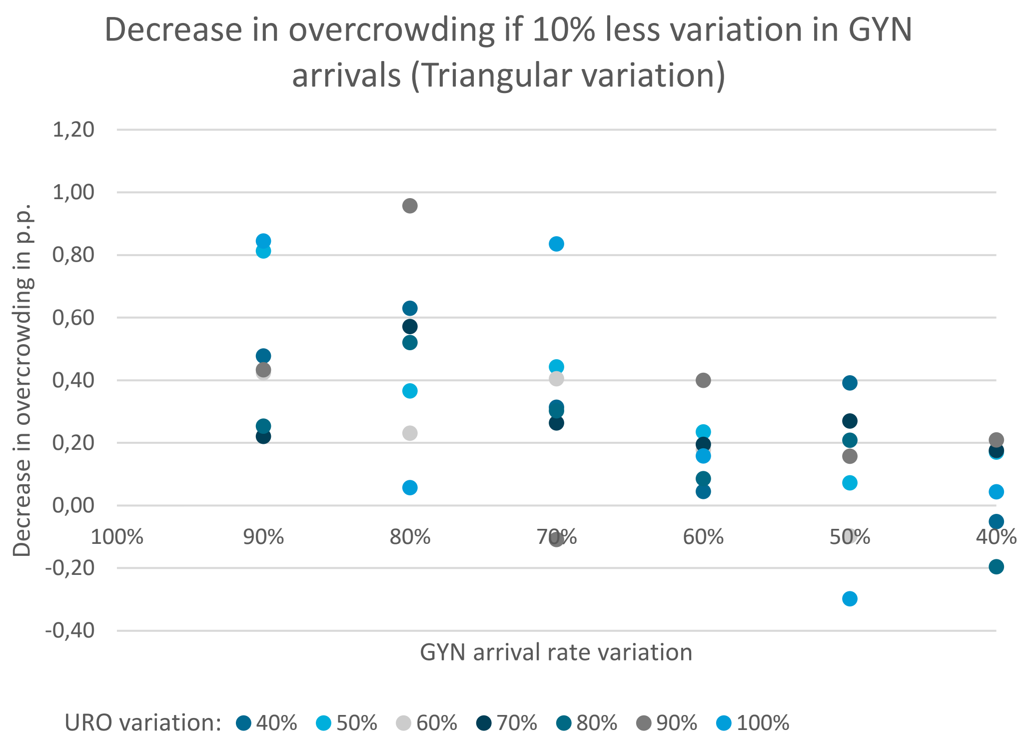
Appendix VII



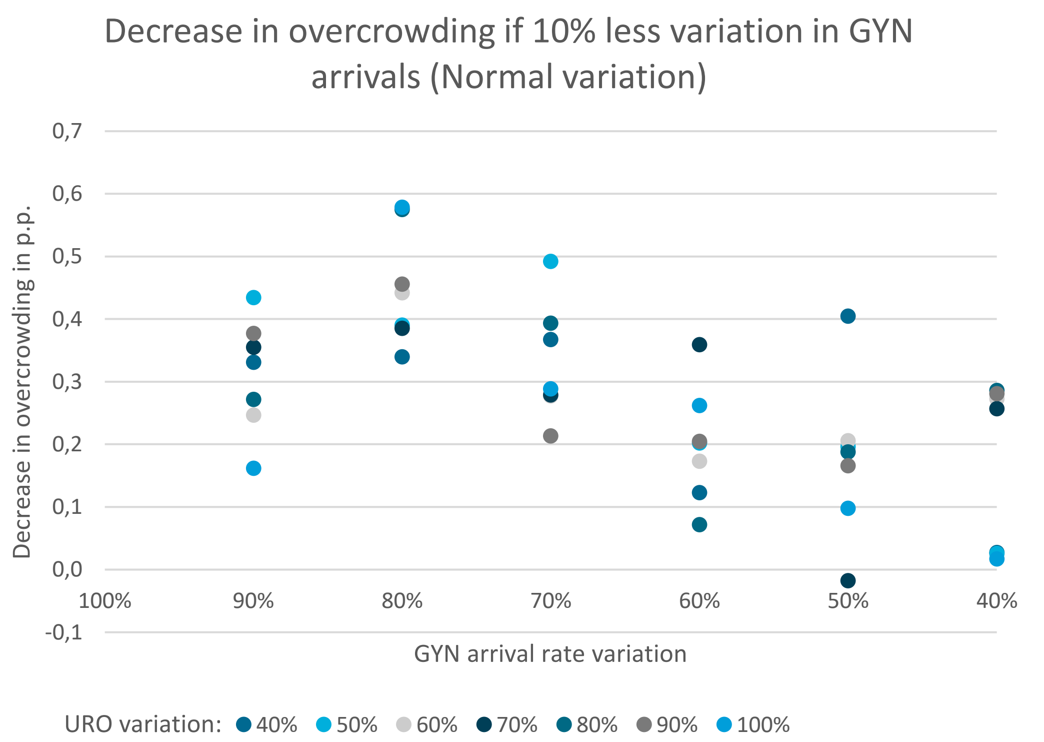
Graph 19. Decrease in overcrowding of decreasing the variation of URO arrivals by 10% (Triangular arrival rate distribution)



Graph 20. Decrease in overcrowding of decreasing the variation of URO arrivals by 10% (Normal arrival rate distribution)

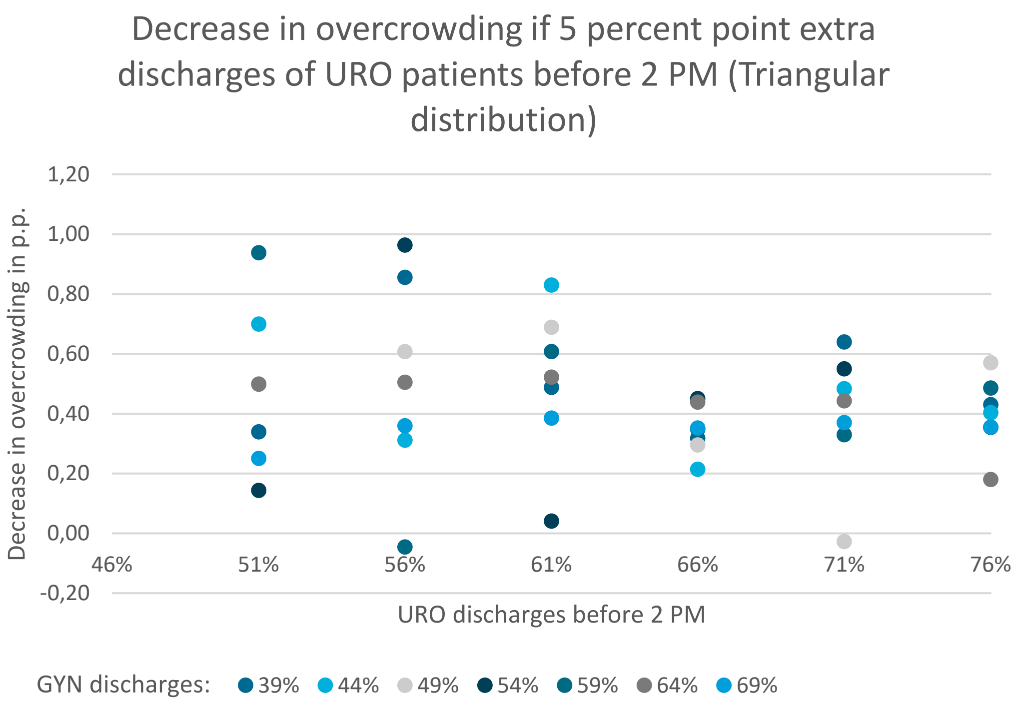


Graph 21. Decrease in overcrowding of decreasing the variation of GYN arrivals by 10% (Triangular arrival rate distribution)

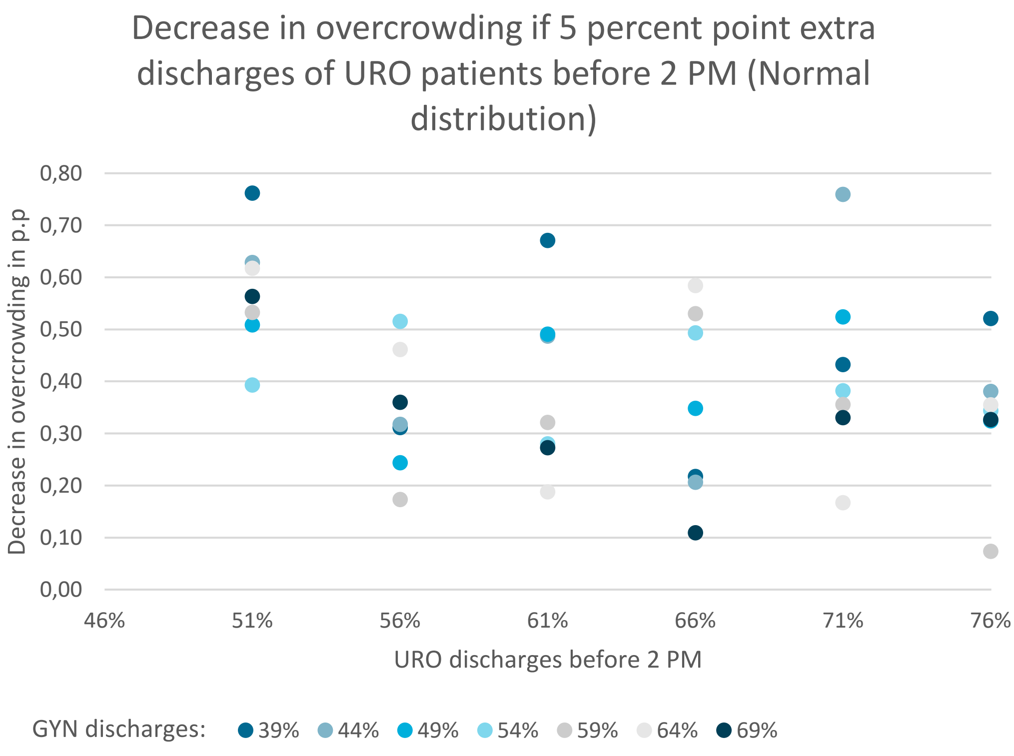


Graph 22. Decrease in overcrowding of decreasing the variation of GYN arrivals by 10% (Normal arrival rate distribution)

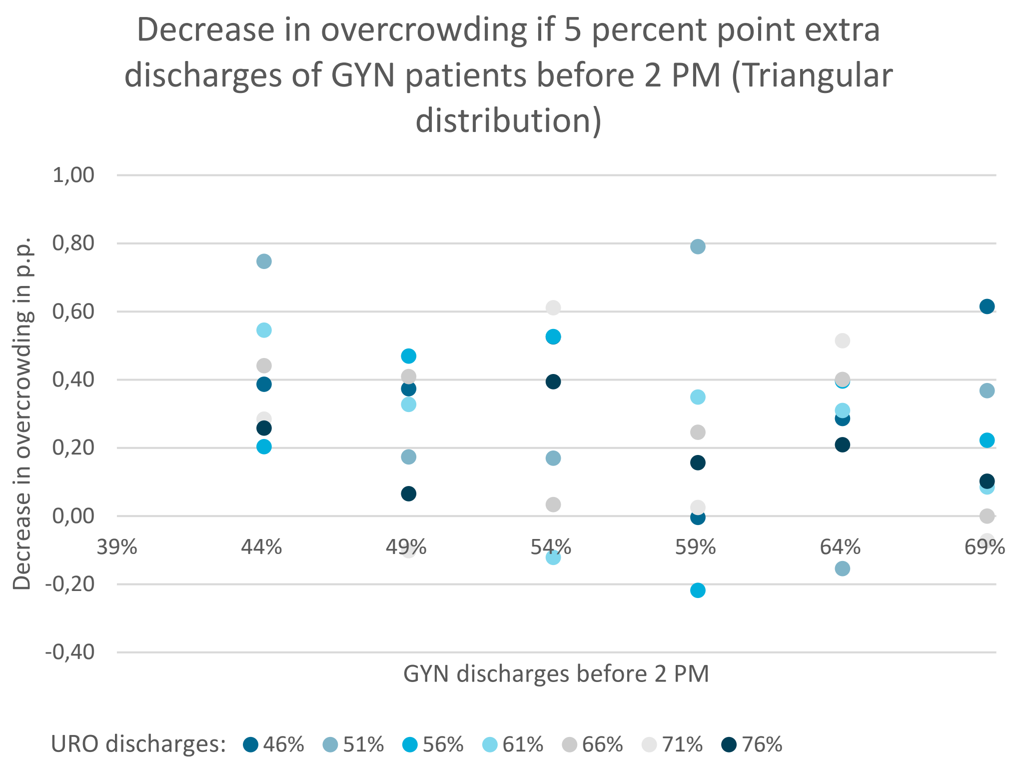
Appendix VIII



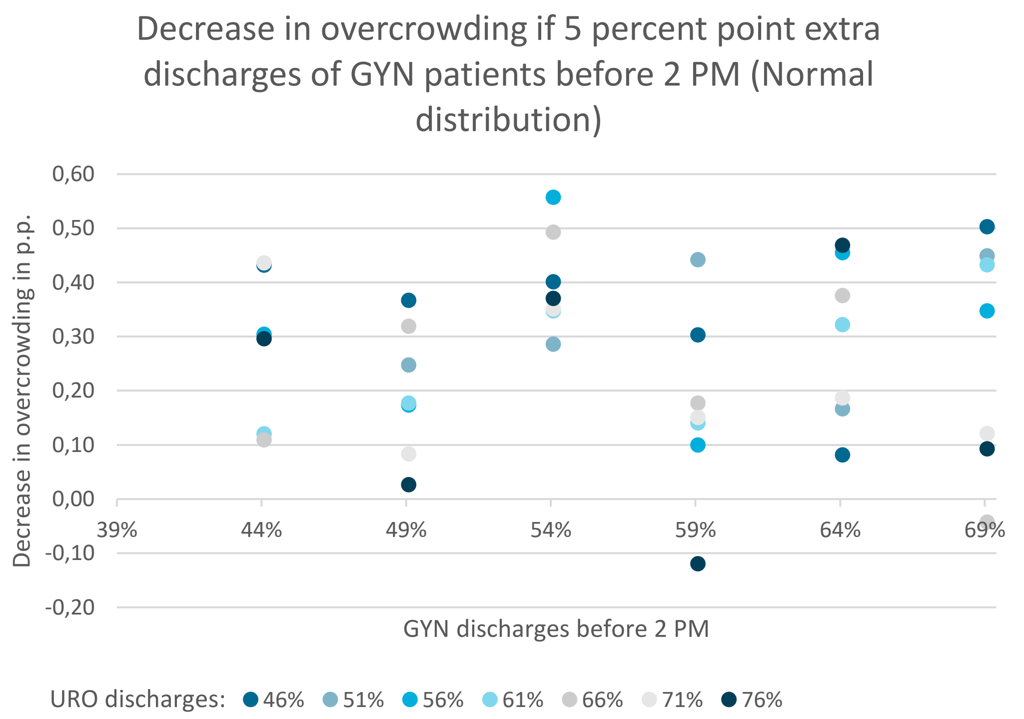
Graph 23. Decrease in overcrowding when 5 percentage point more URO patients are discharged before 2 PM (Triangular arrival rate distribution)



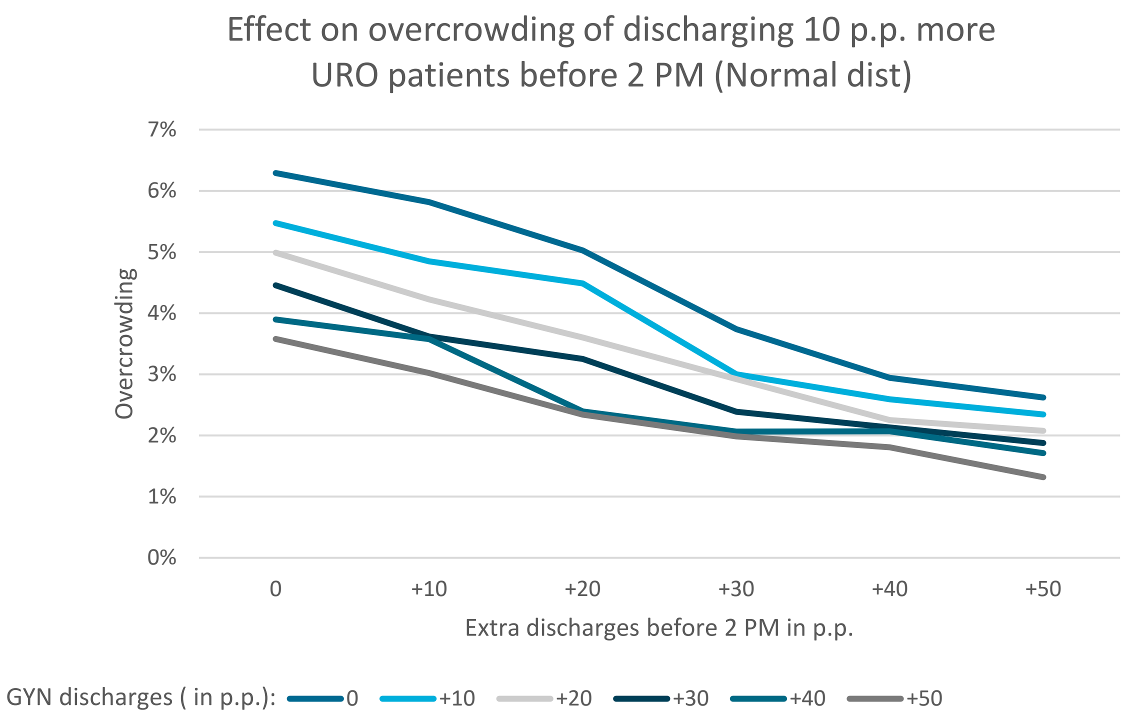
Graph 24. Decrease in overcrowding when 5 percentage point more URO patients are discharged before 2 PM (Normal arrival rate distribution)



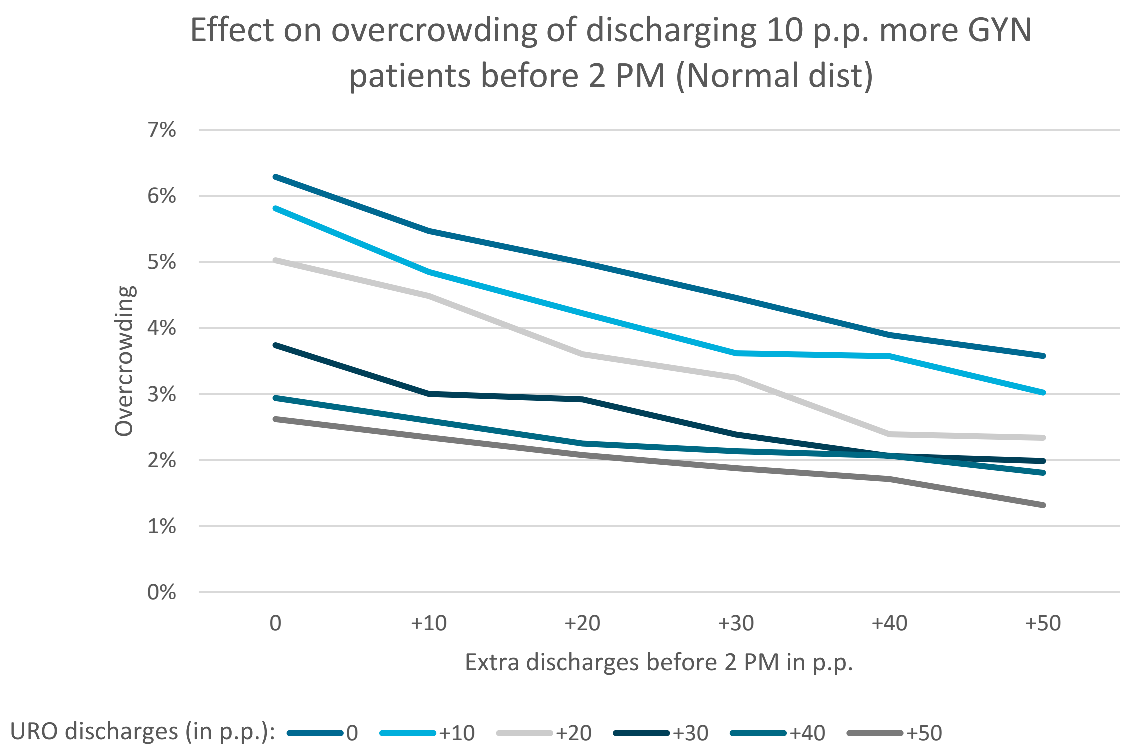
Graph 25. Decrease in overcrowding when 5 percentage point more GYN patients are discharged before 2 PM (Triangular arrival rate distribution)



Graph 26. Decrease in overcrowding when 5 percentage point more GYN patients are discharged before 2 PM (Normal arrival rate distribution)



Graph 27. Effect on overcrowding of discharging more URO patients before 2 PM given the extra discharges before 2 PM of GYN patients (t = 200 runs)



Graph 28. Effect on overcrowding of discharging more GYN patients before 2 PM given the extra discharges before 2 PM of URO patients (t = 200 runs)

Appendix IX

The Erlang-B formula is given by

$$P_b = B(E,m) = \frac{\frac{E^m}{m!}}{\sum_{i=0}^{m} \frac{E^i}{i!}} \quad ,$$

where

P_b is the probability of blocking, i.e. off-servicing a patient to another ward,

m is the number of servers, i.e. beds at the ward, and

$E = \lambda h$ is the offered traffic in *Erlang*, i.e. the number of arrivals per day.

It is often calculated recursively as

$$B(E,0) = 1$$
$$B(E,j) = \frac{EB(E,j-1)}{EB(E,j-1)+j} \quad \forall j = 1,2,\ldots,m$$