UNIVERSITY OF TWENTE.

UMC Utrecht
Wilhelmina Kinderziekenhuis

Cooperative solutions for challenges in neonatal care: A case study of the region of Utrecht

Public version

MSc Thesis
Industrial Engineering and Management

M.A.J. Rikkert
Utrecht, December 8, 2022

Author
M.A.J. Rikkert

Educational Institution
University of Twente
Faculty of Behavioural Management and Social Sciences
Department of Industrial Engineering and Business Information Systems
Centre for Healthcare Operations Improvement and Research

Educational Program
MSc Industrial Engineering and Management
Specialisation: Production and Logistics Management
Research Orientation: Operations Management in Healthcare

Supervisors
dr. ir. A.G. Maan-Leeftink
University of Twente
Centre for Healthcare Operations Improvement and Research

dr. D. Guericke, MSc.
University of Twente
Centre for Healthcare Operations Improvement and Research

B. van den Berg, MSc.
Directie Strategie en Beleid, Integraal Capaciteitsmanagement
University Medical Center Utrecht

dr. K.A. de Bijl-Marcus
University Medical Center Utrecht
Wilhelmina Children's Hospital, Department of Neonatology

Abstract

This research aims to demonstrate the effect of cooperative solutions to tackle neonatal care capacity challenges. The research motivated by the observation that neonatal wards are critically overloaded, resulting in worse performance outcomes for both patient and healthcare employees, and there is a clear need to seek for solutions outside the hospitals themselves. Two earlier articles that proposed analytical models for resource planning and optimisation in intensive care units, based on overflow models in telecommunication systems, form the inspiration of this research.

In neonatal care, multiple patient types are involved that differ in their severity of care. We provide a mathematical model to enhance capacity planning that distinguishes three different patient types, that we refer to as the three-patient-type model (TPTM). So far, the TPTM was applied in two policies in intensive care wards: a virtual MC (VMC) policy by Litvak et al. (2008) and a threshold policy by Chan et al. (2018). We made adaptions to the TPTM in order to apply the model and its policies to the context of neonatal care. The goal of the TPTM is to reduce the number of patients who are refused admission while maintaining adequate levels for patients whom cannot be refused.

This research aims to demonstrate the effect of cooperative solutions to tackle neonatal care capacity challenges. Two earlier articles that proposed analytical models for resource planning and optimisation in intensive care units, based on overflow models in telecommunication systems, form the inspiration of this research. We employ a model that distinguishes three patient types, that we refer to as the three-patient-type model (TPTM). So far, the TPTM was applied in two policies in intensive care wards: a virtual MC (VMC) policy by Litvak et al. (2008) and a threshold policy by Chan et al. (2018). We made adaptions to the TPTM in order to apply the model and its policies to the context of neonatal care. We propose a numerical approach for the VMC policy, in which we model the VMC as a separate location which differs from the analytical approach as proposed by Litvak et al. (2008). Our model is a Continuous-Time Markov Chain (CTMC) that we solve numerically. Due to the solution space size when more than two hospitals are involved, the CTMC approach becomes inaccurate. Therefore we also developed a Discrete Event Simulation (DES). In a case study from the city of Utrecht concerning four hospital the results showed that for two hospitals using the CTMC approach results in a joint decline of 132 lesser refusals per year at the costs of 78 more overbed-days per year for the two hospitals involved, while maintaining proper occupancy levels of 85%. Extending to four hospitals with the DES approach, we observe a joint decline of 49 beds at the costs of 24 more overbed-days.

A threshold-based patient referral policy, inspired by the threshold policy thatChan et al. (2018) developed for a network of ICUs, has shown that we can achieve a higher patient acceptance level and lower fraction of time that the hospitals will spend in overbeds. In this policy we propose a solution method based on CTCM. We also propose a greedy optimisation method to optimise threshold reservation settings. In the case study we demonstrate that an our greedy optimisation technique results in a collective decline of 45 refusals per year and 35 overbed-days per year, for 4 neonatal care locations involved. We conclude that sharing beds and distributing patients over the hospitals in the network according to the threshold reservation policy, is beneficial both for the patient as well as from a nursing perspective, especially keeping in mind the related trends in healthcare regarding a shift towards integrated care and the development of applications that facilitate better communication and capacity insights between hospitals. Furthermore, our simulation results show that the introduction of the VMC lowers the probability of refusing but is at the cost of having overbeds. Therefore, we do not recommend to introduce a virtual MC policy. The threshold policy exhibits good practical applicability, but it is necessary to look at ways to enhance the model in order to make the model more robust. A sensitivity analysis on the input parameters reveals that the threshold policy outperforms the current situation.

Acknowledgements

My research "cooperative solutions for neonatal care capacity problems" is the final step towards obtaining my masters degree in Industrial Engineering and management. I wish you an enjoyable read.

My gratitude goes towards my colleagues from the capacity management of the UMC Utrecht who inspired me during the past nine months. First of all, to my supervisor Bart, who gave a broad perspective on capacity management and from whom I developed lots of professional and personal skills. I learned a lot on what it takes to motivate and inspire people to tackle problems in a different way. I experienced that it is not solely about the mathematics and the computer programs, but that it requires motivation and action to make our research to work into practice. To this, I am also grateful for Arjan who took me away from my computer and let me participate an active role in operational processes the UMC Utrecht contains and to Erik and Irene for guiding me through this process. I am also very grateful to Angelique, my colleague that guided as a personal mentor through this entire process.

Moreover, I would like to thank Gréanne and Daniela for the many discussions we had and the structured feedback they gave which supported me to deliver this work in its current state. Gréanne learned my to critically evaluate my own work and inspired me a lot. Furthermore, I am grateful to Erwin who inspired me to take a look at the operations research in healthcare. This offered a very different perspective in health care which I believed that would not even exists, seven years ago at the start of my bachelor in Technical Medicine.

Last but not least, I am grateful to my friends and family for patiently listening and inspiring me throughout the entire process.

Marieke Rikkert
Borne, December 2022

Contents

Glossary

CTMC Continuous Time Markov Chain.

DES Distrete event Simulation.

ED Exponential Decomposition.

HC High care.

MC Medium care.

NICU Neonatal Intensive Care Unit.

TPTM Three Patient Type Model.

UMCU University Medical Center of Utrecht.

VMC Virtual MC.

Chapter 1

Introduction

Recently, the UMC UTrecht, one of the biggest hospitals in the Netherlands, announced that "Birth care is under pressure" (UMC Utrecht, 2021). Pregnant women, as well as maternity care workers, feel burdened by the existing pressure in the maternity care chain. The shortage of nurses in the Netherlands expects to grow, and an increase in demand is expected; thus, the healthcare capacity will not be adequate to address the future demand (Ministerie van Volksgezondheid en Welzijn en Sport, 2022). There is a clear need to critically enhance neonatal care capacity planning. In this chapter, we present an introduction to the research. The research is motivated by the capacity management of a neonatal department's Medium Care (MC) unit from the Wilhelmina Kinderziekenhuis (WKZ). More context to the organisation of neonatal care in the WKZ is provided in Section 1.1. The complex hospital chain logistics intensify the neonatal capacity problem, particularly in hospitals that know both second and third line responsibilities. The management of this department is complex, as multiple types of arrival streams are involved. The capacity problem is not only present in the WKZ, but it also affects several hospitals in the Utrecht region. Section 1.2 presents information on other hospitals that we include in this research.

1.1 Context introduction: Neonatal care in the WKZ

The WKZ maternity centre covers all aspects of pregnancy for both the mother and the child. The neonatal department consists of the Neonatal Intensive Care Unit (NICU), the post Intensive Care or High Care unit (Post-IC or HC) and the Medium Care unit (MC). Babies born too early or requiring medical attention after birth admits to one of these wards. The neonates' age, weight, respiratory support and other conditions form the overall severity of care that determines the unit the neonate treats. The WKZ is part of the University Medical Center of Utrecht (UMCU) The organizational structure is presented in Figure 1.1.

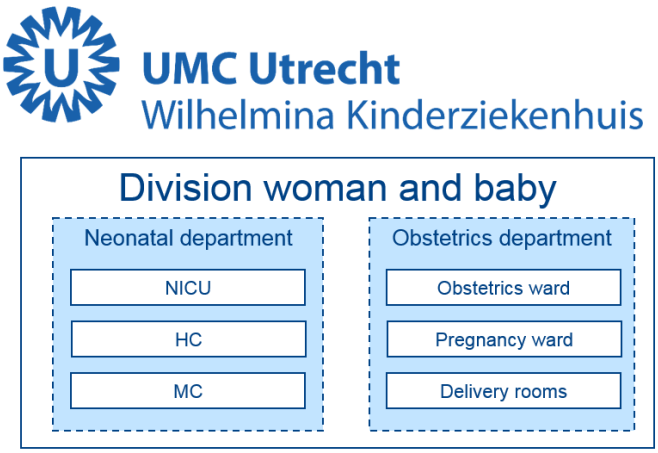


Figure 1.1: Organisational structure of the division woman and baby of the WKZ

Regarding terminology, we refer to the *neonatal department* as the department consisting of the NICU, HC and MC unit. Each individual unit is also referred to as a *neonatal ward*. For example, the *obstetrics department* consists of the delivery rooms, the obstetrics ward, and the pregnancy ward. To the extent of bed availability, we distinguish three concepts: an *available*, *operational*, and an *over*-bed. An available bed is the physical presence of the bed, including the resources to operate that bed, whereas a bed is called operational in case a patient lays on that bed. An overbed is an extra bed which is initiated when full capacity is reached. When we operate more beds than this limit, one speaks of overbeds. Overbeds are unattractive for three reasons. First, from the patient perspective, when more beds are presented than the nurses can handle, there is a risk of lower quality of care (Ogboenyiya et al., 2020). From the medical staff point of view, overbeds result in a high workload. Having overbeds is undesirable from a management point of view since it is challenging to obtain adequate funding to operate if the department produces more than what is specified in the production agreements.

1.1.1 Medium Care

The NICU has a long-standing capacity issue, and solutions are being explored for that department. The hospital undervalued other departments in favor of NICU capacity coordinating optimization. As a result, not only the NICU but also other wards experience capacity problems. One of the downstream departments, the MC, experience the problems as well which is mostly due to the MCs' many functionalities. Those are the subject of discussion in this section. The function of the MC is threefold. First, the MC unit is a scale-down option for patients discharged from the Neonatal Intensive Care Unit (NICU) or High Care unit (HC). Second, WKZ is obligated to fulfil second-line care to meet regional demand. Third, the MC functions as an admission possibility for babies having mothers in third-line care and babies with a high potential for intensive care admission. As a result, it is difficult for the management to balance capacity concerning the unit's functions. This is reflected in the observation that many different patient groups arrive at the MC as reflected in Figure **??**. The following section describes the logistic processes.

1.1.2 Logistical processes in the medium care unit of the WKZ

In this Section we address the situation in the Wilhelmina Children's Hospital (WKZ). Data for this case study is derived from registrations in the health information system (HIS) over the year 2020 and 2021 with a focus on neonatal patients, born in 2019, 2020 ad 2021. The data is obtained from the

years 2020 and 2021, because in these years, the management of the neonatal ward started keeping track of the number of refusals. In the model we use actual data from the neonatal department of the WKZ and Diakonessenhuis and and estimated data from two peripheral hospitals in the Utrecht region, the Antonius hospital and the Meander hospital. Below, we present the patient flow of the neonatal department of the WKZ. The total flow of patients in and out of this department is presented in Figure **??**. Note that there is a small readmission rate, both from MC to HC as well as from MC and HC to NICU, which is indicated in dark red. The total flow of patients in- and out the neonatal department is presented in Table **??**.

1.2 Neonatal care in the Utrecht region

Often, solutions are opted inside the own healthcare organisations. The capacity problem in neonatal care is not limited to one hospital; many hospitals in the Netherlands experience similar events (Ministerie van Volksgezondheid en Welzijn en Sport, 2022). But, there is a saying that "A little flexibility goes a long way", meaning that introducing some flexibility opens opportunities beyond their own hospital. Figure 1.2a presents the location of the four hospitals included in this research.

Figure 1.2: Hospitals included in this research - three hospitals are located in the city of Utrecht and one hospital (the Meander hospital) is outside the city, but still located in the province of Utrecht

(a) Map of the Netherlands, highlighted is the province of Utrecht

(b) Province of Utrecht with the included locations (1=WKZ, 2=Diakonessenhuis, 3=Meander, 4=Antonius)

1.2.1 Previous research in neonatal capacity locations in Utrecht

The neonatal care in the region of Utrecht has been subject to limiting resources for the past years. Not only an increasing demand, but the addition of decreasing resources, leads to a shortage of beds that is noticed in daily practice. The city of Utrecht is particularly well-liked among young adults. As a consequence, it is anticipated that the need for maternity care in the area will grow in the forthcoming years.

Simply increasing the resources would solve the problem at once. However, in practice this is hard to find. Therefore, we have to look at managing the current resources. Previous research on this topic has shown that personnel planning based on acuity measures may lead to better results in terms of overloaded shifts (Hoek, 2015), as well as pooling NICU and HC beds in terms of occupancy (Weernink, 2018). The solutions proposed for the locations themselves are often inadequate to meet the region's growing demands, since the solutions only optimize hospital logistics with a given demand from the locations themselves (Hoek, 2015; Weernink, 2018) and by constraining the amount of incoming patients, with the consequence that other locations are burdened with extra loads. Therefore, a regional solution should be found.

Other previous research showed that pooling resources will lead to a lower amount of transports, which will possibly result in better health care outcomes for preterm infants (Morris, 2021). However, practical barriers are faced when bringing this research in practice. For example, the pooling of nurses goes along with many regulations, let aside the personal preference of the team the nurse is working for. Even though there is evidence that collaboration of hospitals may lead to improvements, (Cattani & Schmidt, 2005; Van Dijk & van der Sluis, 2009), many hospitals lack a strategy of doing so. Without a policy that handles concerns of capacity coordination, most hospitals handle the transfer of overflow patients on an ad-hoc basis.

The problem cluster in Appendix A presents how the various initial problems result in the core problem: a lack of control regarding capacity among the patient groups involved.

1.3 Contribution of the research and research gap

This research concentrates on strategies for cooperation among MCs in the region of Utrecht. We provide a mathematical model to enhance capacity planning. Due to the numerous duties the MC has to fulfil, we discriminate several patient types that the MC employs. In this research, we propose a model that includes three patient types. We refer to this model as the "three-patient-type" model. To this, we extend the three-patient-type model (TPTM) from Litvak et al. (2008) and Chan et al. (2018) which apply the TPTM in intensive care. The goal of the TPTM is to reduce the number of patients who are refused admission while maintaining adequate levels for patients who cannot be refused care in an emergency. The patient types include patients who can be rejected and admitted elsewhere, those who can be refused and do not come back, and finally those who must be admitted. If this last patient type arrives to a full ward, it is admitted by creating an overbed.

We apply two policies to the three-patient-type model: the *virtual ICU* policy, proposed by Litvak et al. (2008), and the *threshold* policy, introduced by Chan et al. (2018). The *virtual ICU* policy, in which ICUs in a region jointly reserve beds for the admission of a specific patient type, may lead to the wastage of valuable ICU resources. A patient may be refused at a given hospital, even if a bed is empty at the same hospital due to that bed being reserved for the virtual ICU. This motivated Chan et al. (2018) to design a model in which no bed is explicitly reserved for a particular patient type. Instead, each patient type admits until the total unit's capacity exceeds a *threshold* value. By tuning this threshold value, they maximise the sharing of beds while protecting resources for other patient types. For both models, we demonstrate the effect on the blocking probability and number of overbeds in the network when hospitals in a network collaborate.

The research gap is twofold. First, we apply the models in the context of neonatal care. The origin of patient types and the actions that correspond to the type differs from intensive care patients as considered in the literature so far. Second, we apply different solution methods than applied to the TPTM so far. We propose a numerical solution method for the virtual MC policy based on Copntinuous-Time Markov Chains (CTMC) which distinghuises from the solution approach from Litvak et al., 2008 because they use a analytical approach in which they model the VMC as an

Equivalent Random unit (ER). They calculate the expected load to this unit which is used to calculate the blocking probability to the VMC. We however propose a CTMC that considers the VMC as an extra location and solve the CTMC for all locations, including the VMC, simultaneously. The load of the VMC in our CTMC is of similar magnitude as the load of type 1 patients that the locations receive; we do not make adjustments to this load.

Second, we propose an greedy optimization for the threshold policy to find the optimal setting of thresholds that yields the most promising results. This differs from the optimisation process of the threshold policy of Chan et al. (2018). Namely, they optimize the set of thresholds by setting constraints to the blocking- and overbed probability and punishing the amount to which the constraint is exceeded. We do not punish outcomes after a fixed value is exceeded but rather select the option that yields the best outcomes for the collective blocking- and overbed proportions. Chan et al. (2018) selects optimal reservation thresholds that yields profits for every location. We however select the best thresholds for that yields the most promising results for all locations together and check for every location if the individual locations are not harmed by the proposed setting.

The overall managerial goal is,

To establish strategies for facilitating and improving capacity sharing amongst neonatal departments at hospitals in the Utrecht area.

Our contributions are the following:

1. **To learn from literature about the most recent methods for tackling capacity issues in neonatal, emergency, and intensive care, that include solutions in collaboration**
   We start by finding ways to overcome capacity problems in literature. The focus is on stochastic arrivals for neonatal, emergency, or intensive care units. We use the Scopus database and filter the relevant articles by reviewing the title, keywords, and abstract based on predefined search terms. The result of this literature review is presented in Chapter 2. We also strive to find performance outcome measures that indicate the quality of the found methods. Section 2.5 provides this.

2. **To design a model for capacity sharing in neonatal departments that include different patient types**
   We formulate a general model that we refer to as the 'three-patient-type' model. This model allows us to vary the actions of patient groups when this patient finds a full ward. Given these actions, we can decide on capacity allocations among these groups. To this model, we apply two initiatives for regional collaboration. First, we extend on the *virtual ICU* policy, second on the *threshold* policy. The three-patient-type model of the previous step is extended by introducing the concept of a joint capacity reservation with peripheral hospitals. Chapter 3 formulates both policies in a model. Litvak et al. (2008) and Chan et al. (2018) illustrated a reduction of respectively the blocking probability, deferral probability, and a number of overbeds for the ICU in case of sharing capacity with several hospitals, which forms the motivation to apply these methods to the neonatal care context.

3. **To apply the virtual MC policy on the three-patient-type model, by adapting the policy from Litvak et al. (2008)**
   The first policy to apply is the virtual ICU policy. Here, several hospitals jointly reserve a small number of beds, which we refer to as the *virtual MC*. Litvak et al. (2008) used an analytical approach to determine the load and variance of this MC; we formulate a queuing model, and use Discrete Event simulation to find the blocking probability and number of overbeds in this system Litvak et al. (2008) showed that cooperation in this form reduced a lower blocking probability and a number of overbeds.

4. **To apply the threshold policy on the three-patient-type model, by adapting the policy from Chan et al. (2018)**
   The risk of under-utilisation of resources, including the possibility to refuse a patient even though a free bed is available, inspired Chan et al. (2018) to develop a model that does not set beds aside for specific patient groups. Instead, each group is allowed to enter until the capacity exceeds a certain *threshold* level. They showed that this form of collaboration yielded better results than the policy introduced by Litvak et al. (2008). This forms the motivation to also apply the policy from Chan et al. (2018) in this research.

5. **To show the practical applicability of the developed model in a case study of WKZ and three peripheral hospitals in the region of Utrecht**
   The model is applied in a case study of the MC of the WKZ. This research aims to set guidelines for the MC unit of the WKZ and similar units from three peripheral hospitals in the Utrecht region. First, we describe the input data for the model based on data from the included hospitals, and use the input data in the models. To state if the models are of good practice, we test the quality of the solution based on an objective that consist of the collective blocking probability, and the fraction of time that the hospitals run one or more overbeds. Furthermore, we make a translation between these probabilities and the actual data.

6. **To show the sensitivity of the aforementioned policies to various input parameters**
   The best settings for the policies are tested with a fixed value for the arrival rate and length of stay. In reality however these values might change. For instance, an increase in the arrival rate may result as a consequence of an RS virus outbreak. We want to discover the behaviour of our models in scenarios with different input values and find out to what extend our models are how sensitive to the inputs because we deisre our models to show a robuust behaviour towards a change in input parameters. By this, we can find whether our proposed settings are still beneficial compared ot the current situation or if different settings should be chosen.

The remainder of this report is as follows: in Chapter 2, we present literature found on pooling strategies. Chapter 3 describes the problem and presents the solution approach. In Chapter 4, we apply the problem to a case study of the neonatal ward from the WKZ. Chapter 5 describes the experiments, Chapter 6 the results and Chapter 7 the conclusions and recommendations.

Chapter 2

Literature review

This chapter presents the literature on capacity issues in healthcare. This chapter contributes to the first research objective, that is "*to learn from literature about the most recent methods for tackling capacity issues in neonatal, emergency, and intensive care, that include solutions in collaboration*". Section 2.1 gives a helicopter view of capacity allocation decisions in healthcare. Next, Section 2.2 provides relevant work on the stochastic modelling of patient arrivals and service times. Then we discuss relevant work regarding the concept of pooling in Section 2.3 and introduce the modelling of hospital wards by means of Markov chains in Section 2.4. To determine how we identify the quality of our models, we propose relevant work on performance outcome measures in Section 2.5 and in Section 2.6 we summarize and conclude our litreature review. This section sums up the identified gaps, such that we can indicate the contribution to the scientific body of knowledge.

2.1 Capacity allocation decisions

Maternity facilities are heavily stochastic systems where capacity requirements such as dimensioning the ward size (De Bruin et al., 2010), cannot be set deterministically (Pehlivan et al., 2012). As a solution, Winston and Goldberg (2004) state that queuing theory can be used to find performance measures that set an analytical relationship between capacity and the required service level. Queuing theory is used to capture relationships between the patient arrivals, flows through several wards, the utilization of servers, waiting times and refusals (Green, 2006). An application of queuing theory is Markov models, which is subject to Section 2.4. Worthington (1991) suggests that increasing service capacity (the traditional method of attempting to reduce long queues) has little effect on queue length because, as soon as patients realize that waiting times reduce, the arrival rate increases, which increases the queue again. As a consequence, sharing resources is often opted as a solution (Ata & Van Mieghem, 2009; Bekker et al., 2017). Furthermore, redesigning the blueprint can help to improve the wards' performance in terms of access to care (Huang et al., 2014). A blueprint supports the principle of assigning patients into categories based on similar characteristics. Huang and Kammerdiner (2013) suggests classifying patients to obtain a group with a low coefficient of variation of their characteristics. The motivation for this is that a scheduling system, in that case, performs cost-efficient. In blueprint schedules, we distinguish two types: *dynamic* and *static* blueprint schedules. Dynamic schedules are blueprints that are able to respond to variability in both demand and supply of resources supply (Hulshof et al., 2013). They have shown promising performances (Huang et al., 2014; Hulshof et al., 2013). Static blueprints include long-term cyclic plans (Hulshof et al., 2013). A challenging assignment is to balance fixed and flexible capacity in static blueprints (Kortbeek et al., 2014). Kortbeek et al. (2014) showed that for a pre-operative appointment clinic, 20 % to 40 % of the capacity should be flexible in order to cope with fluctuations in patient arrivals. Blueprints offer great opportunities for tactical planning purposes but the implementation in neonatal care goes along with the limitation the

neonatal department mostly encounters emergency arrivals (Hoot & Aronsky, 2008). However, the idea of classification classification of patient types and the application of triage principles to those types forms promising solutions for neonatal care.

2.2 Modelling patient arrivals and service times

Modelling neonatal patient wards is challenging compared to modelling other hospital flows (Kanai & Takagi, 2021) due to the urgency of arrivals and unpredictable progression. Kanai and Takagi (2021) has applied a discrete-time Markov chain model in s non-terminating simulation with a simulated run length of 2 years, to characterize the flow of neonatal inpatients over several wards in a hospital. Their probability distributions on the number of patients staying in each ward can be used as input for modelling the neonatal ward and examine the effect of interventions.

Various situations can occur in case a patient arrives at a station that is completely filled (Keskinocak & Savva, 2020). These include the following (Hoot & Aronsky, 2008):

- The arriving patient is rejected and leaves the system;
- The arriving patient waits for a bed;
- In case there are no beds, the patient in the best condition is being discharged and a new patient is admitted;
- The arriving patient is admitted by creating an *overbed*;
- The arriving patient is admitted to a different ward than was initially intended.

Frequently, a combination of the above events occurs (Hoot & Aronsky, 2008). In case of urgent arrivals, no queuing is possible (Zonderland & Boucherie, 2012) thus patients are either served or blocked; we say that the system is *overloaded* and the amount of patients being deferred equals the *overflow*. Overflow systems are typically difficult to solve (B. van Dijk, 2022) due to the system's complexity. Applications are e.g., healthcare (Litvak et al., 2008), call centers (B. van Dijk, 2022), and telecommunication networks (Asaduzzaman et al., 2010). Overflow models in the neonatal wards have been studied before by Asaduzzaman et al. (2010) and they allow the size of the overflow to be adjusted. In practice, however, the number of available beds is limited (B. van Dijk, 2022), thus we cannot create extra beds. Van Dijk and van der Sluis (2009) provides a different approach, by limiting the part to what extent one server accepts overflow from other arrival types. In a bounded overflow scenario, one accepts servers to a fixed bound B; in a probabilistic scenario, in case a server is fully occupied, the other server accepts the arrival with a certain probability P. By deciding to what extent patients are allowed to overflow, optimal capacity allocation decisions are made. The systems discussed by Van Dijk and van der Sluis (2009) only accept calls designated for other servers in case the original server is occupied. In practice, this implies that a patient cannot be refused in case the ward is not yet fully occupied. Thus, the question arises if we can shift calls to other servers even before maximum capacity is reached. Intensive care wards should strive for an occupancy rate of 80 % (Tierney & Conroy, 2014), since this occupancy rate is believed to be the optimal rate where one keeps a buffer for emergency arrivals. The optimal occupancy for neonatal wards may differ from this rate as they the neonatal ward differs from the ICU in terms of patient characteristics (Shah et al., 2015) but an optimal percentage was not found.

2.3 Jointly reserving beds by means of pooling

The concept of pooling is defined in a variety of ways (Cattani & Schmidt, 2005; Laws, 1992; Wischik et al., 2008) and it can be used to achieve a variety of objectives (Hoot & Aronsky, 2008). We look for models that approach to examine the effect of pooling or joint spreading of arrivals.

Multiple definitions of pooling were found. Laws (1992) defines pooling as a system of two identical parallel queues where arrivals join the shorter queue. It is also defined as "making a collection of resources behave like a single resource" (Wischik et al., 2008). There appears to be a widespread belief that pooling service capacities are advantageous, at least in terms of performance and capacity (Cattani & Schmidt, 2005; Van Dijk & van der Sluis, 2009). For example, Cattani and Schmidt (2005) states that "pooling of customer demands, along with pooling of the resources used to fill those demands" yields operational improvements. For identical servers and arrival types, N. M. van Dijk and van der Sluis (2008) show that indeed, the average waiting time decreases. However, in the case of various arrival types, the waiting times do not necessary decrease (Bekker et al., 2017; Vanberkel et al., 2012). It is therefore an open question if pooling is effective in a network of MC locations.

2.3.1 To pool or not to pool?

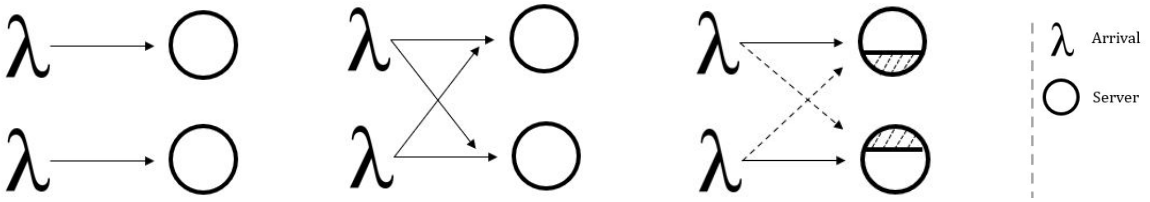
To assess whether pooling should be initiated, one must balance the benefits against the drawbacks (Cattani & Schmidt, 2005). Pooling is a potential method to improve a system's performance without having to add additional resources (Vanberkel et al., 2012). Besides, by sharing resources, the arrivals' variability will be reduced, causing servers to require less capacity to maintain a certain service level (Cattani & Schmidt, 2005; Van Dijk & van der Sluis, 2009; van Doremalen, 2022). Furthermore, to cope with fluctuations in patient demand, patient access times may be improved when resource capacity can temporarily be increased by means of pooling (Vermeulen et al., 2009). Lastly, aggregating services is a way to decrease costs and thus gain efficiency through Economies of Scale (*EOS*) (Vanberkel et al., 2012). As a remark to the benefits, the arrangement must be mutually beneficial for providers to agree to pool their resources. If a supplier's performance suffers as a result of his participation, he has no reason to do so (Nandigam et al., 2019). Pooling is not always beneficial. Resources are usually dedicated to a patient type, hence sharing them is not always possible in practice (Vermeulen et al., 2009). Additionally, employees gain expertise by practising specific procedures frequently so pooling might lead to a loss of experience (Vanberkel et al., 2012). This gain in efficiency is referred to as Economies of Focus (*EOF*). The authors make a trade-off between *EOF* and *EOS* by means of a discrete time-slotted queuing model. To finalize, in case of a large difference in arrival and service patterns, pooling resources increases the arrival variability, thus pooling will not be effective in that case (Bekker et al., 2017; Vanberkel et al., 2012).

2.3.2 Partial resource pooling

Pooling was traditionally seen as a share of complete resources (Laws, 1992). Van Dijk and van der Sluis (2009) and Nandigam et al. (2019) discuss the concept of partial resource pooling by assigning "some resources dedicated to each group and the remaining resources shared between them" (Van Dijk & van der Sluis, 2009). Partial resource pooling opens possibilities for the allocation of resources among different patient types (Bekker et al., 2017). To this extent, Bekker et al. (2017) distinguish four possibilities:

- Each patient type is served by a separate ward, depicted in Figure 2.1a;
- All wards serve all patient types by means of single merging, shown in Figure 2.1b;
- The remaining beds are shared among all patients, with a number of dedicated beds for each patient group. This policy is referred to as the *earmarking policy*, as shown in Figure 2.1c;
- All beds may serve all patient types, but if the number of vacant beds falls below a particular level, each ward will refuse certain patient types. This policy is referred to as the *threshold policy* and will be discussed in more detail in Chapter 3.

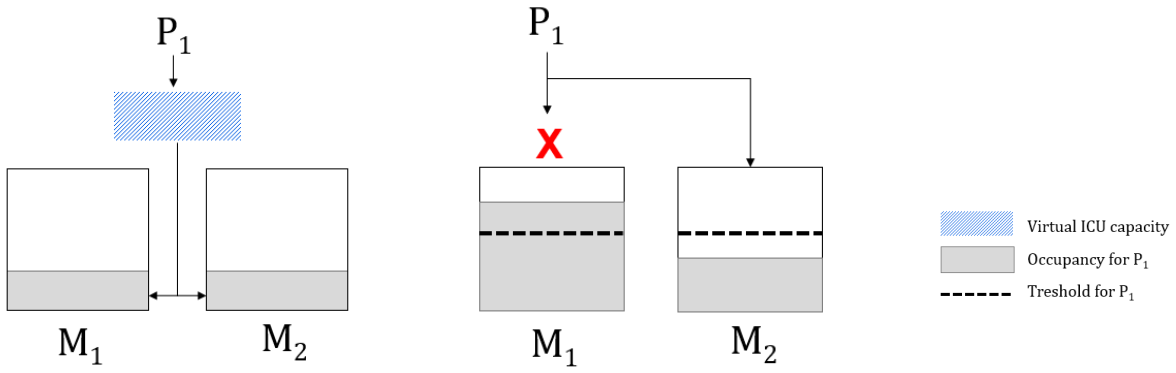
Figure 2.1: General ways of distribution of patient types among several server groups (Bekker et al., 2017)



(a) Separate wards **(b)** Single merging **(c)** Earmarking policy

As an extension to pooling for a single resource type as discussed in (Vanberkel et al., 2012), two remarkable policies are presented by Litvak et al. (2008) and Chan et al. (2018): the *virtual ICU* and the *threshold* policy, as shown in Figure 2.2. In the virtual ICU policy, several hospitals jointly reserve capacity, forming a "virtual" hospital shown in Figure 2.2a. In the threshold policy, some patients are barred from a pre-determined number of beds such that some capacity remains available for other patient groups. When the threshold is reached, some patients are not allowed to enter. This last policy allows for more detailed possibilities by implementing thresholds for specific patient groups.

Figure 2.2: Illustration of pooling for multiple server types



(a) Virtual ICU policy by Litvak et al. (2008) **(b)** Threshold policy by Chan et al. (2018)

The description of the three patient types is as follows. Type 1 patients may be admitted to any location in the network and are only blocked, if all beds in the entire network dedicated for this patient type are occupied. Type 2 patients will trigger an overbed if the location is full, since these patients always have to be admitted. Third, type 3 patients who cannot be treated right after after arriving are blocked. This relates to patients in neonatology who, for instance, require an MRI scan. If the ward is already full, the decision is made to delay the care for these patients because the situation is not life-threatening and does not require immediate treatment. They do not request admission in other locations, in contrast to type 1 patients.

To illustrate the concept of the threshold policy in more detail, consider the following. Shown in Figure 2.2 is a district with two hospitals, each having an MC unit, M_1. and M_2. In situation 2.2a, an arriving patient of type *1*, P_1, is sent to the virtual ICU which is formed by beds that are put aside in M_1 and M_2. In situation 2.2b, each hospital has a certain amount of beds for patients of type *1*, P_1 reserved. Suppose an arriving patient P_1 to location is refused since the occupancy level exceeds the threshold set for that patient type but at location MC_2, this is not yet the case. Then, P_1 will be served by location MC_2. Chan et al. (2018) proved that assigning capacity in this way reduces the overall number of refusals while maintaining appropriate occupancy levels in intensive care.

2.4 Modelling hospital wards: Markov chains

A Markov process is a stochastic model that depicts a series of potential occurrences where each event's likelihood is solely determined by the state it reached in the preceding event (Winston & Goldberg, 2004). Thus, the description of the current state captures all the information impacting on how the process evolves in future states. In other words, a Markov process is said to be memoryless (Norris, 1998).

A Markov chain can be defined on a discrete-time or continuous-time scale (Norris, 1998). A discrete-time Markov model switches between states in specified points in time; so, changes in the system can only happen at one of those fixed time values. In contrast to discrete time steps, a continuous-time Markov chain changes states continuously over time. A continuous-time Markov chain (CTMC) is a continuous stochastic process in which, for each state, the process will change state according to a random variable and then move to a different state as specified by the probabilities of a stochastic matrix. Consider for example a game of chess; the board position only changes at the finish of a player's turn. The system only moves between states if one player has played a move, so at fixed points in time. Now consider a grocery store. The number of customers in the store evolves over time, as some enter the store and other customers exit. Many circumstances, such as the cashier speed, and the grocery list of the customer, make that the length does not change at a fixed point in time, but is a continuous process. Note that both examples hold the memoryless property of a Markov chain. The board position at the beginning of the next player's turn in a game of chess depends only on the current board position and the current player's decision. The number of customers in a store for the next time period only depends on the current number of customers and the circumstances that change during that next time period.

2.5 Performance outcome measures

Performance outcome measures describe the efficiency of a system (Bamford & Chatziaslan, 2009; Hoot & Aronsky, 2008). A system's performance is defined as the sum of its activities during a specified period (Bamford & Chatziaslan, 2009). Bamford and Chatziaslan (2009) found that the issues with performance measurement have progressed from monitoring the incorrect actions to measuring too many actions, resulting in an information overload. We need to find measures that are relevant in describing the actions of the system.

The number of refused admissions is commonly interpreted as a good service level indicator and important for the quality of care (Pehlivan et al., 2012). The number of arrivals does not equal the number of admissions, which is why De Bruin et al. (2010) propose an adapted occupancy rate. Another indicator that is often used in emergency wards is the number of occupied beds (Hoot & Aronsky, 2008). Several definitions of occupancy are found. For example, De Bruin et al. (2010) consider the admissions (including refusals) multiplied by the length of stay, as a fraction of the number of operational beds.

Creemers and Lambrecht (2008) use the total flow time as an outcome measure, that is, the total waiting time plus the processing time. Other models focus on minimizing costs Harrison and López (1999) and Mahar et al. (2011), with the result that these models are highly sensitive to the cost structure. The cost structure of the neonatal facility is complex. Besides, in peripheral hospitals, a different objective holds regarding the profit objective which makes incorporating the cost structure complicated (Harrison & López, 1999). Furthermore, (Mahar et al., 2011) incorporates the quality of service by a constraint that forces a target level. Song et al. (2020) study the effects of off-service placement, that is, the situation in which a patient is laid to an open bed at a unit designated for other patient types due to full capacity at the desired ward. Lastly, Hulshof et al. (2013) use the objectives

to achieve equitable access for patients and to meet production targets or serve the strategically agreed number of patients.

We conclude that various performance measures can be found. Measuring and reporting on the use of capacity is not straightforward, as the measurements are open to bias and misinterpretation (Bamford & Chatziaslan, 2009). Hence, a clear definition of the performance outcome measure is required.

2.6 Conclusion

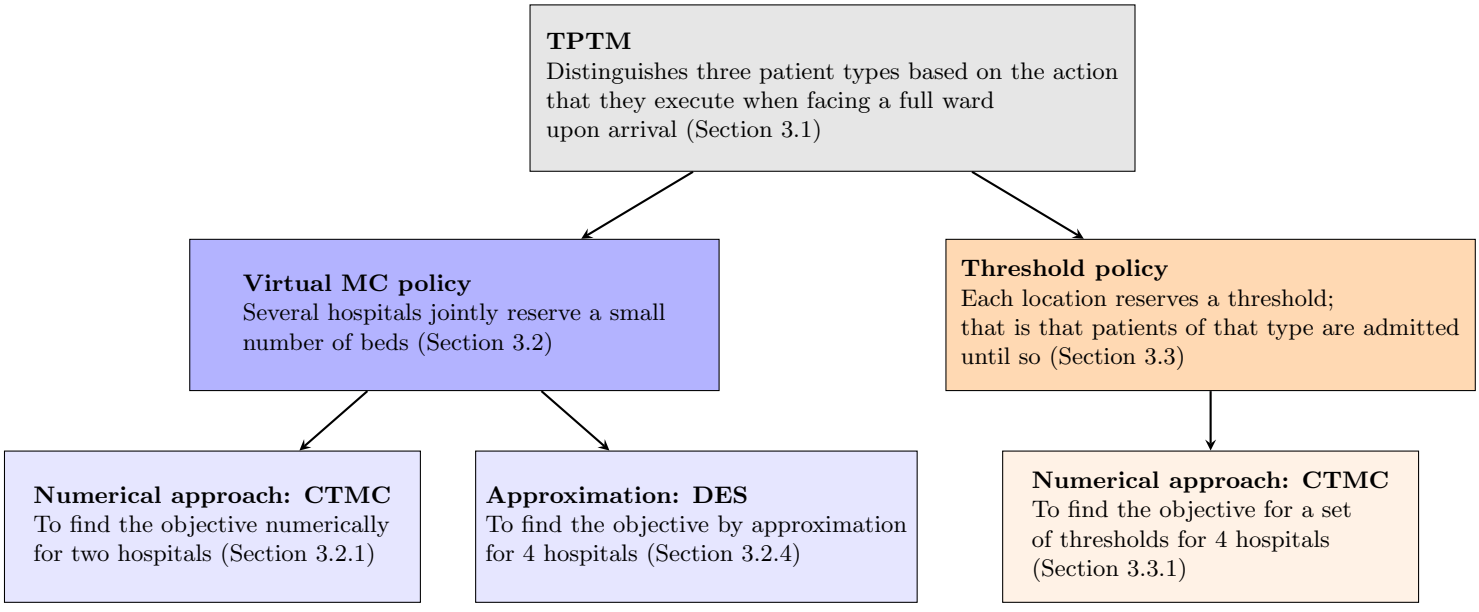
This chapter contributed to the research goal " *To learn from literature about the most recent methods for tackling capacity issues in neonatal, emergency, and intensive care, that include solutions in collaboration*". We saw that applications of partial resource pooling have previously been studied in the context of call centers or telecommunication networks, and recently, the application to healthcare has received increasing attention. Although partial resource pooling models have been studied (Nandigam et al., 2019), (Williams et al., 2015), or similar models that allow patients to be served by other servers (Asaduzzaman et al., 2010; Van Dijk & van der Sluis, 2009), only few consider constraints to the maximum amount of available beds (Chan et al., 2018; Litvak et al., 2008). Many models apply pooling for the overflow, in which patients who are unable to access their original (destination) server are assisted by another server. In case the original server still has the capacity, exchanging patients is limitely analyzed thus forming the theoretical contribution of the research. To the best of our knowledge, the overflow to peripheral hospitals by means of pooling in the neonatal context is not yet discussed in the literature.

Since we are dealing with stochastic arrivals and residence times, applying queuing theory is a promising method to start (Winston & Goldberg, 2004). Given the three-patient-type model from Litvak et al. (2008) this is a promising method to start. The virtual ICU policy may be of great interest to the neonatal care, as it offers options for resource sharing wile keeping appropriate service levels for all patient types (Litvak et al., 2008). Furthermore, the model presented by Chan et al. (2018), can be extend determine to what extent hospitals should jointly reserve beds for this patient arrival type by means of threshold reservation settings. It was found that both policies fits neonatal ward including its various arrival types the most, including the desire to make a particular amount of capacity accessible for pooling. There is a need to distribute the arrivals among the peripheral hospitals which can be achieved by a joint reservation of capacity (Pehlivan et al., 2012). The subsequent sections discuss the modelling of this joint reservation, based on the model proposed by Chan et al. (2018): the *threshold policy*.

Chapter 3

Problem description and solution approach

In this chapter we formulate a model that distinguishes three patient types, based on their behaviour upon arrival. We call this model the Three-Patient-Type-Model (TPTM). To the extent of this model, we apply two policies that incorporate regional collaboration to the three-patient-types. First, we elaborate on the *virtual MC* policy. Second, we discuss the *threshold* policy. For the virtual MC policy, we use two solution approaches. The first solution approach is a numerical approach based on Continuous-Time Markov Chains (CTMC). This approach was found to be a suitable approach for two hospitals, but when increasing the number of locations, the solution space becomes too big to be solved in a feasible time. Therefore we choose to apply a Discrete Event Simulation (DES). The threshold policy is solved numerically using CTMC. The objective that determines the quality of our solution is the sum of all blocking probabilities, and the fraction of time that each hospital spends while having overbeds. A graphical overview of the used methods is presented below.



3.1 The three-patient-type-model

The MC locations are often short of capacity. Typical solutions in case of bed shortage are: transferring a patient to another hospital, refusing a patient, or admitting the patient by creating an overbed. Patients arriving at at any MC in the network are of one of these three classes, which differ in the decision for admittance to the MC. Let $t \in 1, 2, 3$ represent the respective patient types. We assume that once the patients of type 3 are unable to receive treatment, this last group is "removed" from the system. A summery of the patient types and corresponding actions that they face when they have a full ward is as follows.

Type 1	Refusal, and try for admission in other hospital. If no bed is available in any of the locations, this pat
Type 2	Always admitted by creating an overbed
Type 3	Refusal, and no possibility re-try for admission somewhere else in the network

We introduce a network where the set I represents the locations in a network. In total, the network has $i = 1...I$ hospital locations. Each location is responsible for serving its own catchment area of patients. Let $p_{i,t}$ denote the fraction of type t patients at location i. The fraction of type 1, type 2 and type 3 patients is respectively indicated with $p_{i,1}$, $p_{i,2}$, and $p_{i,3}$, where $p_{i,1} + p_{i,2} + p_{i,3} = 1$, $\forall i \in I$.

$\lambda_{i,t}$ denotes be the arrival rate (the average number of patients arriving per day) of patient type t to location i. Let μ_i denote the service rate (the average number of patients treated per day), of location i. We assume that the service rate is similar for every patient in location i, so, we do not discriminate between the patient type and therefore do not include an index for the patient types t. We assume that the time between two subsequent arrivals follows an exponential distribution. This assumption is verified in Appendix B. If the inter-arrival time follows an exponential distribution with rate $\lambda_{i,t}$ time units, then, the average arrival rate per time unit is said to be Poisson distributed with rate $\frac{1}{\lambda_{i,t}}$ (Winston & Goldberg, 2004). Furthermore, we assume that the service time follows an exponential distribution with rate μ_i, which Appendix B justifies.

The arrival rate, $\lambda_{i,t}$ is the fraction of patients of type t at location i, $p_{i,t}$ from the total arrival rate to location i, that we denote with Λ_i. In other words, $\lambda_{i,t} = p_{i,t} * \Lambda_i \forall t \in T, \forall i \in I$. By estimating $\lambda_{i,t}$ in this way we ensure that $\sum_{t=1}^{T} \lambda_{i,t} = \Lambda_i \ \forall i \in I$.

Under the assumptions of Poisson arrivals and exponential service rates, we model the system as an M/M/c queue. This is a stochastic process whose state space is the set {0, 1, 2, 3,...} where each value corresponds to the total number of patients, including those who are currently in service. Figure 3.1 schematically depicts the patient flows for two hospitals.

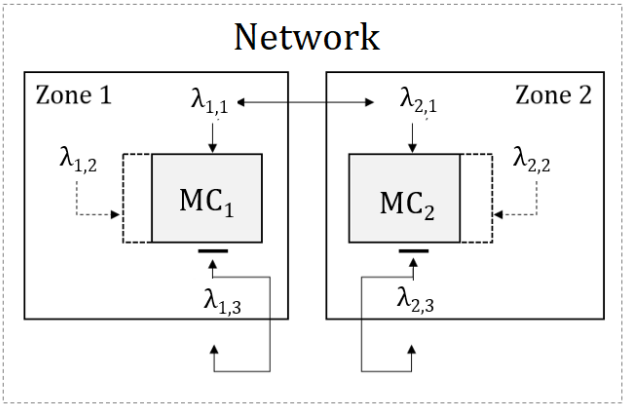


Figure 3.1: Illustration of the three-patient-type model in a network with two locations. Each zone has one MC location. Let $\lambda_{i,t}$ be the arrival rate of patients of type t arriving to location i. External emergency patients arrive at location 1 with a rate of $\lambda_{1,1}$. If MC 1 is full, the patient will attempt admission to the MC 2. Internal emergency patients and elective patients arrive to MC 1 with rates of respectively $\lambda_{1,2}$ and $\lambda_{1,3}$. A temporary overbed will be created in response to an internal emergency patient who cannot be placed in a regular MC bed, and the elective patient will be refused

Litvak et al. (2008) and Chan et al. (2018) refer to the three types as *external emergency*, *internal emergency* and *elective* patients. The models from the authors focus on the intensive care unit (ICU). We apply the TPTM to neonatal care. Since we apply the TPTM to a different context, the assumptions of the TPTM from the authors hold under different conditions. A comparison of our model and the models from Litvak et al. (2008) and Chan et al. (2018) is presented in Table 3.1.

Models from Litvak et al. (2008) and Chan et al. (2018)	**Models in this research**	**Motivation**
Application to the ICU	Application to the Medium Care unit of a neonatal department	The actions that occur for each patient type for accommodation in the ICU network, are similar to the actions in the MC unit
Internal emergency patients emerge from the zone in which the hospital is located and restrict to service solely in the hospital located in their zone of origin	Internal emergency patient can arrive via other hospitals and do not restrict to service in their origin zone	Patients categorize based on their origin type. Transporting newborns includes a lot of risk thus is not desirable; once they require treatment directly after being born, they cannot be rejected thus we scale them under 'internal emergency patient'. Additionally, patients requiring treatment at the WKZ (ones with a close risk of third-line care) cannot be rejected thus are assigned to this category
Elective patients are patients that require ICU treatment after surgery. Since this surgery can be delayed, their ICU admission can be postponed as well	Elective patients come for check-ups, scans or other medical issues that can be postponed	Patients planned via a waiting list scale under elective patients. To the best of our knowledge, these patients are the only patients that can be considered *elective* for the situation in neonatal care
Elective patients are deferred, and re-join the queue at a later moment	Elective patients are blocked and forever lost	In the model from Chan et al. (2018), the arrival rate of type 3 patients is not adjusted for refused type 3 patients, thus the arrival rate does not change if a patient of type 3 re-joins the queue. We therefore consider, the type 3 patients that cannot be accommodated upon arrival, blocked and lost.

Table 3.1: Comparison of the TPTM by Litvak et al. (2008) and Chan et al. (2018), and the adapted version in this research

TPTM Model objective

The objective is to minimize the blocking probability, that we denote with P_i^b, for type 1 and 3 patients, and the probability of having overbeds for type 2 patients, that we denote with P_i^o.

Notation	Name
Sets	
T	Set of patient types
I	Set of hospital locations
Indices	
i ∈ I	Hospital locations
t ∈ T	Patient types
Variables	
P_i^b	Blocking probability for type 1 and 3 patients in location i
P_i^o	Probability of having overbeds for type 2 patients in location i
P^B	Total blocking probability
P^O	Total fraction of time spend in overbeds
Parameters	
$\lambda_{i,t}$	Arrival rate of patients of type t arriving to location i
Λ_i	Total arrival rate to location i
μ_i	Service rate of patients in location i
$p_{i,t}$	Fraction of type t patients at location i

Table 3.2: Overview of the TPTM notations

Regarding resource sharing, we highlight two policies: the *Virtual MC* and the *threshold* policy. Both policies have different ways of assigning capacity to patient groups based on their resource-sharing models. The quality of the policies is determined by an objective, that we denote with Z. Z consists of the following elements:

$$Z = \begin{cases} \text{Fraction of incoming type 1 and 3 patients that cannot admit, hence are refused;} \\ \text{Proportion of time the hospital is running one or more overbeds} \end{cases}$$

The first item is also denoted as the "blocking probability"; to the second item, we refer to as the "overbed probability".

Assumptions on the TPTM

The TPTM goes with several assumptions that are listened below.

1. **There is an unlimited capacity for overbeds**
   To be able to treat all type 2 patients, we assume that there is no capacity restriction on this type. This implies that nurses can take care of an unlimited amount of patients. So, if the current amount of patients in the system $> C_i$, an overbed is created for this patient type.
2. **Patients of type 2 and 3, can only be served in the hospital from their catchment area**
   We assume that patients of type 2 cannot be refused and type 3 patients are neither sent away to an other location. Thus, these patients are always treated in hospital location i, where $z = i$.
3. **Each location should operate at least 1 bed**
   Closing one of the facilities that is a part of the TPTM is not ethically appropriate because each hospital is accountable for servicing its own catchment area and a portion of the patients cannot be refused.

3.1.1 Obtaining the objective in two different policies

With both policies, we obtain Z in a different way. To give an understanding of the difference in the two policies, Table 3.3 describes the difference in terms of a general description, practical meaning and goals.

Table 3.3: Difference of the virtual MC and threshold policy, in terms of general description, practical implication and the goal of the policies

Policy	**Description**	**Practical meaning: How does the policy look like in practice?**	**Goal of the policy: How does the policy find Z**
Virtual MC policy	Hospitals in a network jointly reserve capacity for the admission of type 1 patients	Beds will be distributed over the hospitals in the region, creating a virtual MC	To find the amount of capacity to assign to each patient group at each location, including the size of the virtual ICU, in which Z is minimized
Threshold policy	No capacity is explicitly reserved for a patient type, but type 1 patients are excluded from the last threshold value of beds.	Type 1 patients can admit to any location in the network, as long as the current capacity of that location does not exceed the threshold value.	To find the optimal threshold values, in which Z is minimized

3.2 Virtual MC policy

Consider a network of four hospitals that jointly reserve a small number of beds for the admission of type 1 patients that cannot be placed in their original location. A form of a stochastic model of such a network is a Markov chain, with the state being the number of patients, and the admission and discharge modeled as possible transitions between the states. For an extensive discussion on the concept of Markov chains, we refer to Section 2.4. We formulate a Continuous-Time Markov Chain (CTMC) for the virtual MC (VMC) policy. Since the CTMC is only able to capture the behaviour of the network for a small number of locations, we introduce Discrete Event Simulation (DES) to evaluate a network with more than two locations.

3.2.1 CMTC: Birth- and death processes

Our formulation of the CTMC is based on a birth-death process, that is a special case of a CTMC where the state transitions are of only two types: "births", which increase the state variable by one and "deaths", which decrease the state by one. For an elaborate overview on CTMCs we refer to Section 2.4. We define n_i as the number of patients in a normal bed at MC i, and o_i as the number of patients in an overbed at MC i. Furthermore, we denote v as the number of patients in the virtual MC. The number of patients that can be admitted to a normal bed is constrained by the capacity of MC i, that we denote with C_i. Furthermore, the number of patients that can be admitted to a bed in the virtual MC is constrained by the capacity of the virtual MC, that we denote with C_{vmc}. Thus, we have a state space $S = \{n_1, o_1, n_2, o_2, ..., n_i, o_i, v | n_i \geq 0, n_i \leq C_i, v \leq C_{vmc}, \forall i \in I\}$.

We are interested in the fraction of time spent in each blocking- and overbed state and therefore want to find the steady-state distribution π. We denote Q as the transition rate matrix. In Q, we denote the rate in which the system moves from state S to state S' with the transition rate $q_{SS'}$.

The birth processes in Q are as follows:

$$q_{SS'} = \begin{cases} \lambda_{i,1} + \lambda_{i,2} + \lambda_{i,3} & \text{if S' = (..., } n_i + 1, ...), \text{ and } n_i < C_i, \forall i \in I \\ \lambda_{i,2} & \text{if S' = (..., } o_i + 1, ...), \text{ and } n_i \geq C_i, \forall i \in I \\ \lambda_{i,1} & \text{if S' = (..., } v_i + 1), \text{ and } n_i \geq C_i, \forall i \in I \end{cases}$$

The death processes in Q are:

$$q_{SS'} = \begin{cases} \mu i * S & \text{if S' = (..., } n_i - 1, ...), \forall i \in I \\ 0, & \text{otherwise.} \end{cases}$$

Table 3.4 presents additional notations for the CTMC.

Notation	Name
n_i	Number of patients in a normal bed in location i while being in state S
o_i	Number of patients in an overbed bed in location i while being in state S
v	Number of patients in the virtual MC while being in state S
S	State space
$q_{SS'}$	Transition rate from state S to state S'
Q	Transition rate matrix
π	Steady state vector or stationary distribution
C_i	Capacity of location i
C_{vmc}	Capacity of the virtual MC

Table 3.4: Overview of model notations for the virtual MC

To give a better understanding of the Markov chain, in Figure 3.2 we illustrate an example of a transition rate diagram with two hospitals, each having a capacity of two. The numbers that correspond

to the arrows in this Figure are used to support the following context.

We transit from S={0,0,0,0,0} to a state where one patient is in either one of the locations with rate $\lambda_{i,1} + \lambda_{i,2} + \lambda_{i,3}$, and transit back with rate μ_i **(1, 2)**. We can also transit between states where respectively one or two hospitals have one patient **(7, 8)**. For simplification, we only focus on the first hospital. In S={2,0,0,0,0} full capacity is reached. The only birth and death process is to transit to a state with one patient in the virtual MC **(3)** with rate $\lambda_{i,1}$ and μ_i, or to a state with one overbed **(6)** with rates $\lambda_{i,2}$ and μ_i. A transition between states S= {2,0,0,0,1} and S= {2,1,0,0,0} is also possible with the arrival of a new type 2 patient with rate $\lambda_{i,2}$ or discharge with rate μ_i **(10)**.

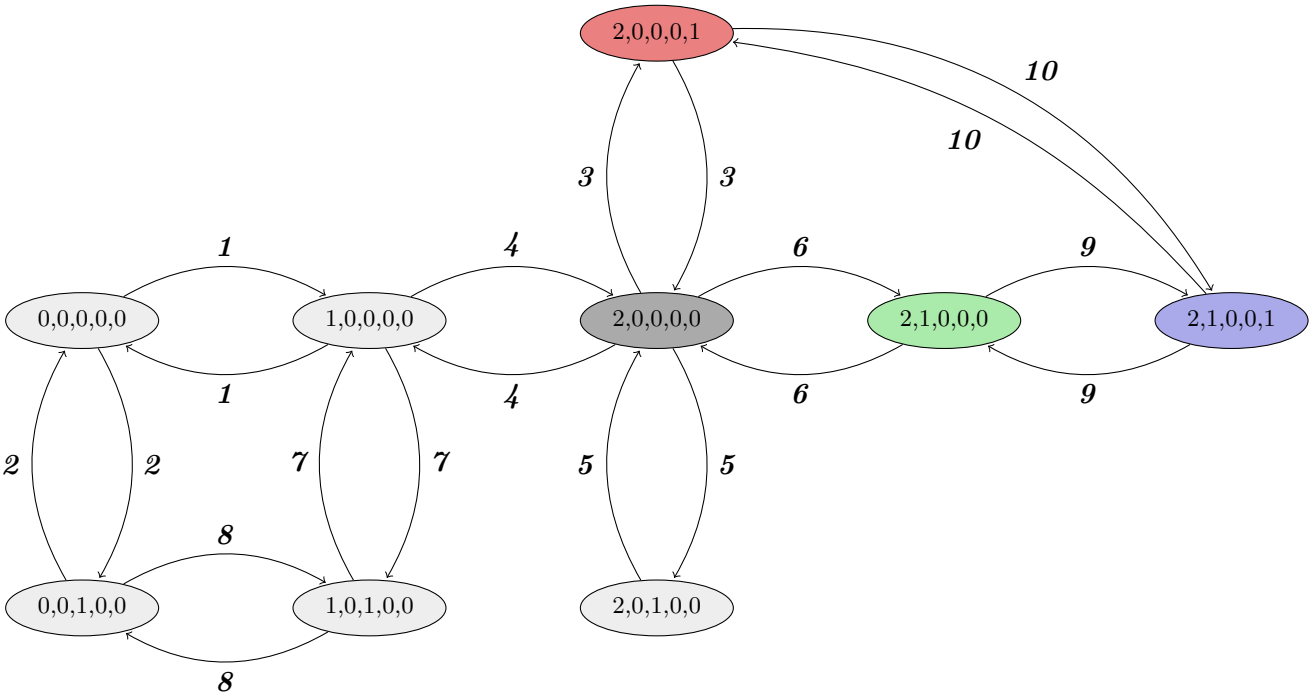


Figure 3.2: VMC policy - Transition diagram of a M/M/s system. Shown are the transition rates from state S to state S' for two hospitals with a capacity of two.

Next, we describe how we fill the matrix Q and to find the steady state vector, π.

3.2.2 Numerical approach for calculating the stationary distribution

Given the transition rates, we are interested in the stationary distribution or steady state vector π such that we can calculate the time spent in each blocking- and overbed state. Therefore, we fill a transition rate matrix Q with the values from the transition rates. Q is a square matrix where states S are displayed on the column vector and the states S' on the row vector.

If Q is positive, it approaches a steady state which is given by the eigenvector π associated with eigenvalue $\lambda = 1$ (Seneta, 1980). In Appendix E we elaborate on the relationship between eigenvectors and Markov chains and explain why this statement is true.

From the eigenvector definition, we see that if $\boldsymbol{v}$ is an eigenvector, $\boldsymbol{Av} = \lambda \boldsymbol{v}$ holds. The transition matrix Q applied to the steady state vector π recognizes as π being an eigenvector of Q with an eigenvalue λ of 1, such that $Q\pi = \lambda\pi \leftrightarrow Q\pi = \pi$. Intuitively it makes sense that this formula is true; consider a system that has a distribution over all of its possible states, and that distribution is equal to the system's steady state. Now suppose that you apply the transition matrix Q to your system. Since your system was already in steady state prior to this action, it should follow that the result

of this action is also equivalent to your steady state. This is reflected in the formula $Q\pi = \pi$. This means we have to solve $Q\pi = \pi$ to find π. Algorithm 1 describes this procedure how we do so.

At the start of our algorithm, the matrix Q is filled with transition rates. In order for Q to become a Markov matrix, we have to normalize the values for each row, such that the sum of each row = 1. This implies that for every state S, all the probabilities of leaving to state S' sum up to 1 such that you ensure you will leave state S. This is executed in steps **(1)...(6)**. Next, in **(7)** we transpose Q for the following reason. Solving $Q\pi = \pi$ equals solving $\pi(Q - I) = 0$, where I represents the identity matrix. This statement is justified in Appendix E. Since Q is now a Markov matrix, it holds that det(Q-I) = 0 (Seneta, 1980). This again implies that, Q = Q.T (transpose), so $\pi(Q - I) \leftrightarrow \pi(Q.T - I)$. We transpose Q and substract the identity matrix I from Q in respectively steps **(8)** and **(9)**.

Furthermore, since π represents our stationary distribution of Q, we have to incorporate a constraints that ensures that the sum of all elements in $\pi = 1$ (Seneta, 1980), which we will ensure by an extra row vector to Q with ones **(8)**. Appendix E describes why this procedure ensures the constraint is fulfilled. Next, initialise the array π with zero's, except the last value which is 1 **(9)**. Finally, after completing all modifications to Q, we are able to calculate $Q\pi$ using a linear solver. We implemented the algorithm on a Lenovo T460s computer with 32gb memory and using the Numpy Linear solver package.

Algorithm 1 Numerical approach for calculating π

Input: transition rate matrix Q
Output: stationary distribution vector v

```
for row in Q: do
    for column in Q: do
        RowSum ← Q[row,column]
    end for
    Q[row,column] ← Q[row,column] / RowSum
end for
Q ← Q.T
Q ← Q - I
Q ← + Array with one's
v ← Empty array with zero's
v[last item] ← 1
solve Qπ
```

3.2.3 Calculating the objectives with π

The objective consists of the system's blocking probability and the fraction of overbeds in the system. From the TPTM formulation, we recall that these objectives are respectively denoted with P_i^b and P_i^o. To find P_i^b, we have to ask ourselves the question: "*When exactly is a patient blocked?*" The answer to that question depends on the patient type. A type 1 patient is blocked if and only if the hospital itself *and* the virtual MC is full. A type 2 patient is never blocked and a type 3 patient is blocked, if the hospital itself is full. We recall that $p_{i,t}$ denotes the fraction of type t patients at location i. The probability that a patient is being blocked while arriving to location i is,

$$P_{i,1}^B = p(\text{arrival of type 1}|\text{we are in a state where type 1 cannot be accommodated}) \quad (1)$$

$$= p_{i,1} * \sum(\pi(n_i, o_i, ..., v)|n_i \geq C_i \ \& \ v = C_{vmc}) \qquad \forall i \in I \quad (2)$$

$$P^B_{i,3} = p(\text{arrival of type 3}|\text{we are in a state where type 3 cannot be accommodated}) \quad (3)$$
$$= p_{i,3} * \sum(\pi(n_i, o_i, ..., v)|n_i \geq C_i) \qquad \forall i \in I \quad (4)$$

Finally, we calculate P^B_i as the sum of the two types,

$$P^B_i = P^B_{i,1} + P^B_{i,3} \qquad \forall i \in I \quad (5)$$

We calculate the second KPI, the "proportion overbeds" as follows. The proportion overbeds is also known as the proportion of time that the system spends in a state where one or more overbeds is presented. We denote this proportion with P^O_i. This outcome measure only holds for type 2 patients, since it is assumed that only this type may be accountable for overbeds. Then,

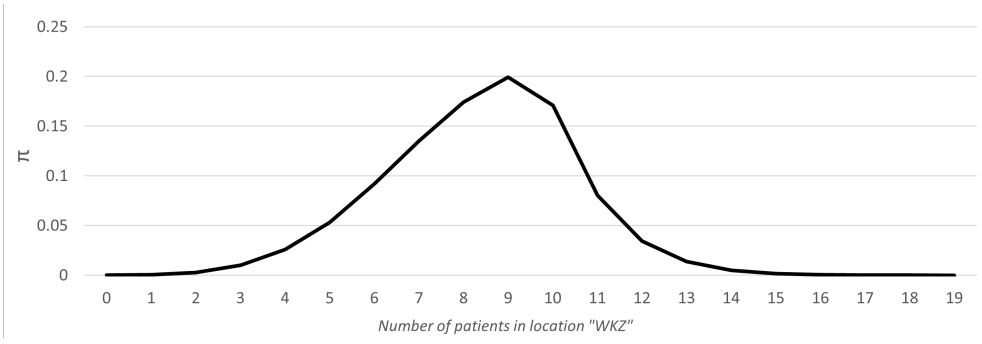
$$P^O_i = p(\text{We are in a state where we have overbeds}) \quad (6)$$
$$= \sum_{i=1}^{I} \pi(n_i, o_i, ..., v)|o_i > 0 \qquad \forall i \in I \quad (7)$$

We remark that states with overbeds only exists where $n_i \geq C_i$. Therefore, we do not have to require that $n_i \geq C_i$ in the states for calculating P^O_i.

Dealing with a large solution space

With more hospital locations and increasing size of the virtual MC, the number of possible configurations increases. With four hospital locations and 10 to 12 beds per location, an analytical solution for the set of equations 3.2.3 to 3.2.3 difficult to realize due to the size of the solution space. Since the hospital only has a certain number of resources to operate overbeds, limiting the maximum number of overbeds is one technique to reduce the size of the solution space. Figure 3.3 presents the value of π_j for $j \in [0, 1, .., 19]$. This Figure shows that the states $(\pi_j|j > 15)$ are very close to zero. In fact, the system spends only 0.078 % of the time in a state π_j where $j > 15$. Therefore we neglect states with more than 5 overbeds.

Figure 3.3: Distribution over the states for location $i = 1$ with a maximum of 9 overbeds allowed. Shown is that the fraction spent in states $\pi(n_i, o_i, ..|o_i > 15)$ is very low



After this reduction, Table 3.5 presents the number of possible states in which the virtual MC consist of respectively 1 and 2 beds per location. When we include four hospitals and each hospital donates two beds to the VMC, Algorithm 1 has to provide π_j for 67321 states. Since the sum of all elements of $\pi = 1$, each state's value is extremely small and hence quite inaccurate.

Description	Number of possible states
1 location	10
2 locations with no virtual MC	225
2 locations, size of the virtual MC = 2	297
2 locations, size of the virtual MC = 4	340
3 locations with no virtual MC	4096
3 locations, size of the virtual MC = 3	4744
3 locations, size of the virtual MC = 6	4671
4 locations with no virtual MC	65536
4 locations, size of the virtual MC = 4	50625
4 locations, size of the virtual MC = 8	67321

Table 3.5: Number of possible states with an increasing amount of locations and capacity of the virtual MC

We conclude that an numerical approach as presented in Algorithm 1 is difficult to realize. Therefore, we we developed a simulation model to study the whole network.

3.2.4 Discrete Event Simulation approach

There are many types of simulation models (Hill, 2007). The behavior of the neonatal ward is most accurate represented by a discrete system, that is a system in which the state variables change instantaneously at separated points in time. We choose to perform a discrete event simulation. An event in this simulation may for example be, the admission or discharge of a patient. The DES was developed in Technomatrix Plant Simulation software[1]. In the subsequent subsections we describe the model settings that are required to execute the DES.

Model settings

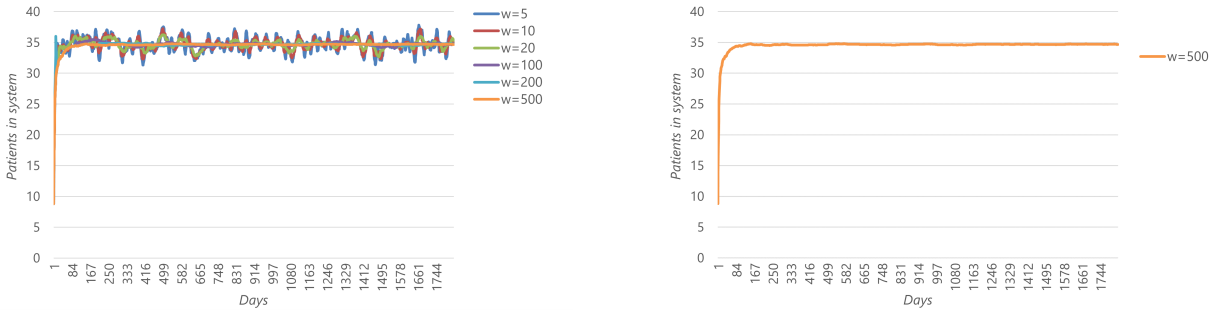
To design the model, we determine: (1) the warm-up period and (2) the number of replications.

Warm-up period
At the start of the simulation, the ward is empty. There are no patients in the system yet so every patient can be admitted, which means that the blocking probability is zero. This data is not a good reflection of reality and should therefore not be included in determining the outcome measures. For this reason, a warm-up period is required, such that the system is in steady state behavior after this period is passed.

To determine the warm-up period, we use Welch's graphical method. Five independent runs of each 10 years were performed, and the average number of patients per day in the system was calculated for every run. Next, we plotted the moving averages with windows of 5, 10, 100, 200 and 500 over time. Figure 3.4a shows the results of Welch's graphical method for the windows of 5, 10, 100, 200 and 500. Figure 3.4b shows the results for w=500 only.

[1]Tecnomatrix Plant Simulation Version 15.2 ©Siemens

Figure 3.4: Average number of patients per day over 5 years



(a) Different window lengths

(b) Window length of 500

We observe that after roughly 167 days, the number of patients in the network stabilizes. To be sure that out outcomes are not dependent on an empty system anymore, we choose a warm-up period of 365 days (1 year).

Number of replications

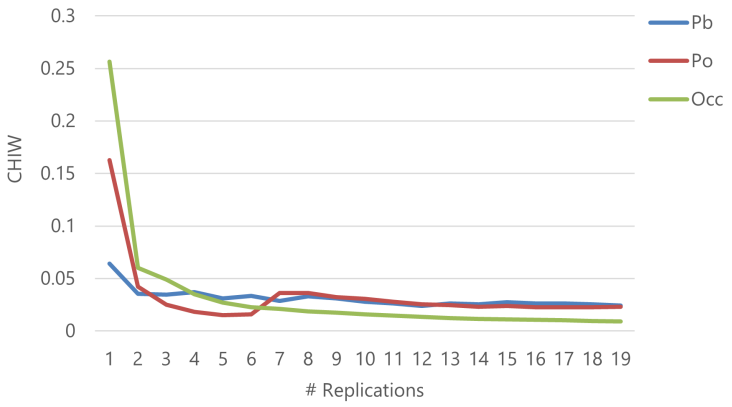
The arrival- and service times are stochastic parameters. We want to ensure that the simulation parameters reflect the true population's parameters. We seek for the sample size so that we can apply the central limit theorem. An aspect of this theorem is namely that the average of the sample mean and standard deviation, equals the population mean and standard deviation.

In every run, we take a sample for the arrival and service rates. In total, we make 20 independent replications. In each replication, we measure three performance indicators:

- The number of refused patients per i
- The number of overbeds per i
- The daily average number of patients as a percentage of the total capacity

The average number of patients as a percentage of the capacity, is what we refer to as the occupancy. We choose a number of replications until the width of the confidence interval relative to the average, or the Confidence Interval Half Width (CIHW) is sufficiently small for both the performance indicators. To decide what is sufficiently small, we choose a confidence level of 95%, and a relative error of $\gamma = 0.05$ thus a corrected target value γ' of 0.0526. Figure 3.5 shows the progression of the CIHW as the number of runs increases. For P_i^o and P_i^o, 2 replications per experiment are sufficient to assure

Figure 3.5: CIHW against the number of replications for the whole network



that that the simulation parameters reflect the true population's parameters. For the occupancy, we need to perform at least 3 replications. To make sure that we perform enough replications to reflect the true population's parameters for all performance indicators, we choose to perform 5 replications per experiment.

Simulation flowchart

A flowchart that shows the logic of the simulation model is shown in Appendix C. This appendix shows various events in the simulation model that all trigger a sequence of steps to be followed. The first event is the arrival of a patient. The arrivals take place based on a derived distribution as presented in Appendix B. The patient type of the arriving patient determines which sequences of actions will be followed. The other event that triggers actions, is the discharge of a patient. When this patient is a type-2 patient and there are patients in overbeds, the patient moves from "overbed" to a "normal bed" because the overbed is not an actual overbed anymore.

Correcting for occupation

Besides the KPI's blocking probability and overbed-fraction, we are also interested in the effect of the VMC on the occupancy rates. We want to make sure that the occupancy rates are high enough to achieve the production agreements but not so high that they endanger the standard of care. According to Tierney and Conroy (2014), conventional hospital wards operate best when they aim for an occupancy percentage of 80%. However, there is no known set value for neonatal wards (Shah et al., 2015), so we suggest that an occupancy of about 80% should be strived for in order for our approach to not negatively impact the quality of treatment.

The occupancy of the VMC is not compensated for the real locations since the VMC is modelled as a separate location at first. The beds in the VMC are essentially those that are given from the real locations, In order to discover the true locations' occupancy, we must correct for the share of beds that belong to the VMC but are in fact located at hospital i. The simplest way how we do so, is to illustrate an example using numbers. Suppose that location 1 donates **2** beds to the VMC and location 2 donates **3** beds. The capacity of locations 1 and 2, are **10** and **12**, respectively. Now suppose the VMC has an occupancy of **80** % and both locations have an occupancy of **90** %. Location 1 has **10-2 = 8** beds left of its own and **2** beds of the VMC. Then, we calculate the total occupancy as (8/10) * 90 % + (2/10)*80 % = 88 %. By calculating the occupancy in this way, we correct for the share of beds that different locations donate to the VMC.

3.3 Threshold policy

In this section we formulate the policy inspired by the threshold policy from Chan et al. (2018). We formulate a one-dimensional CTMC with transition rates, $q_{j,j+1}$ and $q_{j,j-1}$ based on the threshold constraints. For that CTMC we find the steady state vector π and calculate the same two objectives as in the virtual MC: the blocking probability and overbed probability, P_i^b and P_i^o.

3.3.1 Transition rates and calculating the steady state vector

Chan et al. (2018) introduce the *threshold policy*, in which a *threshold* that specifies to what occupancy level patients are allowed to enter. Let $r_{i,t}$ be the threshold value of location i, for patient type t. Suppose there are j patients in total in the hospital; a patient of type t will only be accepted if,

$$j < r_{i,t} \qquad \text{for } t = \{1, 3\},\ i = 1, 2, \ldots, I \quad (8)$$

If $j \geq r_{i,t}$ patient of type t,$t = \{1, 3\}$, cannot enter anymore, Type 1 patients will then proceed to an other hospital. In this Section we seek to optimize the values for every $r_{i,t}$ such that the objective, the sum of all blocking- and overbed probabilities, is minimized. Table 3.6 provides an overview of the model notations for this policy.

Notation	Name
$\pi_{i,j}$	Fraction of time spent in state $\pi_{i,j}$, with j patients in location i
j	Number of patients in the system
n	Number of times a patient tried to admit somewhere else, after refusal at the original location
$\Gamma_{i,n}$	Routing policy for patients from location i that attempted n times for admission before
S	State space
$q_{j,j+1}$	Transition rate from state j to state $j+1$
$q_{j,j-1}$	Transition rate from state j to state $j-1$
C_i	Capacity of location i
$r_{i,t}$	Threshold for patients of type t in location i
x_i	Arrival intensity of type 1 patients to location i in patients per day
$a_{i,n}$	Arrival rate of type 1 patients from location i that attempted for admission n times before
$b_{i,t}$	Local blocking probability of location i for type t, for $t = 1, 3$
$p_{i,t}$	Fraction of type t patients arriving to location i
k	Iteration counter

Table 3.6: Overview of model notations in the threshold policy, new

We set up a CTMC to find P_i^B, and P_i^O. We are interested in the states in which an arriving patient finds a full ward. Unlike to the CTMC in the virtual MC, we do not set up a multi-dimensional Markov chain for the whole network, but we create an one-dimensional Markov chain per location in the network. The reason for this is as follows. In the virtual MC policy, we model the network as $n+1$ locations, where the last location is the virtual MC. The size of the virtual MC depends on the size of the other locations, so the locations cannot be treated independent. For the threshold policy however we do not share beds among each other. We treat the hospitals as independent locations. We model the system with the following characteristics. Let state j denote the number of patients in the system. Let $q_{j,j+1}$ denote the transition rate of state j to state *j+1*. Then, the transition diagram looks as follows:

Figure 3.6: Threshold policy - Transition diagram of a M/M/s system. Shown are the transition rates from state j to state *j+1*

In this policy, multiple patient types having different arrival- and service rates, apply for a bed at one location. Similar to Chan et al. (2018), we combine the different arrival- and service rates in this one-dimensional Markov chain. The transit from state j to state j+1 means that we are in a state with 1 patient more in the system. Instead of expressing the transition rates in rates of λ and μ amounts per time unit, Chan et al. (2018) uses the load $\rho = \frac{\lambda}{\mu}$ to transfer from state j to j+1, and uses j as the transfer rate to transfer from a state j to j-1.

If we are in state j, and we transfer to state $j + 1$, patients type 1 and 3 can only enter when

$$(j < r_{i,t}) \qquad \text{for } t = \{1, 3\}; \forall i \in I \quad (9)$$

So for example, if condition 9 does not suffice for $i = 1$, you can only reach a higher state if a patient of type 2 and/or 3 is admitted. To model this, we use an indicator function $\mathbb{1}\{\cdot\}$, which is 1 in case condition 9 is true and 0 otherwise. Given that we are in state j, we transit from state j to state *j+1* as follows.

$$q_{j,j+1} = x_i \mathbb{1}\{j < C_i - r_{i,1}\} + \lambda_{i,2} + \lambda_{3,i} \mathbb{1}\{j < C_i - r_{i,3}\} \qquad \forall i \in I \quad (10)$$

Where x_i represents the arrival rate of type 1 patients to location i. Type 1 patients can enter an other hospital in case the hospital that belongs to their catchment area, has exceeded its threshold. The variable x_i covers both the type 1 arrival rate from within the catchment area, as well as from outside their catchment area. the patients from outside the catchment area progress trough various locations via a routing policy. In Section 3.3 we discuss how we set up such a routing policy and how find x_i. For now, we continue with the transition rates and setting up the CTMC.

If we are in state j, we transit to state *j-1* if one patient leaves the system;

$$q_{j,j-1} = u_i * j \quad (11)$$

The implementation of Equation 10 is given in Algorithm 2. Given the location i, x_i and $\lambda_{i,2}$ and $\lambda_{i,3}$, we calculate $q_{j,j+1}$ based on the value of j **(4-10)**. After finishing all the birth rates, we continue with the implementation of Equation 11 which is given in **(14)**.

Algorithm 2 Calculating $q_{j,j+1}$ and $q_{j,j-1}$

Input: location i, x_i and $\lambda_{i,2}$ and $\lambda_{i,3}$
Output: $q_{j,j+1}$ and $q_{j,j-1}$, for $j = \{0,1,...,15\}$ for i

for j = 1,2,...,20: **do**
  for $q_{j,j+1}$: **do**
    if $j < r_{i,1}$ and $j < r_{i,3}$: **then**
      $q_{j,j+1} \leftarrow x_i + \lambda_{i,2} + \lambda_{i,3}$
    else if $j < r_{i,1}$: **then**
      $q_{j,j+1} \leftarrow x_i + \lambda_{i,2}$
    else if $j < r_{i,3}$: **then**
      $q_{j,j+1} \leftarrow \lambda_{i,2} + \lambda_{i,3}$
    else if $j \geq r_{i,1}$ and $j \geq r_{i,3}$: **then**
      $q_{j,j+1} \leftarrow \lambda_{i,2}$
    end if
  end for
  for $q_{j,j+1}$: **do**
    $q_{j,j+1} \leftarrow j * \mu_i$
  end for
end for

To solve the model within a feasible time, we reduce the solution space by neglecting states in which more than 15 patients **(1)**. Figure 3.7 shows that the values of state π_j, — $j > 15$ for $i = 1$ are very close to 0; in fact, the system spends only 0.066% of the time in a state π_j where $j > 15$, which supports the decision to neglect these states. An additional advantage is that now both the CTMC of the VMC and the threshold policy both allow a maximum of 5 overbeds so it is valid to compare both policies later on.

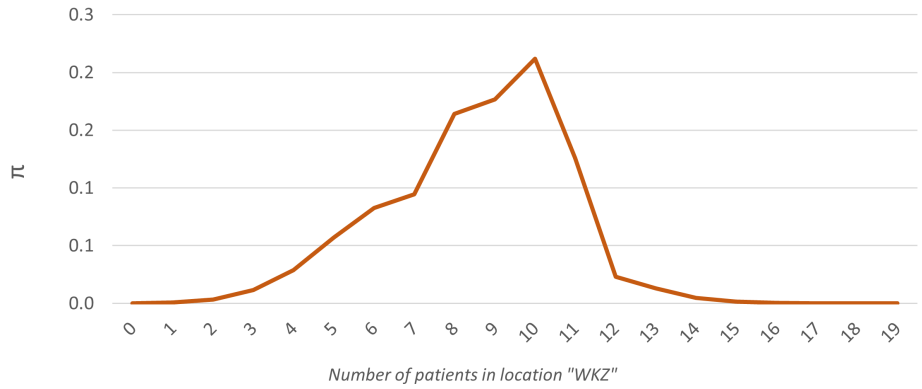


Figure 3.7: Distribution over states for *i=1* with $r_{1,t} = 0$, $\forall t \in T$. The time spent in states ≤ 15 is very small (0.066 %), thus we neglect those.

From the transition rates we calculate the probability for each state that we denote with π_j, by using the fact that $\sum_j^I \pi_j = 1$. If we found so, we continue to calculate the objectives. The blocking probability for a location $b_{i,t}$, is calculated as the sum of the states in which the hospital cannot accept an arriving type t patient anymore, $t \in 1, 3$. Thus,

$$b_{i,t} = \sum_{j=r_{i,t}}^{\infty} \pi_j \qquad t \in 1,3, \forall i \in I \quad (12)$$

$b_{i,t}$ is denoted as the "local" blocking probability because this is the blocking probability for a location, but not for the patient. In calculating the global blocking probability P_i^B we ought to note the following. We know that, to calculate x_i for one location, we require $b_{i,1}$ of all other locations since these blocking probabilities determine to what extent the locations will receive each others type 1 patients. However, at the same time, $b_{i,1}$ for one location is obtained after executing Algorithm 2 and obtaining the steady states π_j. Before we dive in the details of ED, we first explain how we incorporate the routing policy for type 1 patients in order to find x_i and b_i.

Routing policy of type 1 patients

The following applies to patient type 1. When a patient from the catchment area of location i, cannot enter their own hospital, this patient will try to enter elsewhere. The order in which patients from one location proceed to another location, is determined via a predefined *routing policy*. Only if all locations in the network no longer have a place, then, this patient is refused. The routing policy is an adapted version of the three-patient-type model from Chan et al. (2018). The routing policy gives an order in which patient from catchment area z progress trough the network. The order is given by the matrix $\Gamma_{z,n}$, where z is the catchment area and n depicts the number of times that the patient has previously attempted to enter a hospital. For type 1 patients, we define $a_{i,n}$, that is the arrival rate of type 1 patients from location i, that have attempted admission to other hospitals n times before.

It is important to make a distinction between the blocking probability that the patient experiences, and the blocking probability from a hospitals' point of view. A hospital from a certain location could have a very high blocking probability, but the blocked patients arriving to that hospital have the opportunity to enter somewhere else. Thus, the blocking probability that they experience, differs from the hospital from their location. Let $b_{\Gamma_{i,n-1}}$ be the blocking probability at the hospital that patients from the catchment area of location i attempted for admission the $(n-1)$-th time. Then,

$$a_{i,0} = \lambda_{i,1} \tag{13}$$

$$a_{i,n} = a_{i,n-1} b_{\Gamma_{i,n-1}}, \qquad n > 0 \tag{14}$$

Let x_i denote the total arrival intensity of type 1 patients to location i, that is

$$x_i = \sum_{z=1}^{G} \sum_{n:\Gamma_{z,n}=l} a_{i,n} \tag{15}$$

To explain how the routing policy behaves, we illustrate an example. Suppose that location 1 has 10 type-1 arrivals per day from their own catchment area, and this location has a blocking probability of 20 %. Then, 2 patients per day cannot enter location 1 and proceed to location 2, which will receive 2 patients per day extra. Because we assume Poisson arrivals, we can sum the arrival rates from all locations. A detailed overview of the application of matrix $\Gamma_{i,n}$ can be found in Appendix D.

In a network with four locations, there are numerous possible routes. For instance, we can decide on a set procedure where patients are transferred first to the hospital that is closest to them, then to the next closest, and finally to the hospital that is farthest away. However, in reality, there is not a fixed policy. The choice of which hospital to call first depends on a variety of circumstances that are outside the purview of this study. Hospitals call one hospital after the other. We randomly select the routing policies and select a different routing policy for each run to represent the real-world scenario most accurate. The *PolicyNr* from Algorithm 3 is the indicator which routing policy is chosen for that run.

Algorithm 3 Update x_i

Input: $b_i^{(k-1)}$for $i = 1 \cdots I$, location to update x_i for i, PolicyNR
Output: x_i
1: $x_i \leftarrow$ RoutingPolicy(PolicyNR)

3.3.2 Application of the routing policy in the network

The output, b_1, depends on x_i but at the same time, x_i depends on the result of b_1. To handle this, we use a similar approach as Chan et al. (2018) proposes. The authors use exponential decomposition (ED) to tackle the problem of finding x_i and b_i simultaneously (Franx et al., 2006). ED is applied to the network by treating x_i for each MC as mutually independent (Franx et al., 2006). "Mutually independent" means that the locations are considered independent when calculating their steady-state vectors, but their input depends on each other, so we calculate the steady-state vectors one after another. The goal of the ED is to find $x_i \forall i \in I$ such that the resulting $b_{i,1} \forall i \in I$ will return the same x_i. The steady-state vector π_j that returns the $b_{i,1}$ from Equation 12, is the output of Algorithm 4. In this algorithm, we solve x_i for one location i, which provides b_i for that i, and using that b_i in order to calculate x_i for a new location. We do so until the stopping criterium is met. We use the same stopping criterium as Chan et al. (2018) use in their ED, namely, that $b_i^{(k-1)} - b_i^{(k-2)} < 10^{-8}$.

Algorithm 4 Calculating the corrected π_j given $r_{i,t}$ for every i and t

Input: $r_{i,t}$ for every i and t; Rn = Random routing policy nr
Output: Corresponding steady state vector π_j
1: **while** $b_i^{(k-1)} - b_i^{(k-2)} > 10^{-8}$ for every i: **do**
2: i $\leftarrow$ *Random Location*
3: $x_i \leftarrow$ Update(x_i, Rn, $b_{i,t}^{(k-1)}$)
4: *Transition rates array* $\leftarrow$ CalculateTransitionRates(x_i, i)
5: $\pi_{i,j} \leftarrow$ CalculateStates(*Transition Rate Array*, i)
6: $b_{i,t} \leftarrow$ CalculateObjectives(*States array*, i)
7: **end while**
8: Return π_j

It important that the *PolicyNr* does not change during the executing of this algorithm for the following reason. If we would apply different routing policies, we would have that the stopping criteria are based on different routing policies. The values x_i depend on the routing policy that is applied, thus it would not be fair to change routing policies while optimizing b_i.

Although we handle the locations independently of one another, they are dependent on one another in order to calculate x_i. To incorporate this, we take a location for which we have an estimation on b_i, and we take that b_i as the input for a new location $i+1$. Then we recalculate x_i for $i+1$ based on $b_{i,1}$ from the previous location i. Until the stopping criterion is reached, this procedure is repeated. Figure 3.8 shows the progression of b_i after multiple replications, for every $i \in I$. We observe that roughly 3 iterations per location are required to fulfill the stopping criteria, $b_i^{(k-1)} - b_i^{(k-2)} < 10^{-8}$. This implies that we need at least 3*4 = 12 iterations in total. Since the next *location to update* is chosen from a random uniform distribution, 12 iterations is a bare minimum. We choose to perform 100 iterations at least.

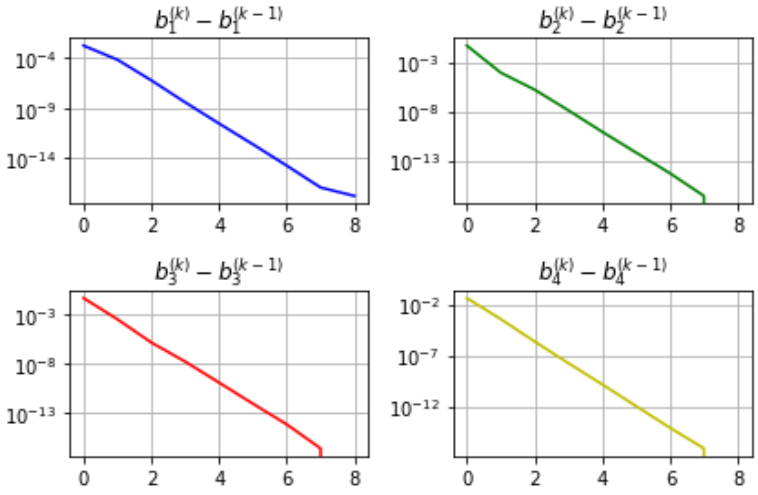
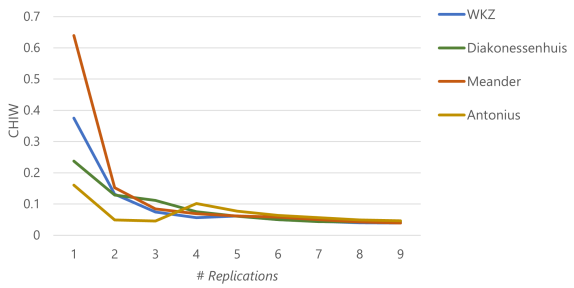


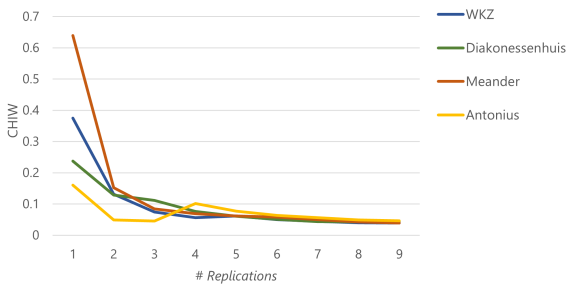
Figure 3.8: Progression of b_i for every i, which shows that we need at least 3 iterations per location to satisfy the stopping criterion.

In reality, there is no fixed preference regarding the decision where to the patient will try to enter after being denied at a previous location. It relies on a variety of factors, including the level of acuity, the travel distance, the availability of ambulances and many more. All of these factors are not captured within the model. To mimic a real-world scenario to the best extent, we need to evaluate multiple routing policies per set of thresholds $r_{i,t}$. In each run, we change the routing policy. To determine how many runs are required, we apply Welch's graphical procedure which is shown in Figure 3.9. We pick a random routing policy and optimize b_i given that routing policy, until we have reached the stopping criterion $b_i^{(k-1)} - b_i^{(k-2)} < 10^{-8}$. Then, we apply Algorithms 5 and 6 to find $P_{i,1}^B$ and $P_{i,1}^O$. In Section 3.3.2 these algorithms are discussed, but for now we focus on the procedure of finding a minimum number of runs. We store these and we continue by selecting a new routing policy and optimize b_i again. We choose a number of replications until the width of the confidence interval relative to the average, or the Confidence Interval Half Width (CIHW) is sufficiently small for both the performance indicators P_i^b and P_i^o. To decide what is sufficiently small, we choose a confidence level of 95%, and a relative error gamma = 0.05 thus a corrected target value gamma' of 0.0526. From Figure 3.9a we observe that 6 runs are enough to ensure the average result of $P_{i,1}^B$ is stable; that is, the effects of different routing policies are accounted for well enough. It was found that we need at least 7 runs to obtain stable results for the $P_{i,1}^O$.

Figure 3.9: Welch's graphical procedure to determine the number of runs in the CTMC



(a) Average P_i^b per location after multiple runs



(b) Average P_i^o per location after multiple runs

Calculating the objective

When the steady state vectors are found, given we have executed at least 8 runs to ensure the stopping criterion from the ED and 7 runs to capture differences in the routing policies, we can calculate the global blocking- and overbed probabilities, that we respectively denote with $P^B_{i,1}$ and $P^O_{i,1}$. First, we discuss how we calculate $P^B_{i,1}$.

The fraction of the time spent in a state where more patients are present than the threshold value gives, is the same as the probability that an arriving patient finds the system in that state. However, not all arriving patients are blocked; patients are only blocked if the threshold for that type for that location is exceeded. The blocking probability for type 1 patients is then, the probability that a type 1 patient arrives to location i, *given* that location i is in a state where $j \geq r_{i,t}$ and all other locations are also in a state where $j \geq r_{i,t}$. This is because, once a type 1 patient is refused, there might still be a bed available somewhere else; thus, the blocking probability is the product of the blocking probabilities from all locations in I. Furthermore, The probability that an arriving patient to location i is a type t patient is given by $p_{i,t}$. Therefore,

$$P^B_{i,1} = p_{i,1} \prod_{i=1}^{I} b_{i,1} \qquad \forall i \in I \quad (16)$$

Algorithm 5 presents how we implement Equation 3.3.2 to obtain the blocking probability from the steady state vector. This algorithm requires the steady state vectors π_j from all locations, but will return only P^b_i for one location. This is because we require the local blocking probabilities $b_{i,t}$ from all locations in the network for the blocking probability from a patient arriving to a location in their catchment area i.

For every location i **(1)**, we use the state vector π_j to add up the states for which condition 9 holds; this gives us the local blocking probability $b_{i,t}$ for that location. We perform this two times, for $t = 1$ and 3 **(4), (6)**. Next, after having obtained $b_{i,t}$ for every i, we apply Equation 3.3.2 to calculate the product of all locations which gives us the probability that all locations are full for patient type 1. We know that the probability of a patient type 1 being refused, equals the probability that an arriving patient is a type t patient while we are in a state where $j \geq r_{i,t}$ $\forall i \in I$. Therefore, we then multiply this product with the probability that an arriving patient to location i is of type 1 for, which is denoted by $p_{i,1}$. Lastly, the total blocking probability for patient from the catchment area of location i is given by the sum of the blocking probability of types 1 and 3 **(11)**.

$$P^b_i = b_{i,1} + b_{i,3} \qquad \forall i \in I \quad (17)$$

Algorithm 5 Calculate blocking probability

Input: *states* from Algorithm 2; $r_{i,t}$ for every i and t
Output: $P^B_{i,t}$ for a given i and $t = 1, 3$
for $i = 1, 2, \cdots, I$: **do**
  for j in *states*: **do**
    if *states*[j] $\geq r_{i,1}$: **then**
      $b_{i,1} += states[j]$
    else if *states*[j] $\geq r_{i,3}$: **then**
      $b_{i,3} += states[j]$
    end if
  end for
end for
$P^B_{i,1} \leftarrow p_{i,1} \prod_{i=1}^{I} b_{i,1}$; $P^B_{i,3} \leftarrow p_{i,3} b_{i,3}$ $\forall i \in I$
$P^b_i \leftarrow P^B_{i,1} + P^B_{i,3}$ $\forall i \in I$

We use a similar approach in calculating the fraction of time that is spent while having overbeds. This however is from a hospital point of view; we do not incorporate the probability that a type 2 patient arrives in calculating so. The fraction of time that a hospital spends having one or more overbeds is denoted by P^o_i and is given by:

$$P^o_i = \sum_{j=C_i+1}^{\infty} \pi_j \tag{18}$$

The hospital is having one overbed, if the current number of patients in the system j exceeds the capacity with 1. This is why we sum over states from $j = C_i + 1$ on. The Algorithm 6 describes how P^o_i is found, given the steady state vector π_j for a location i.

Algorithm 6 Calculate overbed probability

Input: *states* from Algorithm 2
Output: $P^O_{i,t}$ for a given i
for j in *states*: **do**
  if *states*[j] $\geq C_i$: **then**
    $P^O_{i,t} += states[j]$
  end if
end for

So far, we have seen how we calculate P^b_i and P^o_i with a given set of thresholds. We are interested in a set of thresholds for every i and t that minimizes the overall blocking- and overbed probability. It was found that evaluating all possible thresholds in a range from 0 to C_i was not possible due to the size of the solution space. Therefore we propose a combination of two approaches, to balance between exploring all possible thresholds and exploiting promising sets. This is discussed in Section 3.3.3.

3.3.3 Model settings

In this Section we discuss two approaches to find the best possible set of thresholds. Ideally we would evaluate all possible sets of thresholds for every i and t but that would take the algorithm too long to compute. This is caused by the following factors:

1. Per set of thresholds, we need at least 8 replications to achieve the stopping criterion of the ED;
2. The stopping criterion must hold for every $i \in I$ which causes that we need at least 10 replications per routing policy per threshold;
3. Per set of thresholds, we need to perform at least 7 different routing policies, to capture the variability that is caused by a different routing policy

Table 3.7 shows an estimation on the average running time in days. The estimated running time is calculated by measuring the running time of the first 10 configurations and scale this up to the total number of possible states.

Table 3.7: Estimated running time with an increasing number of states. As the range of possible values per $r_{i,t}$ increases, the running time increases significantly.

Evaluated thresholds in range	Possible states	Estimated running time (days)
r_{i_t} = {0,1} for every i	$2^8 = 256$	0.04
r_{i_t} = {0,1,2} for every i and t	6561	1.52
r_{i_t} = {0,..,3} for every i and t	65536	1.82
r_{i_t} = {0,..,4} for every i and t	390625	10.85
r_{i_t} = {0,..,5} for every i and t	1679616	69.98
r_{i_t} = {0,..,6} for every i and t	5764801	333.61
r_{i_t} = {0,..,5} for every i and t	16777216	2905.03
r_{i_t} = {0,..,8} for every i and t	33093567	6894.49

We do not desire for the algorithm to take multiple days to run. Therefore, we have to come up with a smart solution in which as many possible states are investigated, within a feasible running time. The first action to take is to shorten the range of possible values for $r_{i,t}$.

Since the threshold is set per patient type, and each patient type represents only a portion of the overall capacity ($p_{i,t}$, it is not necessary to test every threshold in the range from $[0, C_i]$. Consider that we want each patient category to have a 0% refusal rate. We can determine how many beds we require to obtain 0% refusal rate for a given patient type, using the Erlang loss model (Weernink, 2018; Zonderland & Boucherie, 2012). If all patients of this type could be admitted with this quantity, a $r_{i,t}$ higher than this quantity of beds would be of little use. Table 3.8 shows an overview of the in- and outputs for the Erlang loss model and Equation 19 desribes how we calculate P_b with a given number of beds m.

Table 3.8: Notations for the Erlang loss model (Zonderland & Boucherie, 2012)

Input	Value
λ	The arrival intensity of patient type t to location i = $p_{i,t} * \Lambda_i$
μ	The service rate
ρ	Expected load = $\frac{\lambda}{\mu}$
m	Number of operational servers
i	Number of occupied servers
Output	**Value**
P_b	Blocking probability for patient t
N	Occupied beds as a fraction of total operational servers

$$P_b = \frac{\frac{\rho^m}{m!}}{\sum_{i=1}^{m} \frac{\rho^i}{i!}} \tag{19}$$

Since type 2 patients cannot be refused, we require a slightly different approach to use the Erlang loss model. We increase the number of beds until a blocking probability of 0% is achieved . Table 3.9

presents the results of the Erlang loss model, in which each number indicates the amount of beds that is required to obtain a blocking probability of 0 %.

Table 3.9: Required capacity to obtain a blocking probability of 0 % for that patient type t in that location i. Evaluating $r_{i,t}$ higher than these numbers is not necessary.

	Required capacity of $t = 1$	**Required capacity of** $t = 3$
WKZ	8	4
Diakonessenhuis	7	5
Meander	7	5
Antonius	7	5

Given these ranges for $r_{i,t}$, $\forall I \in I, \forall t \in T$, we obtain still a high running time. Therefore it is not feasible to evaluate all sets of thresholds. To strike a balance between exploring the solution space and exploiting promising areas, we provide a constructive heuristic which returns a promising set of thresholds that is used as input for a greedy improvement heuristic that we propose. The construction heuristic is displayed in Algorithm 7. In this Algorithm we generate 100 random sets of thresholds for every i and t. Given each set, we execute Algorithms 5 and 6 to find Z. This Algorithm returns a list of a set of thresholds and its corresponding Z. From this list we select the set that correspond with the lowest Z but taken into account that all occupancy rates are lower than 85 %. Taking this condition into account is necessary to ensure the quality of care is guaranteed (De Bruin et al., 2010).

Algorithm 7 Constructive heuristic to find a promosing set of thresholds

1: **for** k in 100: **do**
2: Generate set for every $r_{i,t}$, $\forall I \in I, \forall t \in T$
3: Z $\leftarrow$ Sum of all blocking- and overbed probabilities
4: Execute Algorithms 5 and 6 to find Z
5: store Z
6: **end for**

Greedy optimisation

To the the optimal set that Algorithm 8 returns, we provide a greedy optimisation approach to find if we can find any improvement. To exploit around the most promising set from the 100 runs, we gradually change the threshold of one location of one patient type, by an increase or decrease of 1 in line **(4)**. The amount to which $r_{i,t}$ is changed is indicated by the *ChangeParameter*. Hereby is taken into account that $r_{i,t}$ is in the range of 0 to C_i; if $r_{i,t}$ + *ChangeParemeter* $> C_i$, then a new *ChangeParemeter* is selected.

If the total set of thresholds, including the updated $r_{i,t}$ will return a lower objective Z than was found so far, we keep this $r_{i,t}$ and re-enter the Algorithm. If not, then, we discharge this $r_{i,t}$. This ensures that we do not evaluate worse solutions.

You could argue that this might keep us stuck in a local optimum. After all, we don't allow hill-climbing moves by accepting a worse solution. However, by evaluating 100 random sets of thresholds first we assume that we have explored the whole solution space widely enough to ensure that we are in a local optimum.

Algorithm 8 Improvement heuristic

for k in *RunLength*: **do**
    Select random location i and type t, $i \in I$, $t \in 1,3$
    ChangeParameter $\leftarrow$ -1 if randomnumber < 0.5; 1 else
    $r_{i,t} \leftarrow +$ *ChangeParameter*
    Use Algorithms 5 and 6 to find Z
    if Z ¡ Best Z: **then**
        Best $Z \leftarrow Z$
    else if $Z \geq$ Best Z: **then**
        $r_{i,t} \leftarrow$ - *ChangeParameter*
    end if
end for

3.4 Conclusion

First of all, this chapter contributed to the third goal, "*To apply the virtual MC policy on the three-patient-type model, by adapting the policy from Litvak et al. (2008)*". To this extend, we constructed the TPTM and discussed two policies to apply to the TPTM: the virtual MC (VMC) policy and the threshold policy. Two solution methods are applied to the VMC. The first solution method involves Contiguous-Time Markov chains and allows for a numerical analysis of the performance of the network that includes a VMC. Since this policy does not allow for evaluation when more than two hospitals are involved, a Discrete Event Simulation has been developed as well. Furthermore, this chapter has contributed to the fourth research goal, "*To apply the thresholdpolicy on the three-patient-type model, by adapting the policy from Chan et al. (2018)*". To this second policy, we proposed a solution method based on Continuous-Time Markov chains as well. Contrary to the CTMC from the VMC, we use one-dimensional Markov chains that allow for easier evaluation as opposed to using the eigenvector definition in the VMC policy. To optimise the set of threshold yielding the most profitable results, we proposed a greedy optimisation technique that combines exploring all threshold-sets, while exploiting the promising sets.

Chapter 4

Case study

The previous chapter presented a model formulation that will be applied in a case study. This chapter presents input data for the model in which we address the situation in the Wilhelmina Children's Hospital (WKZ) and three peripheral hospitals in the region of Utrecht, the Diakonessenhuis, Meander and Antonius hospital. This chapter contributes to the fifth research goal, "*To show the practical applicability of the developed model in a case study of WKZ and three peripheral hospitals in the region of Utrecht*". Data for this case study is derived from registrations in the health information system (HIS) over the year 2020 and 2021 with a focus on neonatal patients, born in 2019, 2020 and 2021. The data is obtained from the years 2020 and 2021, because in these years, the management of the neonatal ward started keeping track of the number of refusals. In the model, we use actual data from the neonatal department of the WKZ and estimated data from three peripheral hospitals in the Utrecht region, the Antonius hospital, respectively the Diakonessenhuis and the Meander hospital.

4.1 Input parameters

This section describes how we obtain the input parameters. The parameters are obtained by historical data over the years 2020 and 2021. We look over this period for the following two reasons. First, in these years there was a loss of staff due to the Covid pandemic. We thus mimic a tight scenario. Second, in 2020 and 2021 there was a proper record of the refusals. If we neglect the refusals, the arrival rate in the model is lower than the actual arrival rate at the neonatal department and our models do not show accurate results. To obtain the true arrival rate, we must add the refusals to the total arrival rate. From the WKZ and the Diakonessenhuis, we are able to find data directly; for two other peripheral hospitals, we make estimations on the arrival rate $\lambda_{i,t}$ and service rate $\mu_{i,t}$ based on data from the Diakonessenhuis.

4.1.1 Probabilities for transport, refusal and overbed in a full ward

We denoted $p_{i,t}$ as the fraction of type t patients at location i. Based on data of 2020 and 2021, we are able to estimate the ratios for the WKZ and the Diakonessenhuis. We assume that the ratios of the Diakonessenhuis hold for the other peripheral hospitals included in this research as well. Table 4.1 presents the probabilities that, given that we are in state *full*, a patient is respectively transported, placed in an overbed, or refused. Appendix **??** presents context to this estimations.

4.1.2 λ and μ

For both the WKZ and the peripheral hospitals, we cannot accurately determine the inter arrival time λ, because there is no time registered for the indirect refusals and the overbeds. For this reason we

Probability indicator	$p_{1,t}$	$p_{2,t} = p_{i,2} = p_{i,3}$
$P(transfer \mid no\ place)$	0.3	0.54
$P(overbed \mid no\ place)$	0.5	0.4
$P(refusal \mid no\ place)$	0.2	0.06

Table 4.1: $p_{i,t}$ per location i per type t

calculate the λ based on the average number of arrivals in our scope of 2 years and we assume that it follows a Poisson distribution. The number of arrivals, corrected for the refusals, is indicated in Table **??**. We refer to Appendix **??** for an elaborate discussion on the estimation of the true arrival rate. Additionally to the refusals at the MC, a significant number of patients are turned away from the NICU and obstetrics. There is a flow to the MC for these departments. Fewer patients enter the MC through the flow because they are rejected at that primary department. The patient flow that we miss as a result is referred to as indirect refusals and is subject to discussion in Appendix **??**. Furthermore, Appendix B shows an elaboration on the assumptions on Poisson arrivals and exponential interarrival times.

4.1.3 Available capacity for each location i

In this section, we describe the input C_i, the capacity for each location i. The average bed occupancy, which we define as **the number of operational beds as a percentage of the available beds**, was respectively 95.3% and 87.2% for the NICU and the MC over the years 2020 and 2021, the NICU ward including the HC ward.

The capacity of the MC of the WKZ varies between 8 to 12 beds, with an average of 10 beds. A daily assessment moment on the capacity check, reveals this. However, the real occupancy changes during the day, making it challenging to associate an exact number with it. The fluctuation in care intensity, which is challenging to measure in neonatology due to the complexity of the care, is the cause of the occupation's fluctuation. When there are 10 beds available, the daily assessment shows that 10% of the days from 2020 and 2021, one or more overbeds were present. The presence of overbeds contributes to a high occupation. As the neonatal department has an emergency function due to the acuteness of arrivals and almost no planned care, such a high occupancy rate is unacceptable from the emergency care accessibility point of view (Ogboenyiya et al., 2020). This emphasizes the need to consider occupation as an additional KPI in this study. Table 4.2 shows the capacity that we assume each location has. Although the actual capacity fluctuates throughout the day, we assume a fixed value, such that this does not complicate the modelling purposes.

	WKZ	Diakonessenhuis	Meander	Antonius
C_i	10	11	11	11

Table 4.2: Capacity per location i

4.2 Conclusion

In this chapter, we found input data for to apply the models to a case study. We used real data from the WKZ and Diakonessenhuis and estimated data from two peripheral hospitals in the region of Utrecht. This input data is required to achieve the fifth research goal, that is, "*To show the practical applicability of the developed model in a case study of WKZ and three peripheral hospitals in the region of Utrecht*" We presented input values for the available capacity, the arrival rate, length of stay and the share of patient types that each region includes. For the three peripheral hospitals included in this research we assume similar input parameters.

Chapter 5

Results

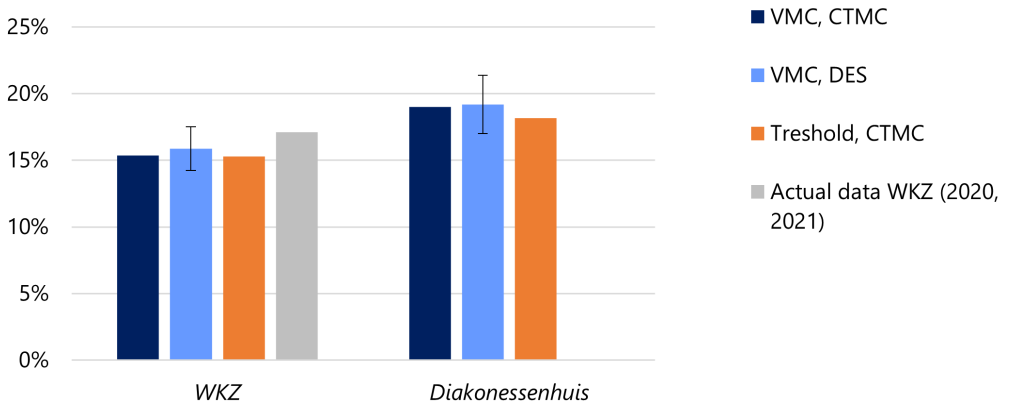
In this chapter we present the experiments. First, we will present a baseline measurement of all three models in Section 5.1. This baseline measure is required to test the validity of the models. Second, we present the results of the virtual MC policy for both the CTMC as well as the DES model. Third, the results of the threshold policy are discussed. We use "percentage point" (p.p.) as an indicator to show differences in policies. Using "percentage point" is a proper measure to indicate an amount of change, while the use of "percent" rather describes a relative change, but we believe that showing the absolute amount of changes is more insightful for the management of the neonatal wards. For example, an decrease from 20 % to 10% means a decrease of 50% while, in fact, the KPI is still 10 %. We end this chapter with a sensitivity analysis on our best models found. The goal of this analysis is to test the robuustness of our models against changes in various input parameters.

5.1 Baseline measurement

To be able to judge the performance of the Virtual MC and threshold policy, we need to compare the results of the policies in which no intervention is set, with the actual performance. The relevant KPI's in this measurement are the blocking probability, the number of overbeds as a fraction of total beds and the occupation. The VMC-DES results are presented with a 95 % confidence interval (CI), meaning that we are 95 % sure that the true value of the KPI is in this interval. The CTMC in the baseline measure has a standard deviation of 0 because the result in every run is the same. Therefore, baseline measure for the CTMC does not include a CI.

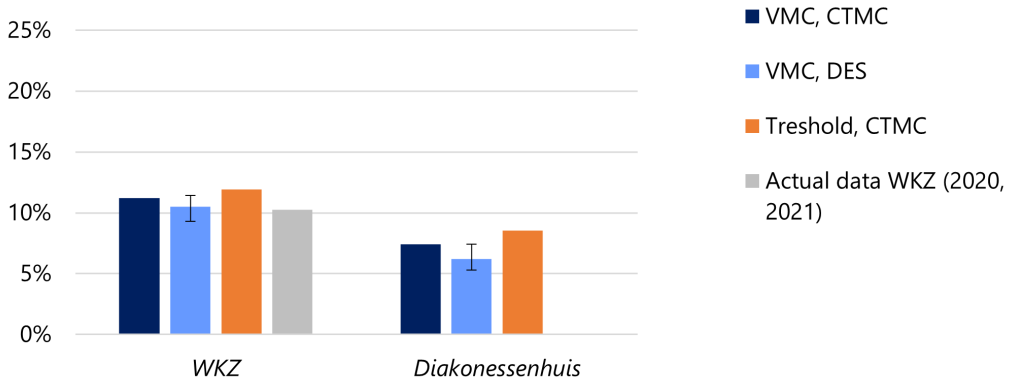
The actual refusal probability is obtained by dividing the number of direct and indirect refusals to the total arrivals over the years 2020 and 2021. In the baseline measurement for the threshold policy, we set $r_{i,t} = C_i \ \forall i \in I, \ \forall t \in T$. Furthermore, we set $a_{i,n} = 0 \ \forall n \in \Gamma$, such that each location will solely serve the patients from its own zone and no collaboration takes place. Figure 5.1 shows the result of the baseline measure on the blocking probability. The Figure shows that the CTMC of both policies show similar results for the WKZ but differ slightly for the Diakonessenhuis. Furthermore, although the average of the DES is slightly higher, the CTMC is in the confidence interval of the DES. The actual data shows that the blocking probability for the WKZ is 1,5 p.p. higher in reality than the models provide. We assume a fixed value for capacity for the models in the baseline measure, which may be the reason why the actual situation differs somewhat from the forecast of the models. In practice, the capacity relies on the acuity of care and the available staff. We noticed that the capacity can range between 8 to 12 beds, which cause that there might be more refusals than the models predict in case of a tight scenario.

Figure 5.1: Baseline measurement of P_i^B for $i = 1, 2$



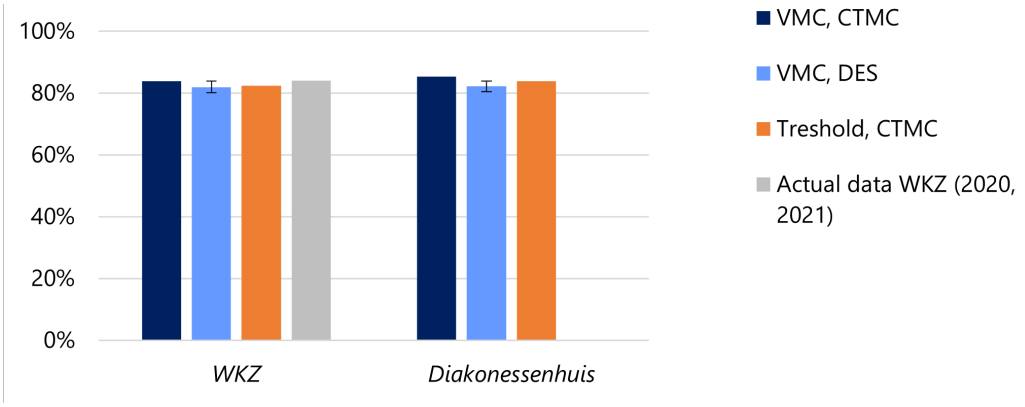
In the daily records that are kept on the number of available and occupied beds, the number of available beds varies, we assume a fixed value for the capacity. Of all days in 2020 and 2010 that the capacity was 10 (128), 13 were found with an occupancy of more than 10 beds, so that means that overbeds were run 10% of the time in real life. Regarding the fraction of overbeds, we notice that the threshold policy says that the baseline measure is 2 p.p. higher than the VMC models predict, as well as the reality for the WKZ. Whereas the threshold policy showed a lower prediction for the blocking probability, the fraction of time spent in overbeds is slightly higher. This is caused by the fact that we balance between refusing and having overbeds, so if one KPI is higher, the other is lower and vice versa. The difference in performance of the baseline measurement and the actual data can again be explained by the variety of available beds which is captured in reality but not in the baseline models.

Figure 5.2: Baseline measurement of P_i^O for $i = 1, 2$



Lastly, we conduct a baseline analysis on the occupation rates in the three policies. We notice a similar occupation in three models, as well as the actual data. This is shown in Figure 5.3.

Figure 5.3: Baseline measurement of the occupation of locations $i = 1, 2$



5.1.1 Conclusion of the baseline measure

We conclude that a minor difference in base performance is present among the different models. This can be caused by variations in available capacity which are captured in the actual data but not in the models, that assume a fixed capacity. Although the blocking and overbed probability show minor percentage point differences amongst the models, we assume that the models in the baseline measure perform similarly. We consider this as a valid assumption because the difference in the overall objective Z, that consist of the sum of blocking- and overbed probabilities, is considerable small. Table 5.1 shows the difference in p.p. between the different policies. We find that these values are very close to 0, why we assume a similar performance on the baseline model.

Table 5.1: Difference in p.p. in Z between the different policies (1= Treshold - CTMC; 2=MC - CTMC; 3=VMC - DES, 4=Actual data WKZ (2020, 2021))

Policy	1	2	3	4
1		0.0063	0.0025	0.0019
2	0.0063		0.0038	0.0019
3	0.0025	0.0038		0.0044
4	0.0019	0.0019	0.0044	

Assuming that the models perform similarly in the baseline measure, we will continue to present the results when examining how the baseline measures performed in comparison to the improvement models.

5.2 Virtual MC policy

This section presents the results from the virtual MC policy. First, we present the results of the policy obtained from a numerical approach in the CTMC in two hospitals. Second, we present the results from the Discrete Event Simulation. By the introduction of the VMC, locations will have to run over beds sooner because they will have to give up capacity due to the VMC's presence. Therefore, the basic setting, in which there is no VMC, achieves the lowest number of overbeds, as both models will demonstrate later on. The introduction of the VMC has shown for both models to increase the time spent with overbeds. It is not the question which setting regarding the VMC size decreases both the blocking probability (P^B) and the time that locations will turn overbeds (P^O) but rather which setting can lower the possibility of refusal the most while increasing the time with overbeds the least. In that, the setting is not only saying which capacity the VMC should have, but also How many beds the locations should give up in order to attain the desired outcome.

5.2.1 CTMC

We compare the performance of the VMC with the current performance as indicated in the baseline measure for two hospitals, the WKZ and the Diakonessenhuis. To this extent, a maximum reduction of overall objective Z of 8 p.p. was found. Table 5.2 presents the optimal number of beds that both locations should keep for themselves, and the size of the VMC, which is that only the Diakonessenhuis donates one bed to the VMC (and the WKZ none). When implementing this setting, the blocking probability lowers but the hospitals have to spend more days in a state with overbeds.

C_1	C_2	C_{vmc}
10	10	1

Table 5.2: Number of beds for the location, and the size of the VMC for the lowest objective

Figure 5.4 presents the results of the baseline measure and the best objective for two locations in terms of blocking probability. The results show that, when the Diakonessenhuis dedicates one bed to the pool, the blocking probability from the patients from the catchment area of the WKZ decreases by 8 p.p. (percentage point), and drops by 15 p.p. for the Diakonessenhuis. An decrease of 8 p.p. in $P^O_{i,t}$ on a population size of 1612 arrivals in two years for the WKZ, means that the number of refusals drops by 128 in two years time, which is equivalent to treating 1.24 patients per week more. A reduction of 15 p.p. means 153 less refusals to the Diakonessenhuis in two years time or, similar, treating 1.47 patients per day more. From the perspective of P^B, the VMC is very profitable for both locations. One might suggest that it is even more profitable for the WKZ because this location does not contribute any beds and the WKZ catchment area is bigger in population size. However, the share of type 1 patients from the WKZ's catchment area is lower in comparison with the share of type 1 patients from the Diakonessenhuis' catchment area (30 % versus 54 %), so this will fade away the consequences of the bigger population size.

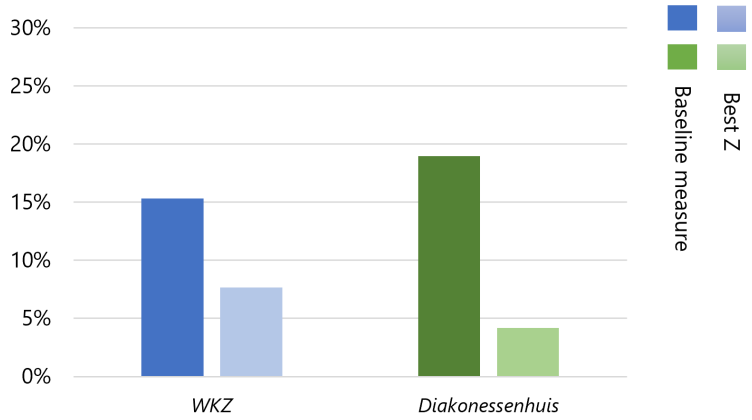


Figure 5.4: Virtual MC Policy - CTMC - Blocking probability (%) for patients from the WKZ and Diakonessehuis catchment area

The results for the fraction of time spent while having overbeds show a less promising picture as shown in Figure 5.5. The results give the impression that introducing a VMC is worse for the WKZ (increase of 4 p.p.) and also for the Diakonessenhuis (increase of 10 p.p.). An increase of 1 p.p. for a timeframe of two years means that 7.3 days more, will have to be spend with one or more overbeds. Turning overbeds is a huge burden for the staff. The WKZ has to spend 29 days more in this circumstance, which is relatively seen, 40% more. For the Diakonessenhuis this is even worse; the overbed-fraction increases with 73 days or relatively seen 143 %. It is logical that the Diakonessenhuis experiences the worst outcome, since they deliver one bed and the WKZ does not.

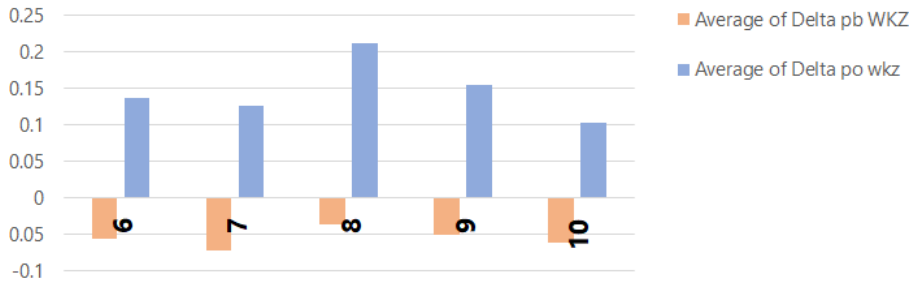


Figure 5.6: Virtual MC policy - CTMC - Average result of P^B and P^O for capacity the WKZ has left

We find it a strange observation that the WKZ's fraction of time spending overbeds increases while this location does not deliver any beds to the VMC. As the input parameters between the baseline measure and the measure with Best Z do not change, it is not to expect that the location delivering no beds to the VMC experiences overbeds.

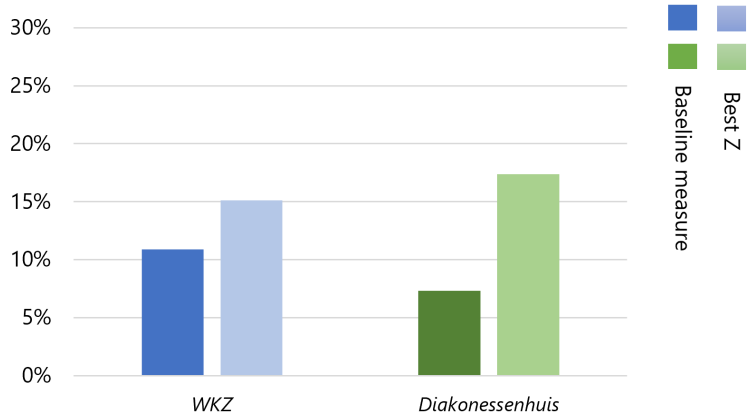


Figure 5.5: Virtual MC Policy - CTMC - Fraction of overbed-days(%) for patients from the WKZ and Diakonessehuis catchment area

The VMC setting that showed the most decline in objective was found when the WKZ does not donate. To investigate the model behaviour more torouhgly, we check what P^O will be when the WKZ does deliver beds. Figure 5.6 presents the average values of P^B and P^O on that the WKZ donates. For every capacity value of the WKZ, we run the capacity scenarios for Diakonessenhuis when they donate 0 to 4 beds. Although P^O on average is the lowest when no beds are donated (equivalent to a capacity of 10) compared to the donation of beds, we still encounter an increase to what we would expect to be zero. By the introduction of the VMC there are more routing possibilities so there will be more states. The objective is a sum of many states; the more states the CTMC consists of, the less accurate the model becomes. From the observation of a rise in P^B for the WKZ we conclude that the CTMC solution approach of the VMC policy is not sufficiently accurate which causes the model to overestimate or underestimate certain outputs.

Finally, the occupation rates increase with 10% for both locations, which comes down to 84 %. This is shown in Figure 5.7. Similar to our observations in a lack of accuracy in P^O we are not certain if this result is valid, however, it is believed that lower than 85 % still form proper conditions in neonatal wards (Ogboenyiya et al., 2020).

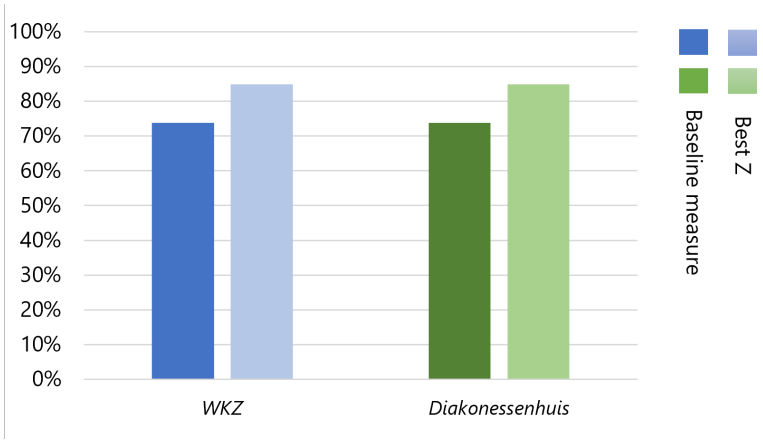


Figure 5.7: Virtual MC Policy - CTMC - Occupation (%) for patients from the WKZ and Diakonessehuis

The numerical approach for a network with two locations has shown that the introduction of the VMC has benefits and drawbacks that vary by KPI. We consider what the VMC accomplishes for the network as a whole to strike a balance between all the variables. In a 2-hospital scenario, the optimal setting was found when only the Diakonessenhuis donates one bed. In a near-optimal setting, the setting where both hospitals donate a bed each, the Diakonessenhuis will profit even more in terms of blocking probability, but this is at the cost of an increase in overbed-fraction as shown in Table 5.3. The WKZ in this case has considerably less beneficial results, why the collective objective is worse than the collective objective where only the Diakonessenhuis donates.

Table 5.3: VMC, CTMC - Difference in objective between the optimal and near-optimal settings

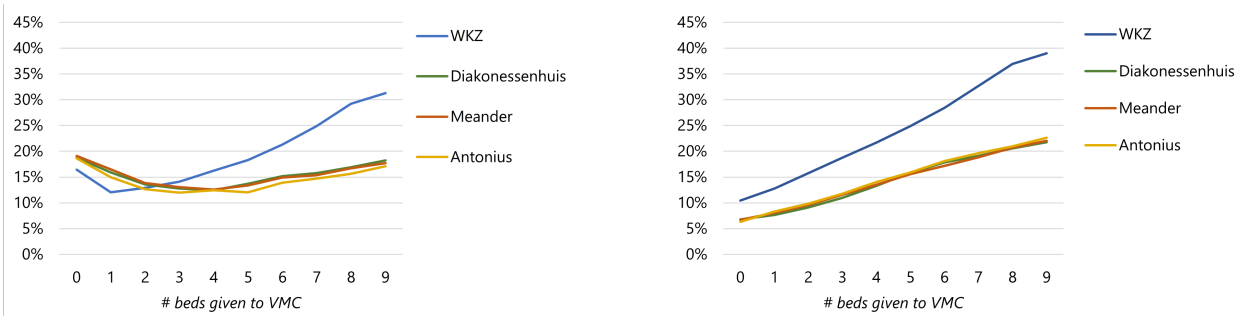
	ΔP_1^B	ΔP_1^O	ΔP_2^B	ΔP_2^O
Best setting (Diakonessenhuis donates 1 bed)	-8	4	-15	10
Near-optimal setting (both locations donate 1 bed each)	-6	10	-16	11

5.2.2 DES

Each location donates a similar amount of beds to the VMC

The impact of each location donating the same number of beds to the VMC was examined first, in a range from 1 to 9 beds. We do not look at the effect of more than 9 beds, because we assume that each location should be opened. For further explanation on this assumption we refer to Section 3.1. Figure 5.8a presents the effect of an increasing amount of donated beds on the blocking probability P_i^B. Figure 5.8b shows the effect of this intervention on P_i^O.

Figure 5.8: Virtual MC policy - Effect on P_i^B and P_i^O when each hospital contributes respectively 1 to 9 beds to the VMC



(a) P_i^B per location (p.p.)

(b) P_i^O per location (p.p.)

When three beds are relocated, we observe that the refusal rate at the peripheral hospitals first falls by roughly 5 p.p. before beginning to rise again. With each additional bed that the WKZ vacates after donating 1 bed, the chance that the WKZ will deny the arrival rises proportional. So, the WKZ would not benefit from jointly reserving beds in this way, solely considering the blocking probability. Figure 5.8b shows a contrary picture for the fraction of time that each hospital spends having one or more overbeds. Every hospital becomes more and more crowded, and this trend lines up with the total number of overbeds listed. This is a logical result, which is caused by the fact that we take away free capacity for type 2 patients, so the more capacity is taken away, the sooner an overbed will be created. Aditionally, it makes sense that the WKZ would experience this effect the most strongly given that they have 1 bed less capacity. There is only 1 bed left when they contribute 9 beds, and an overbed is also employed in 40% of the time.

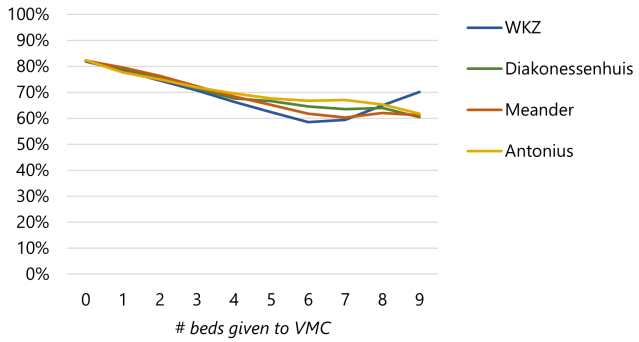


Figure 5.9: Virtual MC policy - Effect on the occupancy (%) each hospital contributes donates respectively 1 to 9 beds to the VMC

Optimizing the amounts of donated beds

The best objective was found with the setting as presented in Table 5.4. Remarkable is that this setting lets the WKZ not put aside any beds, whereas the other hospitals do so. The Diakonessenhuis has to contribute most beds, namely 11-8 = 3. The optimal size of the VMC is 6 beds. This is more than what is expected in line the two-hospital scenario in which each hospital puts one bed aside.

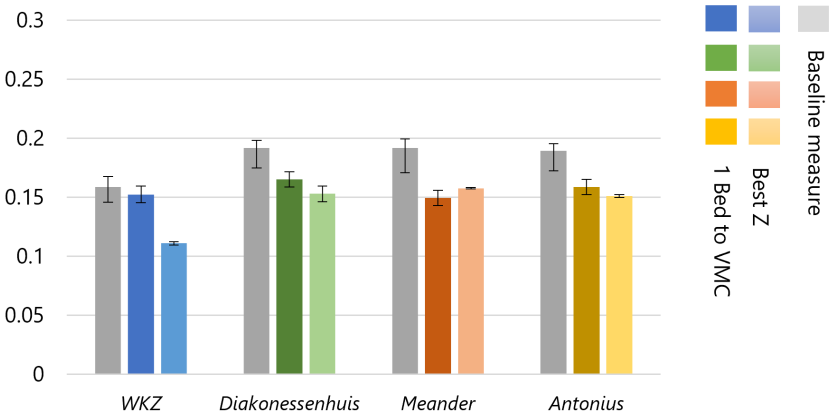
C_1	C_2	C_3	C_4	C_{vmc}
10	9	9	8	7

Table 5.4: Number of beds for the location, and the size of the VMC for the lowest objective

We again delve into each KPI to find what this optimal setting means for each location. For each KPI, we present the outcome in the baseline measure, the configuration in which each location puts one bed aside and the configuration with the best objective. The second factor is included since it illustrates what would be consistent with the CTMC's recommendation for the ideal arrangement of bed distributions.

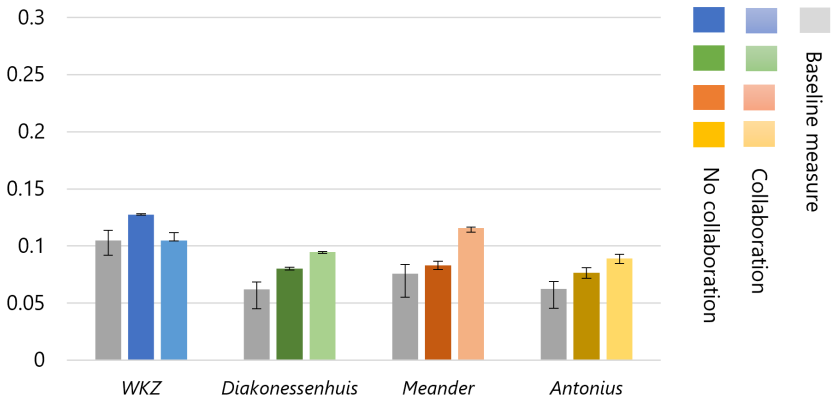
The analysis on the blocking probability for the VMC policy, as shown in Figure 5.10, reveils that this optimal setting is beneficial for all hospitals in terms of P_i^B. When we compare the effect of the best configuration, and the configuration of donating one bed, we observe a decrease in P_i^B for three hospitals (the WKZ, Diakonessenhuis and Antonius) but the Figure shows a slight increase to the Meander hospital as well, although this blocking probability is still significant lower than they currently experience. The CI show that the effect of the VMC on P_i^B result is significant for all locations. A decline of 5 p.p. implies that yearly, 40 patients can be treated more in the WKZ. The 4 p.p. decline in peripheral locations implies that they can treat 20 patients more per year.

Figure 5.10: Virtual MC policy - DES - Blocking probability (%) in the current, one-donated-bed and best setting



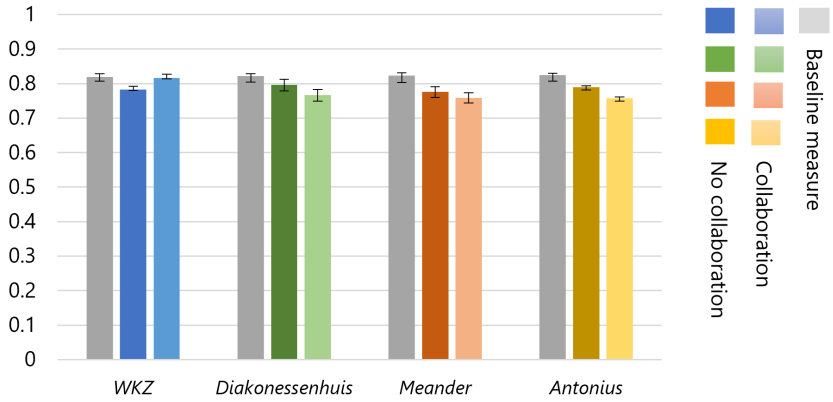
Since the WKZ does not supply beds to the VMC, it is reasonable to assume that the WKZ will treat the majority of patients from its own region. The Diakonessenhuis, on the other hand, would have to give up the majority of the beds, which would force them to utilize the VMC's capacity more, lowering the rejection rate for patients from their region. It has been determined that this is the best setting for the collective probability of refusing and having overbeds. It is anticipated that the chance of overbeds may demonstrate an increase due to that capacity is taken away from the locations so they will have to make overbeds in more percent of the time than they do in the baseline. This is the consequence of the fact that each hospital reserves beds, so less beds are left for their own catchment area of patients. The more capacity that is put aside, the sooner their capacity limit is reached so the sooner they would also have overbeds. The increment is the highest for the Diakonessenhuis (3 p.p which correspond to 11 days per year) and the Meander (4 p.p or 14 days per year). As expected, the WKZ does not experience more overbeds when the VMC is introduced.

Figure 5.11: Virtual MC policy - DES - Fraction of overbed-days (%) in the current, one-donated-bed and best setting



Finally, Figure 5.11 shows a slight reduction of 2 p.p. in occupation for all peripheral locations in case the optimal setting for the VMC is used. This decrease is significant for the Antonius, but not for the Meander and Diakonessenhuis. The WKZ will experience no difference in occupation. This implies that the VMC does not have any worse consequences in terms of occupation.

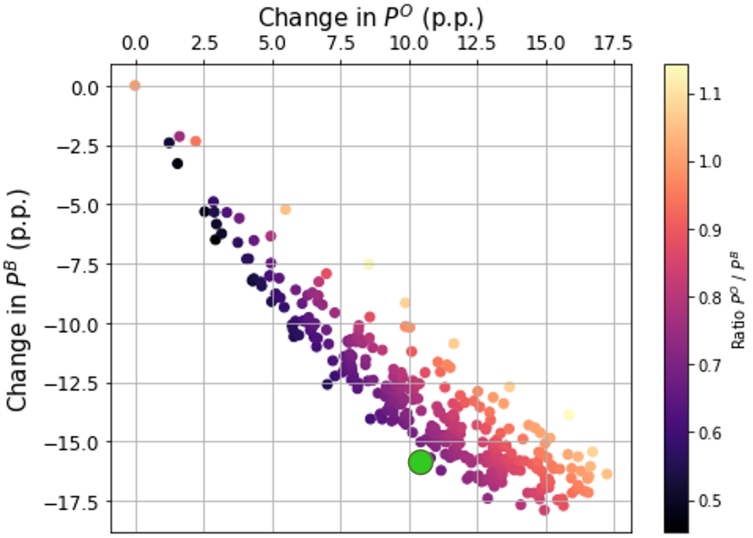
Figure 5.12: Virtual MC policy - DES - Occupation (% occupied beds as a fraction of available beds) in the current, one-donated-bed and best setting



So far we saw the effect on the introduction of the VMC on three KPIs. We have noticed that a reduction of blocking probabilities comes at the cost of increasing time spent in overbeds. To test if the capacity settings from Table 5.4 are capable of striking a proper balance, we check if the solution from this settings in terms of P^B and P^O is on the Pareto frontier.

Figure 5.13 shows, for all evaluated capacity settings, the absolute differences from this capacity setting towards the baseline measure, in P^B and P^O. The points' colors represent the ratio in absolute change in P^B to the absolute change in P^O. The darker the points' color, the lower the increase in P^O in relation with the decrease of P^B, so the more beneficial the capacity settings are. So, the darker the datapoint, the more closer this point is to the Pareto frontier. Points on this line form a set of all Pareto efficient solutions, which allows us to make proper trade-off in case of two KPIs. The points on the X=Y axis correspond to a ratio of 1, meaning that the change from baseline tow the result from this setting in P^B is equal to the change from baseline to this setting in terms of P^O.

Figure 5.13: Virtual MC policy - DES - Pareto frontier of the collective change in overbed-days (%) against the collective change blocking probability (%), n = 221

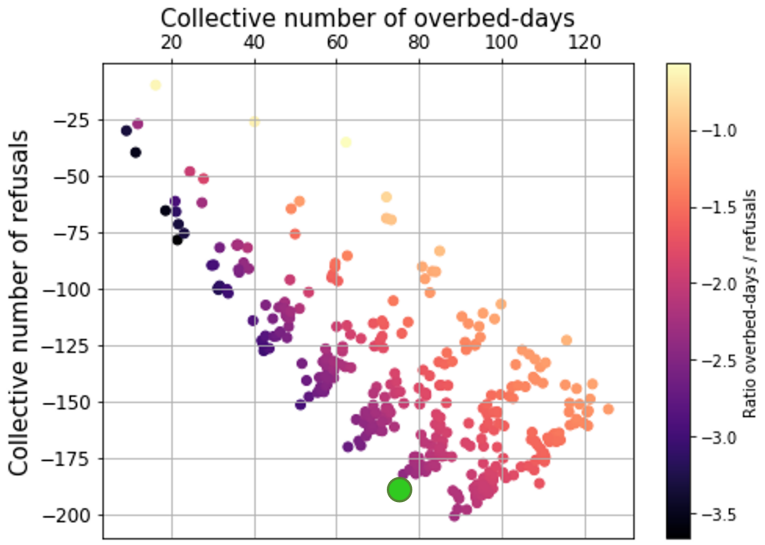


The point at (0,0) indicates our baseline measurement, meaning no difference in P_i^O and P_i^B is present. This dot has the lowest value for the collective fraction of time in overbeds; there are no datapoints where the difference in P_i^O between the intervention and the baseline model, are smaller than 0. This is in line with our expectations because the VMC requires that the locations put aside beds, which makes that the full capacity is reached sooner; having no VMC is then the best when optimizing for overbeds soleley.

In the optimal settings from Table 5.4, P^B decreased from 73% to 57% at the cost of an increase of P_i^O from 31% to 40%. Figure 5.13 shows this point in green. For this point, the ratio of decrease in P_i^B to increase of P_i^O is 0.5625. There is one point however, which ratio is smaller; this is the point where P_i^B drops from 73% to 70.58% and P_i^O increases from 31% to 32.25%, which yields a ratio of 0.5177. This points corresponds to the setting where only the Diakonessenhuis donates 1 bed to the VMC; the other locations keep their original capacity. The reason that our model did not select this solution, is that the overall impact in Z is very small due to that the changes in P^B and P^O are very small.

A dilemma with Figure 5.13 is that it shows the the difference to the current situation in terms of probabilities but not in terms of acutal refused patients and overbed-days. The first viewpoint is easier to use in the model, but it is hard for decision-makers to get an idea of the true meaning of these probabilities. The second viewpoint is more understandable for the management of the neonatal ward. Our objective is determined based on the lowest value for the probabilities, $P^B + P^O$ but we also want to express how good this objective is, from the viewpoint of tangible data and to check if we find a better solution when looking from this last viewpoint. Figure 5.14 shows the increase in overbed-days against the decrease in refusals for two years time. The points correspond to a solution and the green dot indicates our selected solution. From this Figure we notice that our selected solutions is a proper trade-off, when you consider one overbed-day as important as one refusal. This Figure supports the management to make a proper trade-off in case they decided to implement the VMC policy.

Figure 5.14: Virtual MC policy - DES - Pareto frontier of the overbed-days against the number of refusals, n = 221, data for a timeframe of 2 years



The optimal approach features 175 fewer refusals as opposed to 75 more days spent in bed. Our chosen solution is not the best when you evaluate 1 refusal to 1 day in an overbed. The Diakonessenhuis setting, which donates 1 bed (and other locations none), results in 76 fewer refusals compared to 21 more overbed-days, yielding the greatest relative profit in refusals.

5.3 Threshold policy

This section presents the results from the threshold policy. First, we executed greedy constructive heuristic. To this extent, we performed 100 experiments, with each a new random set of thresholds in the range that is given in Table 3.9. After 100 runs, we executed our greedy optimization approach. Figure 5.15 shows the progress of Z against the run number. We started this greedy optimization with the best set of thresholds that was provided from greedy constructive heuristic. The black line shows the average of Z with a window of 5. After roughtly 50 runs, Z stabilizes and no improvement to Z was noticed. To ensure that we exploited the promising areas properly, we continued executing another 50 runs. After this procedure, we found the best objective with a given set of thresholds as presented in the second row of Table 5.5. However, this setting caused the occupation of the Diakonessenhuis to increase to 94 %. Therefore we declined this setting and accepted the second best setting which is indicated in the last row in Table 5.5. In the second best setting, all occupation rates are below 85 %.

	$r_{1,1}$	$r_{1,3}$	$r_{2,1}$	$r_{2,3}$	$r_{3,1}$	$r_{3,3}$	$r_{4,1}$	$r_{4,3}$
After 100 random runs	8	9	9	11	11	9	11	8
Best reservation setting after greedy optimization	9	9	10	10	10	9	10	9
Second best reservation setting after greedy optimization	8	9	9	9	10	8	10	9

Table 5.5: Configurations for every $r_{i,t}$ from the 100 random runs, with the best objective

The total objective from this the best setting that was accepted, had a 19 p.p. reduction in Z compared to the baseline measure.

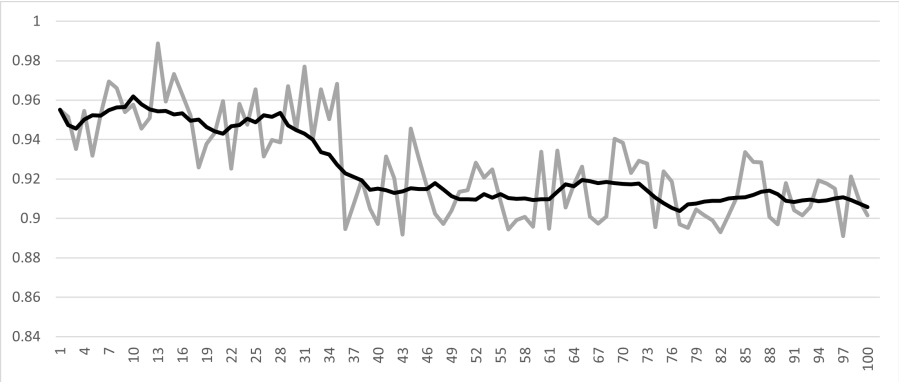


Figure 5.15: Progression of Z in the greedy improvement heuristic

Figure 5.16 presents the results of the blocking probability for each location from the setting that resulted in best Z. as well as the results of the baseline measure that represent the current situation at the neonatal wards. The different colours indicate the different locations. The dark colours represent the baseline measure and the light colours represent the outcomes for the best setting. The error bar represents the standard deviation of the blocking probability from the experiments with that set of thresholds. From this Figure we conclude that the best setting from the 100 runs does result in a lower blocking probability for all peripheral locations. The biggest decline is noticed in the Diakonessenhuis with 4 p.p which is equivalent to 20 patients per year. The results are significant for the peripheral locations but not for the WKZ because the mean of P_1^B of best setting lies within the CI of the baseline.

Figure 5.16: Threshold policy - Blocking probability (%) in the baseline and best setting

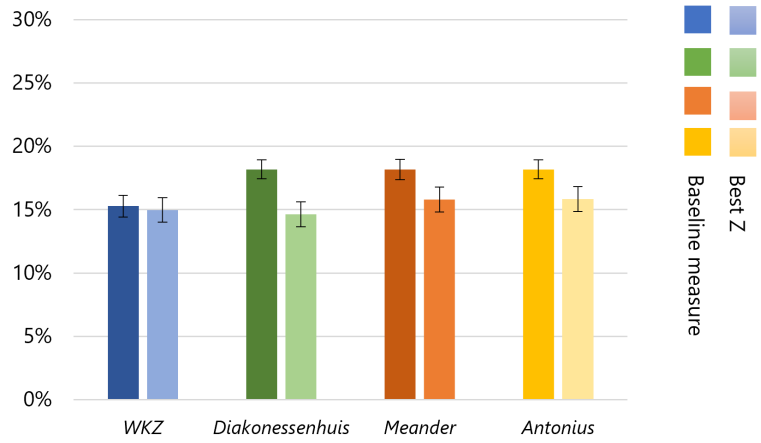
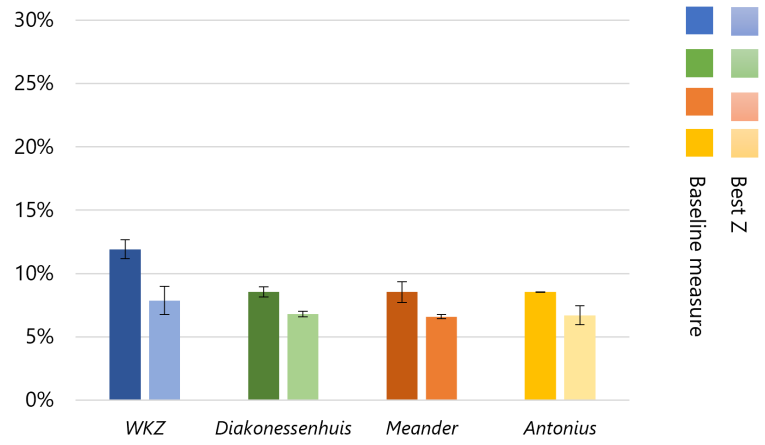


Figure 5.17 presents the fraction of time that the hospitals will to deal with overbeds in the baseline and best settings. A decline of 4 p.p., and 2 p.p., were respectively found for the WKZ and the peripheral hospitals. For the WKZ this implies 15 overbed-days lesser (rounded to two weeks), for the peripheral hospitals 7 overbed-days (or one week).

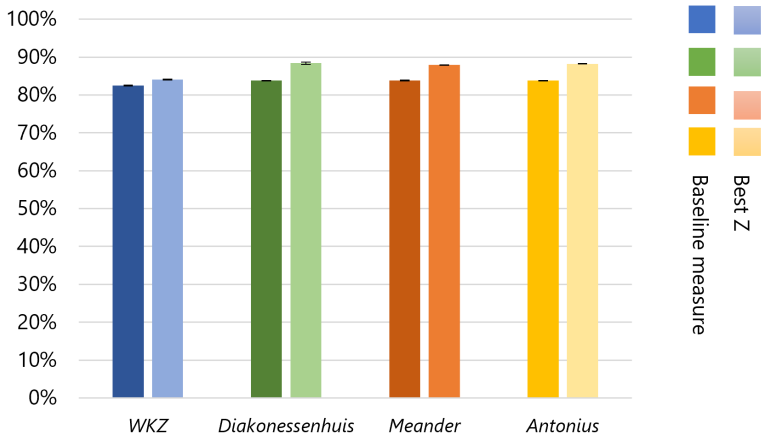
We recall that the blocking probabilities relate to the patients from the hospitals' catchment area, whereas the overbeds relate to the hospital itself. Comparing outcomes for P_1^O and P_1^B we notice that patients from the WKZ' catchment area are relatively most likely to admit to their own hospital. The lower $r_{i,t}$, the sooner you are likely to reject patients or turn overbeds. It is then understandable that the WKZ is relatively least often overburdened.

Figure 5.17: Threshold policy - Fraction of overbed-days in the baseline and best setting



Finally, Figure 5.18 presents the results for the occupation in the baseline setting and the setting with the best Z from 100 experiments. We observed a small increase of 1 p.p. for the WKZ and 3 p.p. for the peripheral locations; the occupation for the peripheral hospitals is 88 % but we have seen that the occupation in the baseline measure is lower than the actual data so we assume that an occupation of 88 % as a result from this policy does not have consequences for patient safety.

Figure 5.18: Threshold policy - Occupation (number of occupied beds as a percentage of the capacity) in the baseline and best setting



In conclusion, using the threshold values shown in Table 5.5 yields the findings shown below. Jointly, across all hospitals, the chance of refusal falls by a total of 9 p.p., and the probability of overbeds drops collectively by 8 p.p., but at the cost of a small increase of the occupation rates. All results are significant. Compared to the VMC policy, there is no need to outweigh the increase in overbeds against the decline refusal probability because the threshold policy causes P^O to decline at points where P^B declines as well. We are still interested in the most promising solution and to find this we plot the difference in baseline and optimal settings. The points in Figure 5.19 represent the outcome of the change in collective P_i^O to the baseline, and the corresponding change n collective P_i^B to the baseline measure. In this Figure we neglected the sets of thresholds that yielded an increase in P_i^O and/or P_i^B. The more the collective decrease of P_i^B and P_i^B, the darker the dot in this scatter plot. This is a different representation of the dots' color than the colors from the Pareto chart in Figure

5.13. Contrary to the VMC policy, we are here are not interested in the ratio because we know there are points where both P_i^B and P_i^B are decreasing compared to the baseline.

The black dots form the Pareto efficient frontier, which consists of all points that Z that were lowest from the baseline and the green dot indicates the solution that was found with our model. It can be seen that our selected solution is on the Pareto efficient Frontier, meaning this is an optimal solution amongst all optimal solutions in minimizing Z.

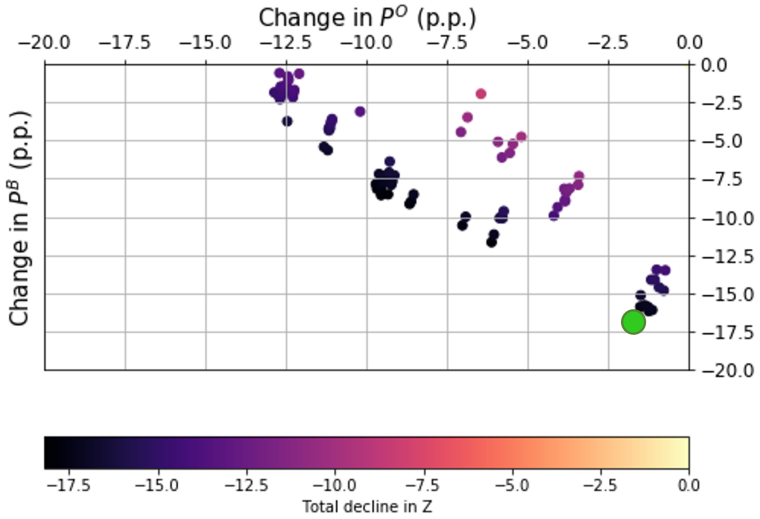


Figure 5.19: Treshold policy - Pareto frontier of the set of solutions provided by the greedy improvement heuristic (n=66). Each point presents the P_i^O with corresponding P_i^B

Figure 5.19 presents a situation in that it depicts the difference from the existing situation in terms of probability but excludes actual denied patients and overbed days. The first point of view is simpler to include into the model, but it is difficult for decision-makers to understand the full significance of these possibilities. The administration of the neonatal ward can directly benefit from the second point of view which is a translation of the probabilities to actual numbers of refused patients and overbed-days. Our goal is set based on the lowest sum of the probabilities, $P^B + P^O$, but there is also a need to investigate whether our selected optimal set of thresholds is most beneficial in terms of actual patient refusals and overbed-days. The green dot indicates our solution, which yields in 70 lesser overbed-days and 90 patients less refused in two years time. We find that our selected solution, out of all the optimal ones, strikes a good balance between patient profits (by reducing blocking probability) and nursing profits (by means of lesser overbeds). For instance, a different solution is provided by -180 fewer refusals and -12 fewer overbed days, which is more promising than our selected solution for the overall image, but this option is not very advantageous from the viewpoint of the nurse because having overbeds is a burden.

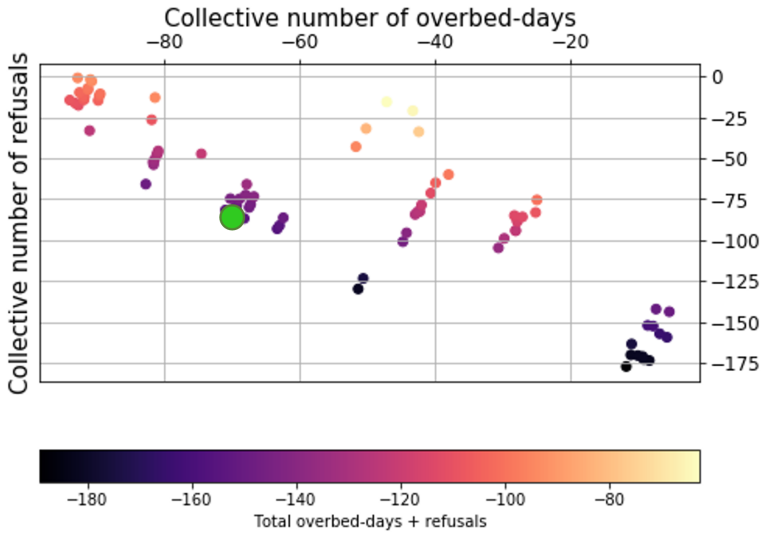


Figure 5.20: Treshold policy - Pareto frontier of the set of solutions provided by the greedy improvement heuristic (n=66, two years time). Each point presents the P_i^O with corresponding P_i^B

From the understandable standpoint, it appears that our solution—indicated by the green dot—is not the ideal one when you take into account that one refusal has the same effect as one day spent in overbed. If you were to optimize for this, the best reservation settings are quite close to our setting after greedy optimisation. This is given in Table 5.6.

	$r_{1,1}$	$r_{1,3}$	$r_{2,1}$	$r_{2,3}$	$r_{3,1}$	$r_{3,3}$	$r_{4,1}$	$r_{4,3}$
Second best setting after greedy optimization	8	9	9	9	10	8	10	9
Setting when you optimize for similar overbed-days and refusals	8	9	10	1-	10	9	9	9

Table 5.6: Configurations for every $r_{i,t}$ from the 100 random runs, with the best objective

5.4 Sensitivity analysis

After running the experiments, the most promising settings will be selected for a sensitivity analysis on certain input parameters. This is done to test the robustness of the model in case estimations of input parameters are not representative for real-life. The input parameters chosen to include in the sensitivity analysis are the arrival rate (λ and the LoS (μ). The experiments selected will be run with a 2.5, 5 and 10 % decrease and a 2.5, 5 and 10 % increase compared to the values used in the experiments. We look at the effect of an increase on the collective chance of refusal and the chance of over-beds.

We compare the behavior of the base model (which mimics the current situation) with that of the optimal model following a change of λ or μ. It seems obvious that the performance of the optimum model degrades with a change in λ or μ, which is why we analyze the behaviour of both the optimal model and the base model. To determine whether our best solution performs better under different conditions other than the current situation does, we want to compare this performance to the same change in lambda or mu at the base measurement.

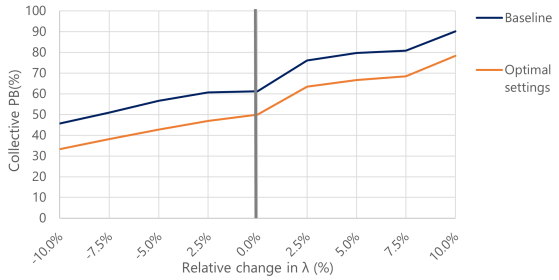
5.4.1 VMC-DES

Figure 5.21 presents the performance of the most promising setting and the baseline model with a change in λ. The x-axis depicts the change in input parameter and the y-axis presents λ. Shown in

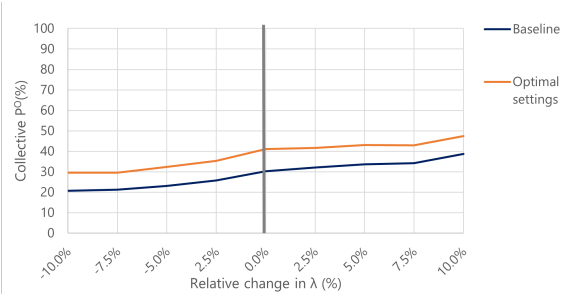
Figure 5.21b is that an change in λ in the model with the best settings ("Best Z") compared to a similar change of the same λ in the base model, the model with the VMC outperfoms the current situation in terms of P^B. To illustrate an example, A 7.5 % decrease in λ in the current situation would yield a similar result as compared to zero decrease in λ for the model with a VMC.

On an increase and decrease on λ we observe that the best model is performing worse than the baseline model does when it comes to time spent in overbeds as depicted in Figure 5.21a. With a 10% increase or decrease in λ the model best model is still performing better than the baseline model would in the current settings. Given that the VMC affects the overbed proportion, it does seem reasonable that changing the input parameters would have a negative impact in the VMC scenario as opposed to the baseline scenario.

Figure 5.21: VMC policy - sensitivity analysis on the number of patients



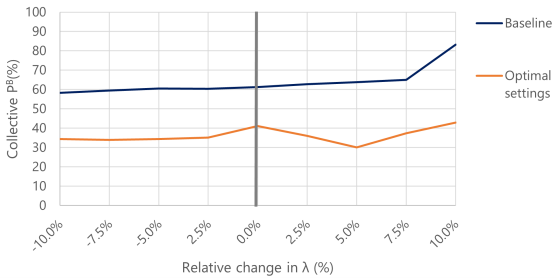
(a) Joint blocking probability (%) to a change in λ



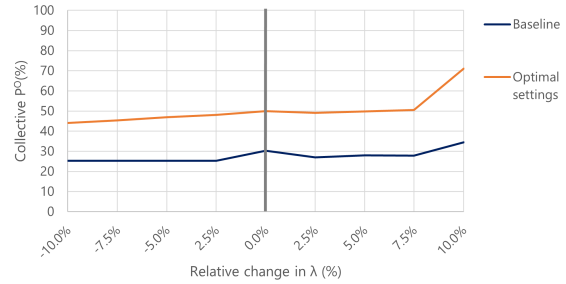
(b) Joint fraction of overbed-days (%) as percentage to a change in λ

Figure 5.22 shows the sensitivity of the VMC model to the baseline model to a change in μ. In terms of the VMC's sensitivity to a change in μ, we see that the VMC continues to perform better in P^B while the same adjustment in terms of P^O performs worse. Given that the VMC affects the overbed proportion, it does seem reasonable that changing the input parameters would have a negative impact in the VMC scenario as opposed to the baseline scenario.

Figure 5.22: VMC policy - sensitivity analysis on the length of stay



(a) Joint blocking probability (%) to a change in μ



(b) Joint fraction of time in overbed (%) to a change in μ

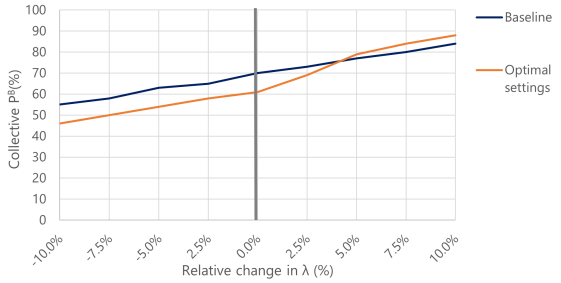
Sumarized, the VMC model yields a better performance in P^B when λ and μ are changed, than the current situation would. This is contrary for P^O, where the current situation outperforms the model with the VMC.

5.4.2 Threshold-CTMC

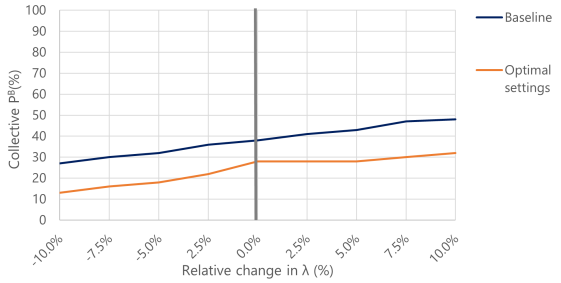
In Figure 5.23a we observe that when a relative change in λ occurs, the baseline model mostly performs better, until λ increases more than 5 %.

Figure 5.23b presents the relative difference in collective overbed probability on both models, by an increasing and decreasing λ. So, contrary to the VMC, when a change in λ the threshold policy outperforms the current situation.

Figure 5.23: Threshold policy - sensitivity analysis on the length of stay



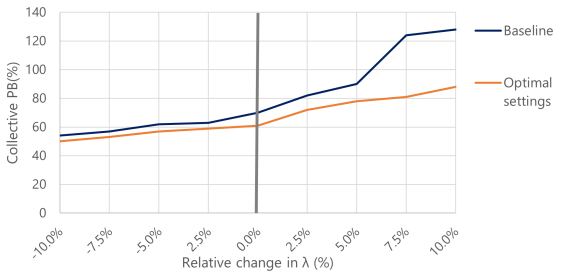
(a) Joint blocking probability (%) to a change in λ



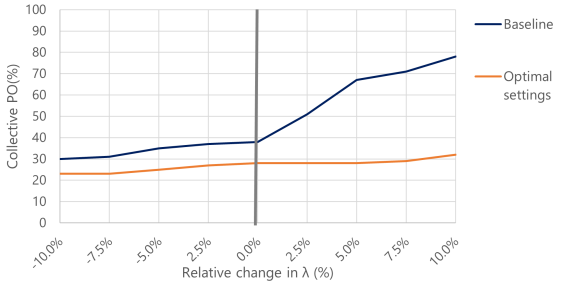
(b) Joint fraction of overbed-days (%) to a change in λ

Lastly we evaluate the effect of a change in μ on the P^B and P^O up to an change in the arrival rate of 10%. The threshold policy outperforms the baseline for both in terms of P^B and P^O. Regarding P^O the threshold policy shows a better performance than the baseline measure with a 10 % decline in P^O does. When μ grows, the baseline measure's P^O performance differs substantially from the ideal conditions when threshold is being used.

Figure 5.24: Threshold policy - sensitivity analysis on the length of stay



(a) Joint blocking probability (%) as percentage of the change in μ



(b) Joint fraction of time in overbed (%) as percentage of the change in μ

5.5 Conclusion

In this Chapter we performed three analysis: a baseline measure to validate the models, an analysis on the three models (VMC-CTMC, VMC-DES and Treshold-CTMC) and a sensitivity analysis. This Chapter contributed to the fourth research goal, that is *to show the practical applicability of the developed model in a case study of the WKZ and its regional hospitals*. We demonstrated that the virtual MC does not cover the practical applicability for two reasons. First, the patient service is improved by a lower blocking probability but this is against the cost of requiring more overbed-days. Second, although the joint objective outperforms the current situation, not every hospital does profit

on every objective and this last subject is very important when it comes to motivating other hospitals to apply the policy. We saw that the CTMC model for the VMC policy gives more extreme results. So, although P^B decreased considerably more in the CTMC approach than in the DES approach, the P^O is remarkably increased. The CTMC only has two hospitals, which helps to explain this. When compared to four hospitals, the impact of reducing the VMC's capacity is larger for two hospitals.

We saw that the threshold policy yields promising results which are not only collectively good but also on the individual location level. Table 5.7 presents an overview of the profits in three models in terms of P^B and P^O. Furthermore, we summarise our findings per policy per solution approach briefly below.

Table 5.7: Overview results for all policies - in tangible data. All results are compared to the current situation in which no policy exists.

Policy type	**Conclusions on interventions**			
	P^B		P^O	
	1 bed aside	best	1 bed aside	best
VMC - CTMC	-132 patients/year	-138 patients/year	+ 78 overbed-days/year	+104 overbed-days/year
VMC - DES	-49 patients/year	-88 patients/year	+24 overbed-days/year	+38 overbed-days/year
TR - CTMC	-	-45 patients/year	-	- 35 overbed-days/year

5.5.1 Baseline measurement

A similar performance among the three models was found. Therefore we conclude that the models perform equal under no intervention.

5.5.2 VMC - CTMC

Although the network benefits as a whole, there are differences per location per objective. This effect is less pronounced in the DES; there, the differences between the locations per objective are smaller. This is due to the fact that a VMC in a network with two locations has a major impact on the capacity of these two locations. With a greater number of locations, there are more possible settings that the VMC can take advantage of.

5.5.3 VMC - DES

When you compare the introduction of a VMC with the current situation, it will always turn out that the current situation offers the least fraction of time that hospitals spend in overbeds, because all capacity is available for every patient type. The introduction of the VMC causes that capacity is taken away from type 2 patients but also opens possibilities for other patient types to experience a lower probability of being refused.

When each location contributes a similar amount of beds to the VMC, the network has to deal with more overbeds which is a logical consequence because less capacity is available for type 2 patients which do cause more overbeds. This effect however is small when each location puts aside just one bed. Thereby, it is very beneficial to donate one bed in terms of blocking probability.

An even more optimal scenario, in which each hospital contributes just a small amount of beds varying in a range from 1 to 3, shows that the network is able to treat 1.03 patients per day more. This is however at the cost of turning more overbeds; 33 more days per year, the network has to operate in overbeds.

5.5.4 TR - CTMC

The optimal setting showed that excluding some patients in a range from one to 3 beds result in lesser refusals and lesser time in overbeds. To be more specific, a decline of 29 days per year lesser overbeds and an increased number of 4 admissions per week was noticed in the whole network when implementing the thresholds as presented in Table 5.5. Furthermore, the network is able to treat 4 patients per week more.

5.5.5 Sensitivity analysis

The sixth research objective is to demonstrate how sensitive the policies are to different input factors. In order to bring this research into practice, it is important to show how the models behave in terms of P^B and P^O because of the following reason. The models are developed using fixed λ and μ values. It is essential that the proposed policies continue to function well in the scenario that λ and/or μ deviate from their distributional basis since the real value might deviate from it (as shown in Appendix B). We showed that, in terms of P^B, the policies outperform the current situation in all cases for λ and μ. However, in terms of overbed-days, the current situation manages changes in input parameters better than the proposed models do so.

Chapter 6

Conclusions

In this Chapter, we present the conclusions of our research. This research was conducted to investigate cooperation possibilities for neonatal care locations. Due to many functions that the WKZ's MC location fulfills, in combination with an increase in demand as well as decrease of resources, the MC is often short of capacity, resulting in refusing patients or admitting them by overbeds. The challenge goes beyond the hospitals own locations, therefore, the managerial goal of the research was,
to establish strategies for facilitating and improving capacity sharing amongst neonatal departments at hospitals in the Utrecht area

We analysed the current situation regarding capacity in Chapter 1, where we showed that the MC of the WKZ has a broad scope of arriving patient types due to the many duties it performs. Chapter 2 showed us that the overflow to peripheral hospitals by means of pooling in the neonatal context is not yet discussed in the literature and two promising models, which were developed to manage overflow in ICU wards, are of great interest to adapt to the context of neonatal care. A mathematical formulation of the TPTM, as well as solution methods for two policies to the TPTM, were discussed in Chapter 3. Chapter 4 provided input data to the case study, and Chapter 5 showed the results of the policies applied to the case study. In this Chapter we provide our conclusions from this research. First, in Section 6.1, we present the scientific contribution. Second, in Section 6.2, we present the practical contribution of this research.

6.1 Scientific contribution

We are the first to apply the TPTM to the context of neonatal care. For the virtual MC policy we proposed solution methods: a numerical solution using Continuous-Time Markov Chains (CTMC) and a Discrete Event Simulation (DES). For the threshold policy, we developed a CTMC as well, in which we used ED as a solution method. A baseline measurement shows that all three models perform equally and show similar outcomes with the current situation that the neonatal ward of the WKZ faces. Sensitivity analysis reveils that the threshold policy is more sensitive to changes, which is caused by the fact that the optimal set of thresholds depends from the locations.

6.2 Practical contribution

This research considered both the KPI per location, as well as the The KPIs however varied per hospital and there is need to ensure that all locations profit from the collaboration.

Although the performance of the VMC is not satisfactory, this still opens possibilities of lowering blocking probabilities if the management would opt to optimize for this objective.

The second policy applied in this research, is the threshold policy. In this policy, each location puts a limitation on the number of patients to accept from each type. Compared to the virtual MC policy, this policy has shown better performance. This is mainly due to a that the policy is more specific to each patient type, instead of limiting all patient types which is the case in the virtual MC policy.

To conclude, the results of this study provide a quantitative basis to support the decision making in a network of hospitals with critically overloaded neonatal care wards. The study provides guidelines for cooperative solutions to tackle the current situation in neonatal care. Among capacity managers there is general saying that "*a little flexibility goes a long way*". Our numerical results have shown that collaborating, by means of small threshold reservation settings, yields significant lower blocking probabilities and less time that the hospitals have to spend having overbeds.

Chapter 7

Limitations and further research

This chapter presents the limitations of the research and related trends in healthcare to which this research relates. Furthermore, we opt for further research ideas. First we present the research limitations in Section 7.1 which concentrates around a lack of data and the model limitations to capture the actual systems' behaviour. Second, in Section 7.2 we compare our results with the results that are known from literature. Then, in Section 7.3, we debate future work opportunities. Finally, Section 7.4 provides suggestions to implement this work into practice.

7.1 Limitations to this research

The research presented in this thesis is subject to several limitations, most of which are centered around a lack of data or uncertainty in the data regarding patient arrivals or length of stays, as well as available capacity from the locations. There are many barriers to face when implementing this research into practice. These barriers include management issues, as well as legal and monetary issues. The cost structure of neonatal wards is complex due to the severity of care. Therefore, legal, as well as monetary issues are also outside the scope of this research.

7.1.1 Lack of data

The arrival rate and length of stay are necessary for the selection of proper model settings. Estimations on the true arrival rate were not known for two out of four hospitals from the case study. To fine-tune the threshold reservation settings it is important that the true on λ and μ are known for every participating location. Although the sensitivity analysis has shown that the threshold policy in most cases outperforms the current situation, we desire that the models' parameters reflect the true population parameters as adequate as possible.

7.1.2 Model limitations in capturing the actual systems' behaviour

There is a number of neonatology characteristics related to behavior of the wards, that are not captured properly by this model. This can cause the model findings to diverge from what actually occurs when the proposed reservation settings are implemented.

The first limitation concerns the available capacity. In this research we assumed a fixed capacity for a time period of two years. Figure 7.1 shows the available capacity from 2020 and 2021 for the MC ward of the WKZ. The dark line indicates the average value per month and the dotted lines indicate the standard deviation around this average. This capacity is the result of factors that include the number of physically available beds, the number of staff, the personnel formation ratios, and the acuity of care required by the patients admitted at that time. For an extensive work related to acuity measurements,

that may serve as an addition to this research, we refer to the work of Hoek (2015). Besides those factors mentioned there are are a variety of immeasurable factors that influence capacity as well; these factors have to do with a personal experience and the willingness of health care workers, as well as patients' families, which may influence the available capacity as well. We assume that the realized available capacity is a product of all the aforementioned factors. Figure 7.1 shows this amount is a reasonable amount to assume.

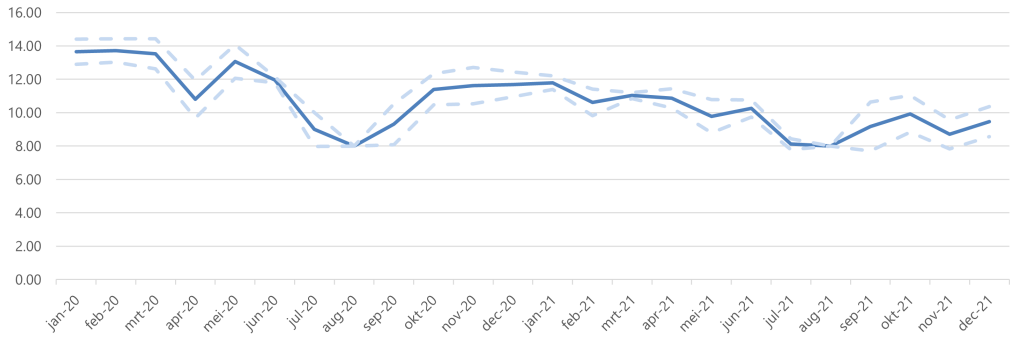


Figure 7.1: Available capacity (average and standard deviation) from the WKZ's MC ward - 2020 and 2021

A phenomenon that argues why the results may underestimate the true effect of the interventions, is the following. The induction of labor is often delayed until a patient has been discharged from the hospital. After the discharge, the new patient will immediately be admitted, resulting in zero inter-arrival time. In our distribution fitting of the IAT of the WKZ, as shown in Appendix B, we observed a peak in the smallest number of IAT in hours, which is the consequence of this phenomenon. This peak causes the IAT distribution to shift to the left, and coherently, the estimation on the arrival rate (expressed in arrivals per time unit) to the right, meaning that the model would overestimate the number of arrivals. When the models are based on a higher value for λ than λ actually is, the results may even be more promising in reality. We refer to the scenario when a patient is placed "on hold" as it occurs when they are ready to be transferred but the new destination is not yet accessible. There is no data data concerning the patients placed "on hold" so we are not able to examine the true effect of this phenomenon on λ and correct λ for this. We recommend the neonatal ward to start keeping track of patients that are "on hold". This would not only increase the accuracy of our results but also reinforce the perception among doctors that often patients are ready for transportation but are unable to leave the hospital because the next location is not yet ready to admit them.

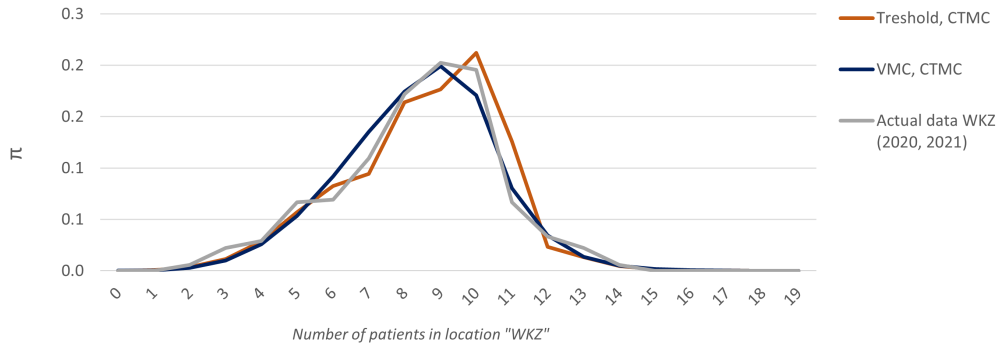
Thirdly, it can be argued that the positive results regarding the decrease in the blocking probability were due to the methods used for analysis. In real-life, it frequently happens that there is not enough capacity to admit an additional patient, but that the patient is placed in an emergency bed nonetheless. Using these beds is not preferred and not a long-term solution, as this results in lower nurse-to-patient ratios and increases the workload for nurses.

7.1.3 Differences in the baseline meausure

In Section 5.1 from Chapter 5 we assumed the three solution methods to the models perform equal under baseline circumstances. We did however notice that the threshold policy showed a lower result for P^B and higher results for P^O than the other models do, and the actual situation of the WKZ does. To clarify the model differences from the baseline measure, we analyze at the distribution over the states in both policies and compare this with the actual distribution from the WKZ. From the WKZ we know the realized capacity from each day of the years 2020 and 2021 since a daily record of the occupied beds is kept. Per quantity, we count the number of days that this quantity is present. Next, we normalize the sum of the days per quantity over the total number of days, such that we get

a percentage of the time that each quantity was presented. Figure 7.2 presents the actual distribution over the states, as well as the distribution in the VMC-CTMC and the threshold-CTMC models. The VMC - DES model is excluded from this overview because this model does not keep track of which time, which amounts of beds was presented.

Figure 7.2: Distribution over states in the baseline measure for the WKZ



The CTMC of the threshold policy concerns all locations individually. This means that there are four one-dimensional Markov chains are made. Since there are fewer states involved compared to the CTMC of the VMC, the accuracy is higher, which explains why the threshold policy is closer to the line of the actual data. The actual data has a lower peak of 1 % in the state with 10 beds, as the threshold policy does. This means that in reality, the hospital does spend 1% lesser of the time in a state with 10 beds than the threshold policy predicts. This may be caused by that the input parameters of the data follow an estimated distribution that does not capture all real datapoints. The state 10 accounted for in obtaining P^O but not in P^B because P^O is the sum of all states where $j \geq C_i$ and P^B are the states where $j > C_i$. There is also a decrease at the point of 12 states of the threshold compared to the actual data and the VMC policy which causes a lower P^B in the baseline measure for the threshold policy compared to the others. State 12 weighs more heavily in P^B than in P^O because the latter probability contains the large value of state 10. This analysis does not fully answer the question, why the models show a difference behaviour in the baseline measure. We attribute this effect to the nature of the models. Although the objectives are calculated in the same way, the structure of the CTMC of the VMC is more complex than the CTMC of the threshold policy. The threshold policy is resolved exactly and the VMC numerically and due to the exact resolution it is expected that the threshold policy will give more accurate results. However, the inputs of both models are derived from a distribution and that distribution deviates from the actual distribution as can be seen in Appendix B because the effect on the collective objective is negligible, we have assumed in this study that the models are the same perform in the baseline. We recommend doing better research into the input values so that the models are better able to accurately represent the current data.

7.1.4 Limitations to the VMC policy

It is to expect that you run more overbed when the VMC is introduced, because you submit your own capacity. You count a state as overbed when the number of patients in that state is greater than your *available capacity*. Then, *your availability capacity* is the number of beds you can operate on, minus the number of beds you hand over to the VMC. Thus, it seems like you operate a lot of overbeds, but this is not actually the case. This implies that, even if our method of modeling makes it appear like the probability of overbeds increases when the VMC is introduced, an overbed would not actually exist if, for example, you give up 9 beds to the VMC and have 1 bed left while there are 10 patients in service at your location. If your capacity is 10, it means that you have the resources to operate 10 beds and, although 9 out of 10 beds are in fact "VMC beds", you still don't run overbeds because

the state that you are in is not exceeding your capacity limit. This limitation may have caused the KPI overbed-fraction to be over estimated. We recommend to adapt our policy to model the VMC not as a seperate location when it comes to overbeds, but as a part of the location itself. A different approach would be to estimate which fraction of all patients in the VMC is of location i and correct the actual patients in location i with this fraction. Note that we did correct for the occupancy for the VMC over the locations by distributing the occupancy of the VMC over the locations and correcting for the magnitude of beds that each location donated to the VMC.

7.1.5 Limitations to the threshold policy

When setting proper thresholds, the threshold that yields most profit for one location depends on the other locations thresholds as well. Table 7.1 presents sets of configurations for every $r_{i,t}$ that resulted an objective that did not differ more than 5 % of the best objective.

$r_{1,1}$	$r_{1,3}$	$r_{2,1}$	$r_{2,3}$	$r_{3,1}$	$r_{3,3}$	$r_{4,1}$	$r_{4,3}$
8	9	9	11	11	9	11	8
9	10	10	11	10	9	9	8
7	9	10	7	11	8	9	9
7	8	9	9	10	11	10	7
8	5	11	9	11	9	11	10

Table 7.1: Configurations for every $r_{i,t}$ from the improvement heuristic within 5% difference with the best objective

From these sets we find that, even if the WKZ would for example introduce a limit of 7 beds to their type 1 patient, it is still important that other locations apply limits which jointly result in a decline of blocking- and overbed probability. Important to bring this research into practice when shifting thresholds, is to take the occupation into account. We have seen that the most promising solution in terms of collective blocking- and overbed probabilities is not a good solution in terms of occupancy, since one of the locations then has to deal with occupancy higher than 90 %.

7.2 Comparison of the results in neonatal care from this research and intensive care from literature

7.2.1 Results of the VMC policy

Litvak et al. (2008) Introduced the virtual ICU policy where the underlying concept is resource sharing by providing mutual overfow. Their goal was to provide sufficient service level for non-overflow calls, in other words, for type 2 and 3 patients that cannot be sent elsewhere. From a patient point of view, we did notice an better service performance due to the introduction of the VMC. The KPI representing the patient point of view is namely the fraction of refusal. For the employees' point of view however, the virtual MC policy did not provide sufficient service levels as Litvak et al. (2008) suggests.

If all available capacity amongst the four hospitals included in this research could be pooled, we find, using Equation 19, that the collective arrival rate, and use estimation on the average LoS, no refusals would not occur at all. Nonetheless, neonatal patients are bound to particular hospitals. For example, the WKZ cannot release all of its capacity because they fulfill a third-line function which other hospitals do not. Hence we are not able to accomplish a zero percent blocking probability. Litvak et al. (2008) is able to achieve the managerial goal of at most 1 % rejections for type 1 patients. They assume that all hospitals have a similar patient stream structure. They assume that the ratios $p_{i,t}$ hold for every location in the network and do not distinguishes academic hospitals from peripheral hospitals in this extend. Furthermore, the ICU has comparable more elective patients than neonatology has.

The neonatal knows a much bigger ratio of type 2 patients, so, it is reasonable that the VMC yields less promising solutions in neonatal care.

7.2.2 Results of the threshold policy

Similar to Chan et al. (2018) we conclude that the threshold policy is more efficient for both the patient and the employees, by allowing an type 1 patient to be assigned to any of the vacant beds in the network as long as none of the vacancy thresholds are violated, than the virtual MC policy is. Since less overbeds contribute towards higher service levels for all patients, we recommend to adapt the threshold policy in neonatal care.

The LoS of elective patients is more predictable and their arrivals can be controlled. The precise knowledge of how many patients of type 3 there are exactly presented in each location may help to decrease the number of cancellations by better planning of elective arrivals and ultimately result in a smaller number of cancellations.

The threshold policy from Chan et al. (2018) suggest that the results of their ICU network are not very sensitive to the patient LoS distribution, as we observed as well.

Chan et al. (2018) recommended to investigate the effect of *maximal* resource sharing. By this they mean to investigate the effect of thresholds such that patients can always admit to any location of the network by setting a threshold value of 0. For neonatal care however, maximal resource sharing is not possible since locations require to serve their own catchment area and closing one of the locations is not a feasible solution. We do however recommend to investigate this effect for other application areas.

A suitable tool is required to quantify the number of IC beds needed for each hospital in different scenarios, taking into account factors like the expected patient volume, the distribution of patient types across hospitals, as well as the percentage of rejected patients and overbeds permitted by management, health insurers, or the government. In addition to providing data on the current capacity, this tool may be extended through being optimized. A quick examination of all possible combinations of the number of beds at each hospital and the number of regional beds is possible due to the CTMC of the threshold policy. This makes it possible to optimize how many beds are distributed among hospitals.

7.3 Future work

This research opens opportunities for both scientific work and practical work. Scientific opportunities are discussed in Section 7.3.1 and practical opportunities in Section 7.3.2.

7.3.1 Scientific work

For now we have chosen fixed sets of reservation threshold settings. An extension of this work could be to introduce dynamic settings. In a dynamic setting, the thresholds should adapt to a response in input parameters. The CTMC compiles fast; therefore, an adaptation in thresholds allows for fast evaluation of the solution quality of that adaptation. In a short time, many threshold values could be evaluated. Our greedy improvement heuristic has shown to be effective in exploiting promising set of threshold reservation settings. The greedy improvement heuristic can then be used to find the most optimal set of threshold given a change in input parameters and thus adapt to the situation quickly. Second, we recommend to use a more powerful computer that allows for calculating the VMC in a numerical way when more than 2 locations are included. We believe that including more locations would yield proportional more benefits.

7.3.2 Practical work

We saw that the suggested models are sensitive to changes in λ and μ. First, we therefore recommend to fine-tune the input parameters to build a model that is more robust. One can think of including seasonality such as holidays and the weather conditions that influence λ and μ, such that our models will find more accurate settings suitable for that situation.

Future work may involve extending the threshold policy model to a network with more than 4 locations. We limited ourselves to a network with four locations. Comparing the VMC's exact model with the DES, we already see an increase in performance when more locations are opened. The Meander is located in Amersfoort. Another hospital that is comparable with the type of care that the WKZ provides, and is relatively close located to the Meander hospital, is the Isala hospital in Zwolle. From the Meander point of view, investigating collaboration possibilities with the Isala may be of interest. But then, the Isala hospital has other hospitals surrounded as well. It would be of great interest to investigate to what extent hospitals should collaborate, because in reality no fixed borders are set.

Another future work suggestion is to check on medical conditions that the patient types differ. If we can relate the medical conditions to the patient types, it becomes easier for the doctor to distinguish patients objectively and to apply proper triage principles which strengthens the implementation of this work into practice.

7.4 Implementation of the research into practice

The concept of pooling is a difficult concept to understand for managers, nurses and doctors. Pooling demand is not always applicable in the context of birth care. Although one might state you assign certain capacity to certain patient groups, in practice, the moment of birth is hard to predict; therefore, the occupation at operational level is hart to predict, making it is difficult to assign patients to beds in time. From the sensitivity analysis of a change in λ we have seen that an increase λ continues to lower P^B and P^O more than the current situation would do, which is beneficial for other wards than the MC. A 10% increase in λ means 162 patients per 2 years can be treated more in the WKZ. This opens possibilities for other wards to let patients move to the MC and hold patients for less time if they have already been treated. If you for example you assume that this concerns NICU patients, you can relieve the NICU with 81 patients/year. However, we lack of data of patients that which should have already been forwarded but did not do so, due to shortages elsewhere.

Type 1 patients who are refused because a bed is not available anywhere in the network, will in practice be sent to a different hospital than the locations included here. So they have to deal with the extra load of type-1 patients from one of these four hospitals. By lowering the refusal rate with the threshold policy, this means that other hospitals outside our case study will also benefit from this policy. An interesting research opportunity would be to investigate the effect of the loads that other hospitals will experience. We refer to the research of (Kanai & Takagi, 2021) that forms a good stating point in modelling the loads of neonatal patients in other wards.

In addition to the above-mentioned problems, there are two trends that are relevant to this research. These trends give more context to the problem and emphasize the added value of this research. First of all, due to a significant amount of time spent on transfer coordination, the management strives to develop an application to map regional bed availability of the post-HC and MC beds. In case of a shortage of beds, the hospitals in the Utrecht region can see each other's capacity, saving the time-consuming process of calling back and forth. On a regional level, the app connects supply and demand for post-IC and MC beds. The development of this application is a contributing factor toward improving the outflow of MC patients and supports the implementation of the thresold reservation

settings.

Second, in line with the concept of pooling, more and more attention is being paid to the practice of integrated care, notably in maternity care. Economies of scale advantages could be gained by collaborating between departments within a healthcare institution, or between healthcare institutions. The Regional Ambulance Care Facility Utrecht, for example, has recently started distributing first- and second-line pregnancies in the region of Utrecht. This coordination used to be executed by hospitals and obstetric clinics itself, but by taking away this task, the clinicians have more time to practice their profession. The high workloads drove the motivation to support regional collaboration. Seeking for more support is of importance to translate this research into practice.

Bibliography

Asaduzzaman, M., Chaussalet, T. J., & Robertson, N. J. (2010). A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research*, *178*(1), 67–76.

Ata, B., & Van Mieghem, J. A. (2009). The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science*, *55*(1), 115–131.

Bamford, D., & Chatziaslan, E. (2009). Healthcare capacity measurement. *International Journal of Productivity and Performance Management*.

Bekker, R., Koole, G., & Roubos, D. (2017). Flexible bed allocations for hospital wards. *Health care management science*, *20*(4), 453–466.

Cattani, K., & Schmidt, G. M. (2005). The pooling principle. *INFORMS Transactions on Education*, *5*(2), 17–24.

Chan, Y.-C., Wong, E. W., Joynt, G., Lai, P., & Zukerman, M. (2018). Overflow models for the admission of intensive care patients. *Health Care Management Science*, *21*(4), 554–572.

Creemers, S., & Lambrecht, M. (2008). Healthcare queueing models. *FBE Research Report KBI_0804*.

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

Franx, G. J., Koole, G., & Pot, A. (2006). Approximating multi-skill blocking systems by hyperexponential decomposition. *Performance Evaluation*, *63*(8), 799–824.

Green, L. (2006). Queueing analysis in healthcare. In *Patient flow: Reducing delay in healthcare delivery* (pp. 281–307). Springer.

Harrison, J. M., & López, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing systems*, *33*(4), 339–368.

Hill, R. (2007). Discrete-event simulation: A first course.

Hoek, W. (2015). Acuity measurement at the neonatal intensive care unit. *Master thesis, University of Twente*.

Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of emergency medicine*, *52*(2), 126–136.

Huang, Y.-L., & Kammerdiner, A. (2013). Reduction of service time variation in patient visit groups using decision tree method for an effective scheduling. *International Journal of Healthcare Technology and Management*, *14*(1-2), 3–21.

Huang, Y.-L., Zuniga, P., & Marcak, J. (2014). A cost-effective urgent care policy to improve patient access in a dynamic scheduled clinic setting. *Journal of the Operational Research Society*, *65*(5), 763–776.

Hulshof, P. J., Boucherie, R. J., Hans, E. W., & Hurink, J. L. (2013). Tactical resource allocation and elective patient admission planning in care processes. *Health care management science*, *16*(2), 152–166.

Kanai, Y., & Takagi, H. (2021). Markov chain analysis for the neonatal inpatient flow in a hospital. *Health Care Management Science*, *24*(1), 92–116.

Keskinocak, P., & Savva, N. (2020). A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. *Manufacturing & Service Operations Management*, *22*(1), 59–72.

Kortbeek, N., Zonderland, M. E., Braaksma, A., Vliegen, I. M., Boucherie, R. J., Litvak, N., & Hans, E. W. (2014). Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance evaluation*, *80*, 5–26.

Laws, C. (1992). Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability*, *24*(3), 699–726.

Litvak, N., Van Rijsbergen, M., Boucherie, R. J., & van Houdenhoven, M. (2008). Managing the overflow of intensive care patients. *European journal of operational research*, *185*(3), 998–1010.

Mahar, S., Bretthauer, K. M., & Salzarulo, P. A. (2011). Locating specialized service capacity in a multi-hospital network. *European Journal of Operational Research*, *212*(3), 596–605.

Ministerie van Volksgezondheid en Welzijn en Sport. (2022). Nieuwe prognose verwachte personeelstekort. https://www.rijksoverheid.nl/documenten/kamerstukken/2022/01/20/kamerbrief-over-nieuwe-prognose-verwachte-personeelstekort

Morris, K. (2021). Reducing neonatal intensive care transports by introducing a regional nurse flex pool. *Master thesis, University of Twente*.

Nandigam, A., Jog, S., Manjunath, D., Nair, J., & Prabhu, B. J. (2019). Sharing within limits: Partial resource pooling in loss systems. *IEEE/ACM Transactions on Networking*, *27*(4), 1305–1318.

Norris, J. R. (1998). *Markov chains*. Cambridge university press.

Ogboenyiya, A. A., Tubbs-Cooley, H. L., Miller, E., Johnson, K., & Bakas, T. (2020). Missed nursing care in pediatric and neonatal care settings: An integrative review. *MCN: The American Journal of Maternal/Child Nursing*, *45*(5), 254–264.

Pehlivan, C., Augusto, V., Xie, X., & Crenn-Hebert, C. (2012). Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. *2012 IEEE International Conference on Automation Science and Engineering (CASE)*, 137–142.

Seneta, E. (1980). Computing the stationary distribution for infinite markov chains. *Linear Algebra and Its Applications*, *34*, 259–267.

Shah, P., Mirea, L., Ng, E., Solimano, A., & Lee, S. (2015). Association of unit size, resource utilization and occupancy with outcomes of preterm infants. *Journal of Perinatology*, *35*(7), 522–529.

Song, H., Tucker, A. L., Graue, R., Moravick, S., & Yang, J. J. (2020). Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science*, *66*(9), 3825–3842.

Tierney, L. T., & Conroy, K. M. (2014). Optimal occupancy in the icu: A literature review. *Australian Critical Care*, *27*(2), 77–84.

UMC Utrecht. (2021). Geboortezorg onder druk. https://www.umcutrecht.nl/nieuws/regionale-geboortezorg-onder-druk

Van Dijk, N., & van der Sluis, E. (2009). Pooling is not the answer. *European Journal of Operational Research*, *197*(1), 415–421.

Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., & Litvak, N. (2012). Efficiency evaluation for pooling resources in health care. *OR spectrum*, *34*(2), 371–390.

van Dijk, B., Schilstra. (2022). On two product form modifications for finite overflow systems. *Annals of operations research*, *310*(2), 519–549.

van Dijk, N. M., & van der Sluis, E. (2008). To pool or not to pool in call centers. *Production and Operations Management*, *17*(3), 296–305.

van Doremalen, R. (2022). Predicting the or capacity to optimize pac scheduling. *Master thesis, University of Twente*.

Vermeulen, I. B., Bohte, S. M., Elkhuizen, S. G., Lameris, H., Bakker, P. J., & La Poutre, H. (2009). Adaptive resource allocation for efficient patient scheduling. *Artificial intelligence in medicine*, *46*(1), 67–80.

Weernink, A. O. (2018). The assessment of pooling intensive care and high care units at the neonatology department of wilhelmina kinderziekenhuis. *Master thesis, University of Twente.*

Williams, J., Dumont, S., Parry-Jones, J., Komenda, I., Griffiths, J., & Knight, V. (2015). Mathematical modelling of patient flows to predict critical care capacity required following the merger of two district general hospitals into one. *Anaesthesia*, *70*(1), 32–40.

Winston, W., & Goldberg, J. (2004). *Operations research: Applications and algorithms.*

Wischik, D., Handley, M., & Braun, M. B. (2008). The resource pooling principle. *ACM SIGCOMM Computer Communication Review*, *38*(5), 47–52.

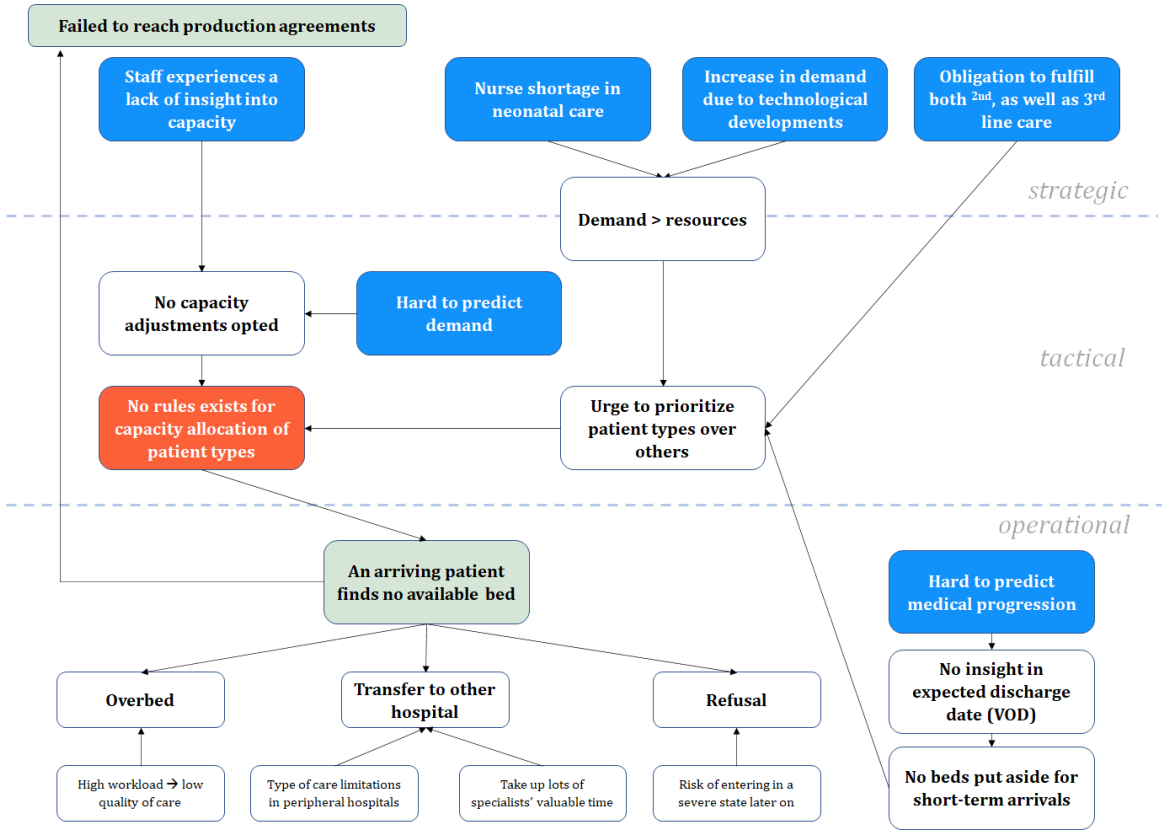
Worthington, D. (1991). Hospital waiting list management models. *Journal of the Operational Research Society*, *42*(10), 833–843.

Zonderland, M. E., & Boucherie, R. J. (2012). Queuing networks in health care systems. In *Handbook of healthcare system scheduling* (pp. 201–243). Springer.

Appendix A

Problem cluster

This Appendix presents the problem cluster on capacity of the neonatal wards in Utrecht. The core problem comes down to the following: an arriving patient does not find an empty bed. This is noticed on operational level.



Appendix B

Distribution fittings to the data

This Appendix presents a justification on the assumptions for the Poisson arrivals and exponential inter arrival times. The actual data is grouped in histograms, where number of bins is calculated as $\sqrt{number of datapoints}$. First, we present the inter arrival times and service times of the MC ward of the WKZ. To assume Poisson arrivals, we test the actual data on exponential inter arrival times. Figure **??** presents the frequency, presented at each bin, and the exponential fitting that corresponds to the bins. The exponential fitting is plotted with a mean of to $\lambda = (730/1612) * 24 = 2.20$ hours. To this extend, we included the refusals to the total arrival rate. We were able to do so, because we have estimations on the (direct and indirect) refusals from in 2020 and 2021 from the WKZ. These numbers are presented in Appendix **??**. The quality of the fitting of the exponential distribution to the actual data is measured in Mean Squared Error (MSE). The MSE for the IAT of the WKZ is 191.

Figure **??** presents the exponential fitting to the LOS of patients in the MC ward of the WKZ. The LoS follows an exponential distribution with mean $\mu = 4.50$ days. The number of datapoints is lesser than the number of datapoints used for the IAT distribution, because we do not know the LoS of the refusals. The exponential fitting is plotted with mean $\mu = 4.50$ hours and has a MSE 675.

Next, we present an elaboration on the estimations on λ and μ for the peripheral hospitals. Since this research only feautures data from the Diakonessenhuis, we limit ourselves to the fittings on data from this hospital only. This data is from 2019, whereas the WKZ data is from 2020 and 2021. According to the management of the Diakonessenhuis, the data of the neonatal ward from 2019 forms a more reliable picture, that is better representing the actual situation, than the data of the next two years did.

Figure **??** shows the fitting of the distribution of inter arrival times to the actual data. We find that the inter arrival times to the Diakonessenhuis follow an exponential distribution with mean $\lambda = 19$ hours. Contrary to the WKZ, the IAT from the Diakonessenhuis does not contain the refusals since this data is unknown. For the λ the arrival rate without the refusals is used. The MSE is 50. Next, by adding the anticipated number of rejects, we scale up the lambda. We can scale this to the size of the Diakonessenhuis as we know that the WKZ rejects 157 patients at 2 years owing to a full obstetrics department. We do not consider those refusals since the Diakonessenhuis lacks a NICU.

Appendix C

Simulation flowchart of the VMC DES

In this Appendix we explain the logic we built into our simulation model. We show two flowcharts, one concerning the arrival of a patient and one the discharge. Figure C.1 shows the arrival process. The blue boxes indicate the counters that are used to determine the objective. When a patient is created, first, it is checked whether the location of the origin of the patient still has capacity left. If so, this patient will be admitted. If not, we check which patient type the patient is. If it is a type 1 patient, and the VMC has still capacity left, the patient will admit to the VMC. If the patient is not a type 1 patient but a type 2, then, it is admitted to the original location. If neither of the above two conditions is true, it automatically means the patient is a type 3 patient, hence is refused. We increase the "nRefusals" counter with 1. At the end of our simulation run, we divide the "nRefusals" counter to our "nPatientsCreated" counter to obtain P_i^b. Furthermore, we divide "nOverbeds" by "nRefusals" to obtain the overbed probability.

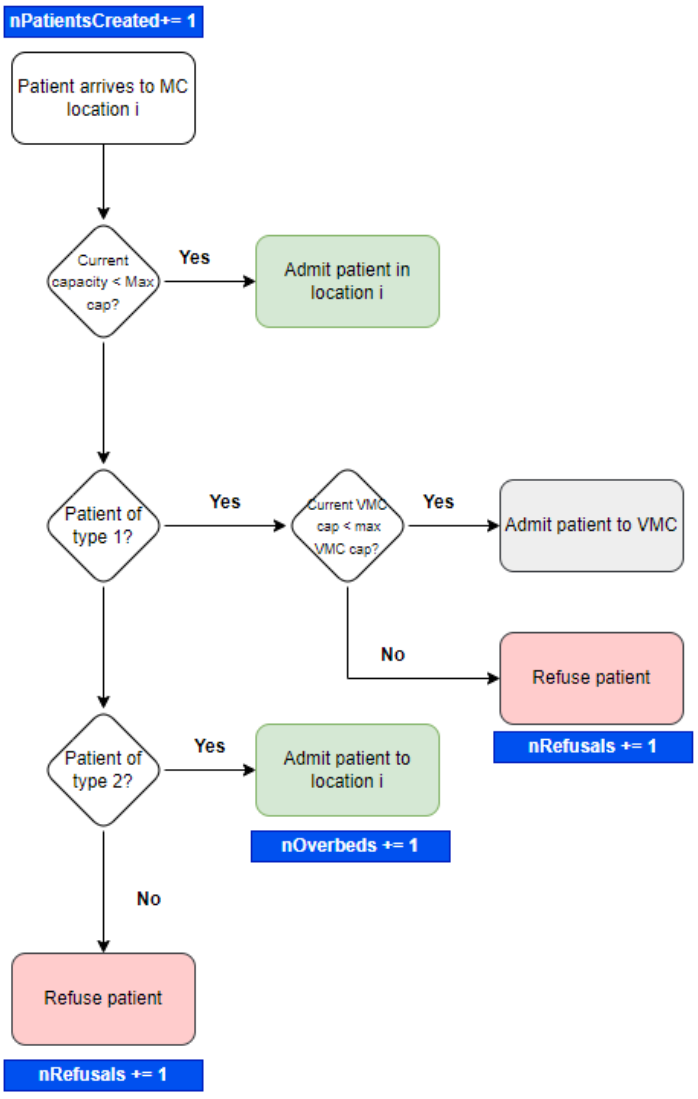


Figure C.1: Flowchart to determine the destination of an arriving patient

The next series of events that trigger a process is the discharge of a patient. If this concerns a type-2 patient, it is checked if the location contains patients in overbeds and if so, the patient is moved from the overbed to a normal bed of location i. This ensures that the overbeds only exists in moments where full capacity exists.

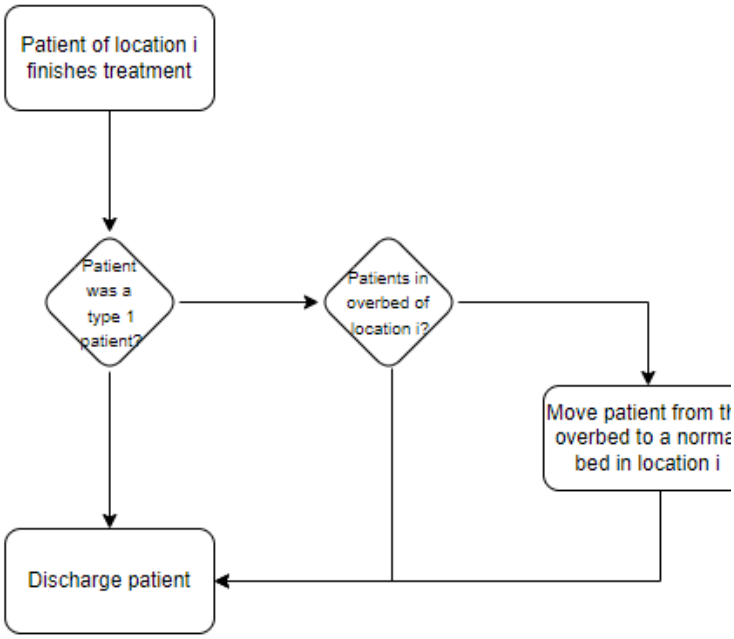


Figure C.2: Flowchart to illustrate the discharge process

Appendix D

The routing policies explained

The routing policy, $\Gamma_{z,n}$, describes in which order external emergency patient from each zone z attempt for admission at other hospitals. Table D.1 describes an example of such a policy. The vertical axis is the current zone, the originating zone the patient is from. On the horizontal axis, the next zone for attempt is displayed.

		Next zone			
		1	**2**	**3**	**4**
Current zone	**1**	1st	2nd	3rd	4th
	2	4th	1st	2nd	3rd
	3	3rd	4th	1st	2nd
	4	2nd	3rd	4th	1st

Table D.1: Example of an routing policy. In this example, patients from a zone z attempt for admission in zone *z+1*

In order to illustrate the working of the policy, suppose the following example, in which we calculate the total load that location 4 receives, x_4. According to the policy, patients from zone 1, will first attempt at their own location, second at $i = 2$, third at $i = 3$ and last at $i = 4$. So, from patients from location 1, their attempt at location 4 is their fourth attempt.

Previously, we defined $a_{i,n}$ as the load of patients from a zone z that have overflowed n times in the network. So, patients from zone 1 attempting at zone 4 are denoted by $a_{1,4}$. Likewise, the hospital from location 4 receives:

- $a_{4,1}$
- $a_{3,2}$
- $a_{2,3}$

We also stated that, in order to obtain the total load to a location,

$$a_{i,0} = \lambda_{i,1}$$
$$a_{i,n} = a_{i,n-1} b_{\Gamma_{i,n-1}}, n > 0$$

Thus, $a_{4,n} = \lambda_{4,1}$, and
$a_{3,2} = a_{3,1} b_3$;
$a_{2,3} = a_{2,2} b_2 = (a_{2,1} b_1) b_2$;
$a_{1,4} = a_{1,3} b_3 = (a_{1,2} b_2) b_3 = ((a_{1,1} b_1) b_2) b_3$

In total, the hospital in zone 4 receives a load from external emergency patients of: $a_{4,1} + a_{3,2} + a_{2,3} + a_{1,4} = a_{3,1} b_3 + (a_{2,1} b_1) b_2 + ((a_{1,1} b_1) b_2) b_3$.

Appendix E

Finding steady state vector based on eigenvectors

This Appendix provides an in-depth explanation on eigenvectors and their relationship with Markov chains. Let $q_{SS'}$ denote the transition rates of state S to state S'. Let m be the total number of possible states. Then, the transition rate Q looks like:

$$A = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,n} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n,1} & q_{n,2} & \cdots & q_{n,n} \end{pmatrix} \tag{20}$$

Now, we introduce a steady state vector v. Initially, v looks like: $v = (v_1, v_2, ..., \ v_n)$. The eigenvector definition states that $Qv = \lambda v$; if $\lambda = 1$, then $vA = v$. This looks like,

$$\begin{aligned} v_1A_{11} + v_2A_{21} + v_3A_{31} &= v_1 \\ v_1A_{12} + v_2A_{22} + v_3A_{32} &= v_2 \\ v_1A_{13} + v_2A_{23} + v_3A_{31} &= v_3 \\ &... \end{aligned} \tag{21}$$

We can write $vA = v$ as $vA - v = 0$, which is similar to $v(A - I) = 0$. The operation "-1" for a square matrix is similar to the operation "square matrix - I" where I is the identity matrix. Furthermore, we need to add the constraint that every element in v sums to 1, so $v_1 + v_2... + v_n = 1$. Then,

$$\begin{aligned} (v_1A_{11} - v_1) + v_2A_{21} + v_3A_{31} &= 0 \\ v_1A_{12} + (v_2A_{22} - v_2) + v_3A_{32} &= 0 \\ &... \\ v_1 + v_2 + v_3 &= 1 \end{aligned} \tag{22}$$

Now we formed a set of equations that we can solve using a linear solver that will return $v = v_1, v_2, ..., v_n$ which is the steady state vector from matrix A.