UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management and Social sciences

Assessment of a Pods System to Reduce Waiting and Throughput Times in an Emergency Department

Aina Goday Verdaguer

M.Sc. Thesis IE&M

September 2019

Supervisory Committee:

University of Twente
Prof. Dr. Ir. E.W. Hans
Dr. D. Demirtas

University of Auckland
Dr. C. Jagtenberg
Dr. M. O'Sullivan
Dr. C. Walker

Assessment of a Pods System to Reduce Waiting and Throughput Times in an Emergency Department

September 2019

Author
Aina Goday Verdaguer
Industrial Engineering and Management
University of Twente

First Internal Supervisor
Prof. Dr. Ir. E.W. Hans
School of Behavioral, Management and Social Sciences
Department IEBIS
University of Twente

Second Internal Supervisor
Dr. D. Demirtas
School of Behavioral, Management and Social Sciences
Department IEBIS
University of Twente

First External Supervisor
Dr. C. Jagtenberg
Department of Engineering Science
University of Auckland

Second External Supervisor
Dr. M. O'Sullivan
Department of Engineering Science
University of Auckland

Third External Supervisor
Dr. C. Walker
Department of Engineering Science
University of Auckland

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

UNIVERSITY OF TWENTE.

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less.

— Marie Curie

To my parents.

To Gerardo.

Management Summary

In the framework of completing the master thesis of Industrial Engineering and Management, the author performed research at the Emergency Department (ED) of Auckland City Hospital (ACH) in New Zealand. This study is built upon the research initiated by scholars from the University of Auckland (UOA).

The ED consists of 8 services: Resus, Monitored, Acutes, Ambulatory Care, Short Stay, Procedure, Consultation and Other. Until December 2018, there was a general pool of physicians where they worked together as one group to treat all services, except for Resus, which had its own dedicated team of two physicians. Nevertheless, the ED decided to unpool the medical staff and to establish teams of physicians and nurses to provide care to specific patient groups in particular geographic areas within the ED. This new system is referred to as the "pods" system, whereas the previous situation was the "no pods" system.

The ED contacted the UoA researchers as they wanted to know the effects on their waiting and throughput times of implementing the pods system. For this purpose, the researchers defined three Key Performance Indicators (KPIs) based on the ED's performance targets: mean triage to sign-on time (time from triage to first physician's assessment), 95th percentile of the total time spent in the ED and fraction of patients leaving the ED within 6 hours. Furthermore, they developed a simulation model, with which they tested different scenarios. Their results showed that both increasing the workforce in the pods system and keeping the original no pods system resulted in a similar performance. Nevertheless, they made suggestions for the model to be improved and for other queuing disciplines like Priority Accumulation (PA) to be investigated. With PA patients accumulate priority as a linear function of their waiting time.

The goal of this thesis is two-fold. First, to improve the simulation model by extending, verifying and validating it. Second, to explore the implications of implementing priority accumulation, the effects on the abovementioned KPIs of the no pods and pods systems and how these two behave when adding extra workforce. This goal is translated into the following research question:

How can priority accumulation, the pods system and an increase in medical staff help improve the waiting and throughput times of the Emergency Department in Auckland City Hospital?

To be able to answer the research question we first conducted a literature review about priority accumulation and the pods system. We then extended the simulation model and developed a PA prototype as one of the patient sorting methods for the simulation model. We also developed a staffing model which is a Mixed Integer Linear Program (MILP) that allocates a given list of

physicians to pods. This allocation is inputted to the simulation model, with which experiments are run.

We first tackled priority accumulation. We conducted a literature review from which we learned that this queueing discipline can only be applied in stable systems. Moreover, it does not affect the throughput of the ED, but only the order in which patients are seen. Priority accumulation can help reduce waiting and throughput times for low acuity patients, yet to the detriment of times for other patient groups. Nevertheless, we decided not to carry on the research on PA for two reasons. First, as we will see, our pods system with the original workforce shows an unstable behavior, therefore, PA cannot be applied. Second, to the best of our knowledge priority accumulation is sensitive in front of staff capacity changes, which is not desirable.

Having decided to not continue with priority accumulation, we focused on having a proper allocation of resources and to investigate the effects of unpooling medical staff and increasing the workforce. For this purpose, we run four experiments:

- Experiment 1: no pods, original workforce
- Experiment 2: pods, original workforce
- Experiment 3: no pods, increased workforce
- Experiment 4: pods, increased workforce

We evaluate the KPIs on the four main services: Acutes, Ambulatory Care, Monitored and Resus. To make it clear, the no pods system consists of two pods: Resus and General (including Acutes, Ambulatory Care and Monitored). The pods system consists of 3 pods: Resus, General (including Acutes and Monitored) and Ambulatory Care.

Table 1: Point estimates of each KPI per service and experiment.

	Mean triage to sign-on (min)				95^{th} percentile total time (min)				Frac. patients leaving within 6h			
	Exp.1	Exp.2	Exp.3	Exp.4	Exp.1	Exp.2	Exp.3	Exp.4	Exp.1	Exp.2	Exp.3	Exp.4
Acutes	68.1	416.1	36.5	51.7	453.5	2449.4	368.1	407.1	0.86	0.48	0.94	0.90
Amb. Care	111.8	68.5	38.2	38.9	549.7	411.6	327.3	330.6	0.85	0.92	0.97	0.97
Monitored	40.0	82.3	22.1	24.5	413.4	603.3	358.1	373.8	0.90	0.79	0.95	0.94
Resus	16.4	18.6	14.0	14.7	307.9	345.9	285.3	288.2	0.97	0.96	0.98	0.98

Table 1 shows the results obtained with the four experiments. Overall, our results seem to indicate that the no pods system outperforms the pods system as it yields shorter waiting and throughput times and an increased fraction of patients leaving the ED on time. Going into detail, when comparing the transition from no pods to pods while keeping the original workforce (Exp.1 vs Exp.2), we see that the pods system manages to significantly improve the metrics for Ambulatory Care patients. Yet, this happens at the expense of notably deteriorating the metrics of Autes and Monitored. When assessing the transitions of both systems to an increased workforce (Exp.1 vs Exp.3 and Exp.2 vs Exp.4) we observe that adding extra physicians results in a noteworthy improvement in both cases. Moving on to the comparison of the transition from no pods to pods while keeping the additional workforce (Exp.3 vs Exp.4), we can see that the pods system does not improve the metrics for Ambulatory Care patients as much as comparing experiments 1 and 2, and neither does it considerably deteriorate the KPIs for Acutes and Ambulatory Care. What happens

now with the pods system is that the Ambulatory Care pod suffers from the same effect as the others, which is that they can be very busy while other pods have idle physicians. Therefore, none of the metrics changes as much as going from Exp.1 to Exp.2, but they are all worse than Exp.3.

While the pods system does not show an outstanding performance in our results, studies in the literature report a performance improvement when implementing pods. We attribute the differences between their pods system performance and ours to human behavior. These studies compared ED data before and after implementing pods, and thus, were able to take into account the human side. On the contrary, we did not model the human behavior and only focused on quantifiable logistical aspects, thereby showing the extent to which the pods system affects the performance from a logistics point of view.

In conclusion, we believe priority accumulation can only be implemented when increasing the workforce as this will ensure a stable system. Given its sensitive nature in front of staff capacity changes, we consider other more robust sorting methods to be more appropriate for the ED. Moreover, we have seen that from a logistics viewpoint, the no pods system outperforms the pods system as it does not cut the physician capacity, resulting in shorter waiting and throughput times. Lastly, increasing the workforce affects both systems positively.

Based on our results and conclusions, from a logistics perspective we do not recommend the ED to make a transition from no pods to pods. On the other hand, the pods system may engender medical staff behaviors that trigger an improved performance, as shown in the literature. Therefore, we do not have enough evidence to give complete advice, for which we believe the ED management has to evaluate whether or not it is worth trying the pods system. Nevertheless, we do suggest increasing the ED's workforce as this will clearly improve the KPIs and bring them closer to the performance targets' values. Moreover, we suggest to carry out further research on implementing pods but using back-up physician pools to reduce the idle-physician/busy-pod problem.

Acknowledgments

I am sitting on a chair in my room thinking how to start this chapter called acknowledgments. Paradoxically, even though this is the beginning of the chapter, it actually is the end of a four year chapter of my life. I started my journey in August 2015, when I moved to The Netherlands to get the one-year Health Sciences masters. After six months, I realized I did not want to go back home. I felt the need to extend this experience in order to continue growing personally and academically, so my parents encouraged me to start another masters. There I was, not having finished the health sciences thesis, but setting out on my first academic year of the Industrial Engineering and Management masters.

Combining both masters was not as easy as I thought. However, with hard work, effort, and perseverance –and lots of support from all my loved ones- I managed to successfully graduate from the first masters, and now, hopefully, I will graduate from the second one. Looking back, there were many ups and downs. I went through a tough anxiety period, shed many tears due to stress, spent 6 months apart from my people, and also went through a surgery. Without all this, I would not be the person I am today, and the achievement I am about to make would not feel as satisfying.

"I want to stay in Twente" I said; "I don't want to complicate my life" I said. Well, I literally ended up in the other side of the world for six months, in New Zealand. How could this have happened? The answer has one name. Erwin, thank you for realizing way before me that this is what I actually wanted and needed and for encouraging me to take on this adventure. You have never doubted me, unlike myself, and you have always believed in me. You have not only been a professor who has taught me many things, but you have been my moral companion on this journey. Thanks for encouraging me to face my fears, for showing me my flaws and help me improving them. Thanks for reminding me my strong points and make me feel proud of them. Without your support, this thesis would not have been possible.

I also want to express my gratitude to Derya, who enthusiastically became my second supervisor and who also never doubted me. I must remark that I am incredibly amazed with her ability to understand and explain my confused thoughts in a very well and structured manner. Moreover, I want to thank Caroline Jagtenberg, who gave me her support from the very first moment and made things easy for me to go to New Zealand. While there, she not only helped me with the thesis, but she also listened to me and gave me advice in my low personal moments. I would also like to thank Michael O'Sullivan and Cameron Walker, who welcomed me in their research group and provided the topic of the thesis. Their guidance, programming help and interesting points of study were essential to build the thesis I am presenting. Caroline, Mike, Cam, you have all been very patient with me, and I will always be thankful for it. Furthermore, I want to express my gratitude to Ilze Ziedins, who was very enthusiastic to help and give expertise regarding priority accumulation.

Special thanks go to Peter Schuur, who encouraged me to start the second masters and who has helped me and given me eternal support during all these years.

I also want to thank all my friends. From Enschede, I would like to mention especially Juana, Marc, Gio, Samara, Rafa, Jose and Aniruddh. From New Zealand, special thanks to Katherine, Keith, Behdad, Annika and Juanpa. All of you have made of these years an unforgettable experience with incredibly funny moments and many challenges (such as running 10Km!). Also, I have to emphasize how much you guys care about me by always making sure there is chocolate, either Milka, Nutella or Whittaker's.

None of this would have been possible without my parents. I do not have enough words to describe my gratitude to them. Their love, strength, courage and effort are the reason why I am able to stand where I am right now. *Mama, papa... No tinc prou paraules per agrair-vos tot el que heu fet per mi. El vostre amor, fortalesa, coratge i esforç m'han fet ser qui sóc i arribar a on estic. Sou el meu pilar més important i la meva raó de seguir endavant. Us estimo.*

Lastly, the person I want to give the most special thanks. Gerardo, thanks for being my companion during these last three years. Your smile, patience and love have shown me the light in the darkest moments I have gone through and have given me joy in the good ones. You were there to help me face my deepest fears and have taught me to go step by step. I would not have even considered going to New Zealand if it had not been for you. Thanks for encouraging me to challenge myself and to fearlessly pursue what I want. *T'estimo.*

Contents

List of Figures

List of Tables

Acronyms

ACH Auckland City Hospital.

ADHB Auckland District Health Board.

CI Confidence Interval.

CNS Clinical Nurse Specialist.

DES Discrete-Event Simulation.

ED Emergency Department.

HO House Officer.

KPIs Key Performance Indicators.

LB Lower Bound.

MILP Mixed Integer Linear Program.

MOSS Medical Officer of Specialist Scale.

NP Nurse Practitioner.

PA Priority Accumulation.

SMO Senior Medical Officer.

TC Triage Code.

UB Upper Bound.

UOA University of Auckland.

WT Waiting Time.

Chapter 1

Previous Research on the ED

The current thesis builds upon the research initiated by the researchers of the Engineering Science Department at the University of Auckland and requested by the Emergency Department (ED) at Auckland City Hospital (ACH). In order to understand the focus of the current study, this chapter explains the previous research conducted in the ED up to the starting point of this thesis. All the information presented is extracted from meetings with the ED management and reports elaborated by the UoA researchers. The chapter is structured as follows. In Section 1.1 we introduce the background and describe the functioning of the ED. Next, we describe the simulation model developed by the UoA researchers in Section 1.2. Lastly, we conclude the chapter with Section 1.3, where we explain the simulation model's application to the ED of ACH.

1.1 Context

1.1.1 Research Group

The investigation presented in this thesis is a result of a research visit of the author (from the University of Twente in The Netherlands) to the Engineering Science Department at the University of Auckland in New Zealand. The department focuses its research on three different fields: operations research, mechanics and biomedical engineering. The group of researchers formed by Thomas Adams, Michael O'Sullivan and Cameron Walker, from the operations research and computational analytics group, initiated this project. Later on, Caroline Jagtenberg from the same group joined the project together with the author.

1.1.2 Emergency Department in Auckland City Hospital

Auckland City Hospital (ACH) is the largest public hospital and clinical research facility in New Zealand. Operated by Auckland District Health Board (ADHB), ACH has approximately one million patient contacts each year including hospital and outpatient services. It has around 11,000 health and medical staff employed and they train about 1,800 medical staff, becoming the largest trainer of physicians in New Zealand. The ED sees approximately 70,000 patients of different acuity

levels throughout a year, of which around 40% are admitted to hospital. According to the ED Clinical Director, during the past years the demand for ED services has increased by 5% annually.

In the coming subsections we introduce the reader to the functioning of the ED. This system description is necessary so as to understand the conceptual model of the simulation model described in Section 1.2.2.

ED Space Division

The ED is divided in 4 main areas: Resus, Monitored, Acutes and Ambulatory Care. Resus treats the critically unwell patients that require resuscitation; Monitored deals with patients needing consistent observation; Acutes is where patients require a bed but not monitoring and Ambulatory Care treats the least severe cases. There are a total of 8 different services divided into these four areas. The eight services are: Resus, Monitored, Acutes, Ambulatory Care, Short Stay, Consultation, Procedure and Other. Table 1.1 shows the services within each area. Moreover, the ED extended its bed capacity from a total of 69 beds in 2017 to a total of 80 beds in 2018. The beds for each service were re-distributed.

Table 1.1: Services within the four main areas of the ED

	Area			
	Resus	Monitored	Acutes	Ambulatory Care
Services	Resus	Monitored	Acutes Procedure Short Stay	Ambulatory Care Consultation Other

Medical Staff

Physicians are allowed to treat different acuity patients according to their skills. There are a total of six physician categories depending on the level of experience and knowledge. From top to bottom these are: Senior Medical Officer (SMO), Fellow, Medical Officer of Specialist Scale (MOSS), Registrar, House Officer (HO) and Clinical Nurse Specialist (CNS)/Nurse Practitioner (NP).

Rules

Each pool of physicians needs to be staffed by at least one SMO from 8:00 to 24:00. Also, staff will not see a new patient in the last hour of their shift. The specific rules for each area are:

- *Resus*: this area is open 24 hours a day and needs to be staffed by two people (1 SMO and a MOSS or Fellow or Registrar) from 8:00 to 24:00. 60% of the patients will be seen by one physician only and 40% will be seen by both physicians.
- *Monitored*: this area is open 24 hours a day and is staffed by the general pool except for HO and CNS/NP.

- *Acutes*: this area is open 24 hours a day and is staffed by the general pool except for CNS/NP.
- *Ambulatory Care*: this area is open from Monday to Tuesday from 8:00 to 23:00 and Friday to Sunday 24 hours. When closed, its patients belong to acutes. Ambulatory Care is staffed by the general pool and it is only here where the nurse practitioners (CNS/NP) can work.

Performance Targets

The Ministry of Health and ADHB negotiate the performance targets of the services that are provided in the hospital. The main target is for 95% of the patients to be admitted to hospital, discharged or transferred from the ED within six hours. The breach of this target partly depends on uncontrollable factors by the ED, such as waiting for diagnostic imaging or laboratory tests. Instead, the ED focuses on providing an initial physician assessment within one hour from the patient's arrival. They consider that this will allow the patient to be discharged from the ED within the six hours target.

Patient Process

Upon the patients' arrival, they are first seen by a triage nurse and assigned a triage code from 1 to 5 depending on their severity. The time that this occurs is called the "Triage time". Then, they wait to be assigned to one of the areas in the ED to receive assessment and/or treatment. Once a patient is in the appropriate area he/she waits to be seen first by a nurse and afterwards by a physician. The first time a physician sees a patient is termed the "Sign-on time", and it is the time from being triaged until this point that the ED wishes to ensure is less than one hour. If required, the patient will have some tests or scans done. The same physician will continue to observe the patient until a decision is made, which can either be to admit them to hospital, discharge them or to transfer them to another facility. The time when a decision is made is called the "Decision time". Sometimes the decision can be made almost immediately, in other cases it can take significantly longer. While the patient is waiting to be assessed, for a decision to be made, or even after a decision is made, it is possible for the patient to be moved to another area in the ED if their medical condition necessitates it.

Physician Process

At the start of a shift, physicians choose the patient that they will treat. On the one hand, if the chosen patient needs sign-on, the physician will make the assessment, decide whether tests or scans are needed and will fill in forms and information to the system. We refer to this last step as paperwork. On the other hand, if the patient needs a decision, the physician will check the results -if any-, make a decision, notify it to the patient and complete the paperwork.

Junior physicians (Registrars and HO) need advice from a senior physician (SMO) after the first assessment and before notifying the decision to the patient. However, if the SMO is occupied, juniors might decide to continue working and consult with the SMO later on. Also, it is up to the physician to decide when the paperwork is done. Some physicians prefer to visit several patients in a row and then register all their notes in the computer afterwards. Contrarily, some opt to do the

paperwork after each consultation. Physicians take care of the patient from the assessment to the decision, unless their shift is finished and other physicians take over. Finally, in the case that all physicians are busy and Resus needs more people, a call will be made and someone will leave the task that was doing and will take care of the Resus patient. Afterwards, the physician will continue with the left task.

Choice of Patient

When there is more than one patient waiting to be seen, the free physicians decide which patient goes next according to three different factors. These are: 1) the urgency of the patient (triage code), 2) the waiting time and 3) to some extent, the personal preferences of the physician. Also, physicians will tend to choose patients that need to be notified about the decision -if their results are ready- before choosing a patient that needs assessment. This way, patients are discharged from the ED and beds become available for patients that need to be moved to another service within the ED or for new patients. Moreover, the information system highlights those patients that have spent more than 4 hours in the ED. This is done with the aim to get the physicians' attention that these patients need to be seen.

Pooling of Resources

Until December 2018, there was a general pool of physicians where all the medical staff worked together as one large group to treat all the services except for Resus, who had its own dedicated team of two physicians. We refer to this system as "no pods". The ED decided to unpool resources by implementing a "pods" system, where physicians and nurses are assigned to work in teams in a specific geographic area. Therefore, each pod has an own pool of medical staff. The idea behind this decision is that having some staff dedicated to the less severe cases will help getting patients through the ED on time and to comply with the performance targets. In December 2018, they slowly started the transition by having 3 pods: Resus, General - consisting of Acutes and Monitored - and Ambulatory Care. Even though the final ED's intention is to implement 4 pods, in the remainder of this project we do research on the 3 pods system, which we simply refer to as "pods system". Figure 1.1 shows the two systems that will be dealt with in this thesis: the no pods system with 2 pools of physicians and the pods system with 3 pools of physicians. In the remainder of the thesis we refer to pods and pools indistinctively as each pod owns a pool of physicians.

Until December 2018 the ED was using the roster shown in Table B.1 in Appendix B. With the implementation of 3 pods, they intend to use the roster shown in Table B.2 in Appendix B, which provides a total of four more physicians.

Contact with the University

Since the ED wanted to do a transition from no pods to having 4 pods, they contacted the aforementioned research group to have a better understanding of the effect of different staffing models on patient quality of care, including metrics such as waiting time and staff workloads. The next section introduces the reader to the research approach the researchers came up with.

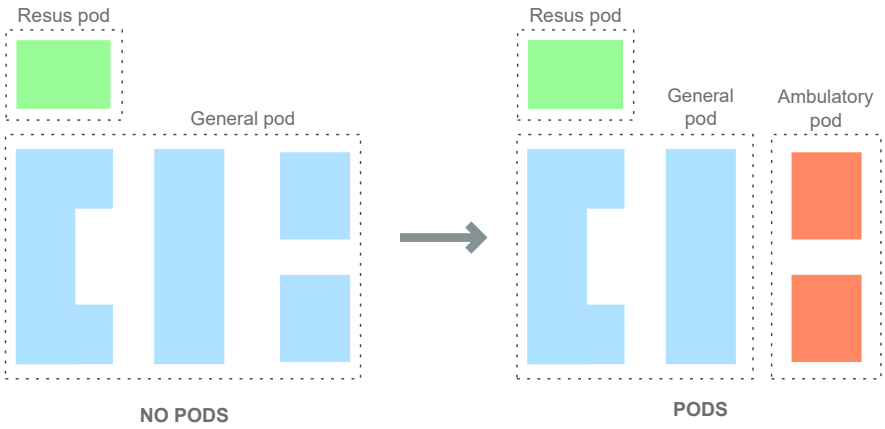

Figure 1.1: Transition from the no pods to the pods system

1.2 Simulation model

1.2.1 Overview

The ED researchers decided to develop a Discrete-Event Simulation (DES) model of the ED. Before, going into detail with the conceptual model, we give an overview of the research approach designed by the UoA researchers (Figure 1.2).

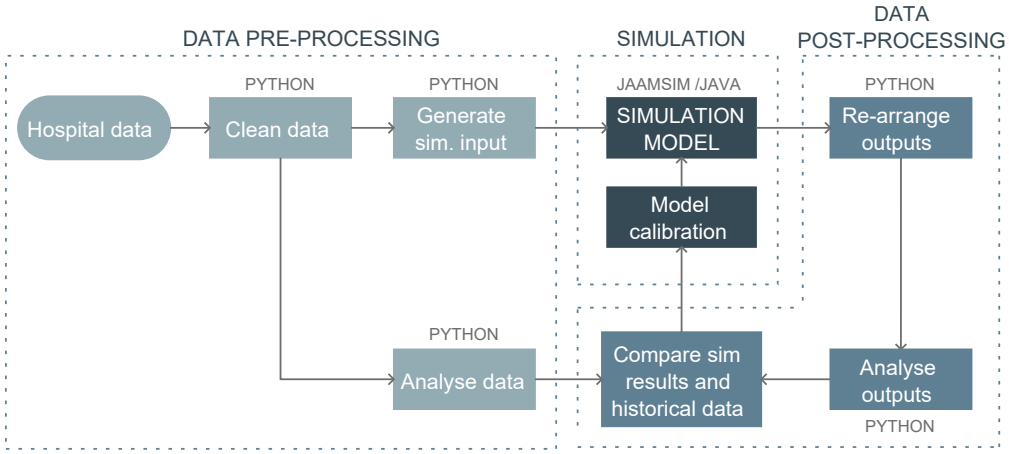

Figure 1.2: Overview of the previous research in the ED

First, the hospital provided historical data that the researchers cleaned and analyzed with Python. This was used to generate a file with input parameters for the simulation. The hospital also informed about the functioning of the ED, such as the patient process, the physician process, rosters, rules, etc., which was used to develop the simulation model of the ED. The simulation model was built with the free open source software program JaamSim, which runs in Java. Once the experiments were run in the simulation and their results available, the researchers analyzed the text file generated by JaamSim using Python. They translated the simulation outputs into point estimates of KPIs and time distributions. All of these were compared to the historical data to be able to calibrate the simulation model.

1.2.2 Conceptual Model

In this section we explain the most important and essential parts of the abstraction of the simulation model from the real world it is representing. The two main roles modeled are patients and physicians. These two go through different activities and have certain behaviors and only interact with each other when a patient needs to be assessed or a decision is notified. During the rest of the activities, patients either remain in their assigned rooms, are waiting in a queue or are undergoing tests and scans. When physicians are not with patients they are doing paperwork, making decisions, interacting with other physicians or they are idle. In the following subsections we introduce the inputs and outputs, modeling simplifications and assumptions, patient and physicians process and the main choice done in the simulation: which patient to choose next and how to dispatch a physician.

Inputs and Outputs

The inputs for the model are:

- Patient characteristics: triage code, care path, arrival time, assessment duration, decision duration, tests & scans duration and after decision duration.
- Physician characteristics: number of physicians, role, shift hours, physician pool, make decision time distribution, paperwork time distribution and advice time distribution.
- Room capacity

The outputs of the model for each patient are: arrival time, triage time, sign-on time, decision time, depart time and assigned service upon arrival.

Model Simplifications

1. Ambulatory Care is always open.
2. A patient is handled by one physician at a time and a physician physically visits one patient at a time. This also implies that 40% of Resus patients are not seen by two physicians but just one, as opposite to the rules.
3. All waiting rooms have infinite capacity.
4. There are no physician interruptions. Thus, if there are no available physicians upon the arrival of a Resus patient, this has to wait.
5. Breaks during staff shifts are not considered.
6. After sign-on and notifying the decision, physicians immediately proceed to do paperwork.
7. The decision making process - when patients' results are ready - before notifying the decision to the patient is not modeled.
8. Patients with triage code 3, 4 and 5 are treated as if all had the same triage code in order to avoid very low acuity patients to wait too much.

9. Nurses are not modeled, except for the triage nurse - who is always available- and two nurse practitioners that take care of Ambulatory Care patients.
10. Travel distances and travel times between rooms are neglected.
11. The process of patients undergoing tests and scans is simplified by making them wait in their rooms a specific amount of time.

Model Assumptions

1. All physicians work at the same pace at all times. Work is not sped up when there are longer queues.
2. Due to the lack of data, the time that physicians take to make a decision, do paperwork or give advice to junior physicians is assumed to follow a normal distribution.
3. In pods, if a patient is moved to another pod and needs a decision, the physician that did the assessment in the previous pod will walk to the pod where the patient is located.

Patient Process

The patient process described in section 1.1.2 can be summarized with the activity cycle diagram shown in Figure 1.3. The essential idea is that upon arrival the patient is triaged, sent to an area, assessed, a decision is made and afterwards he/she leaves the area and the ED. However, several movements between areas are possible and the patient can have the assessment and the decision in different areas, as well as to just stay in an area without seeing a physician. Nevertheless, the patients' behavior is more complex than this, which is shown in the patient's behavioral diagram in Figure 1.5 (left diagram).

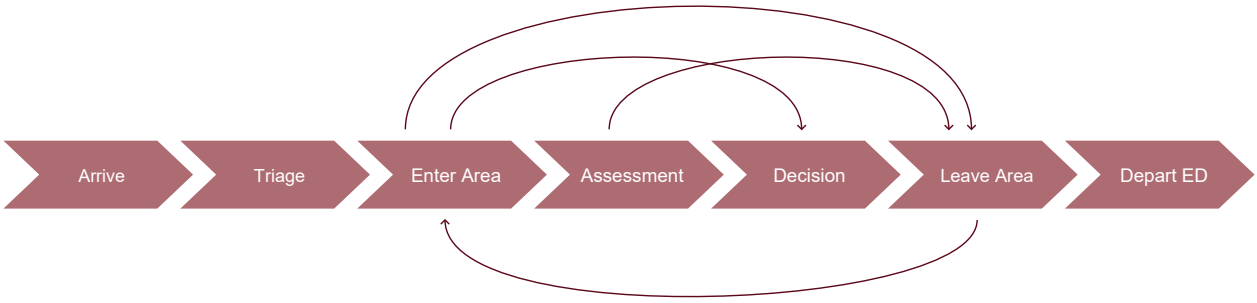

Figure 1.3: Activity cycle diagram for ED patients

Physician Process

Figure 1.4 summarizes the physician process described in section 1.1.2. The basic idea is that physicians visit patients in two steps - assessment and decision - and they do administration work after each of these two steps. After an assessment or notification of decision, the physician can assess another patient or notify decisions to other patients. This is repeated until the shift ends. The process of making decisions is not modeled in this initial model. The behavioral cycle diagram of physicians shown in Figure 1.5 (right) depicts in more detail how physicians work.

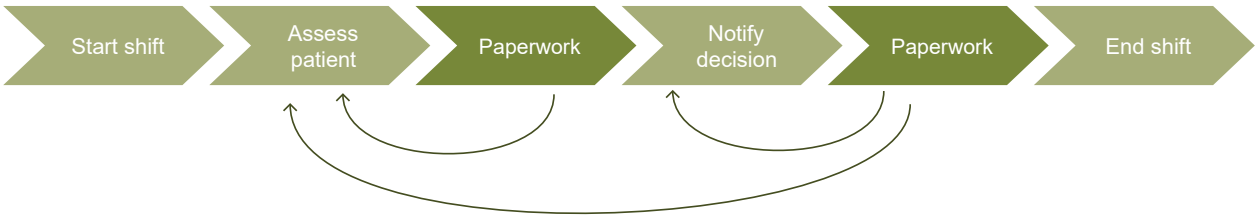

Figure 1.4: Activity cycle diagram for ED physicians

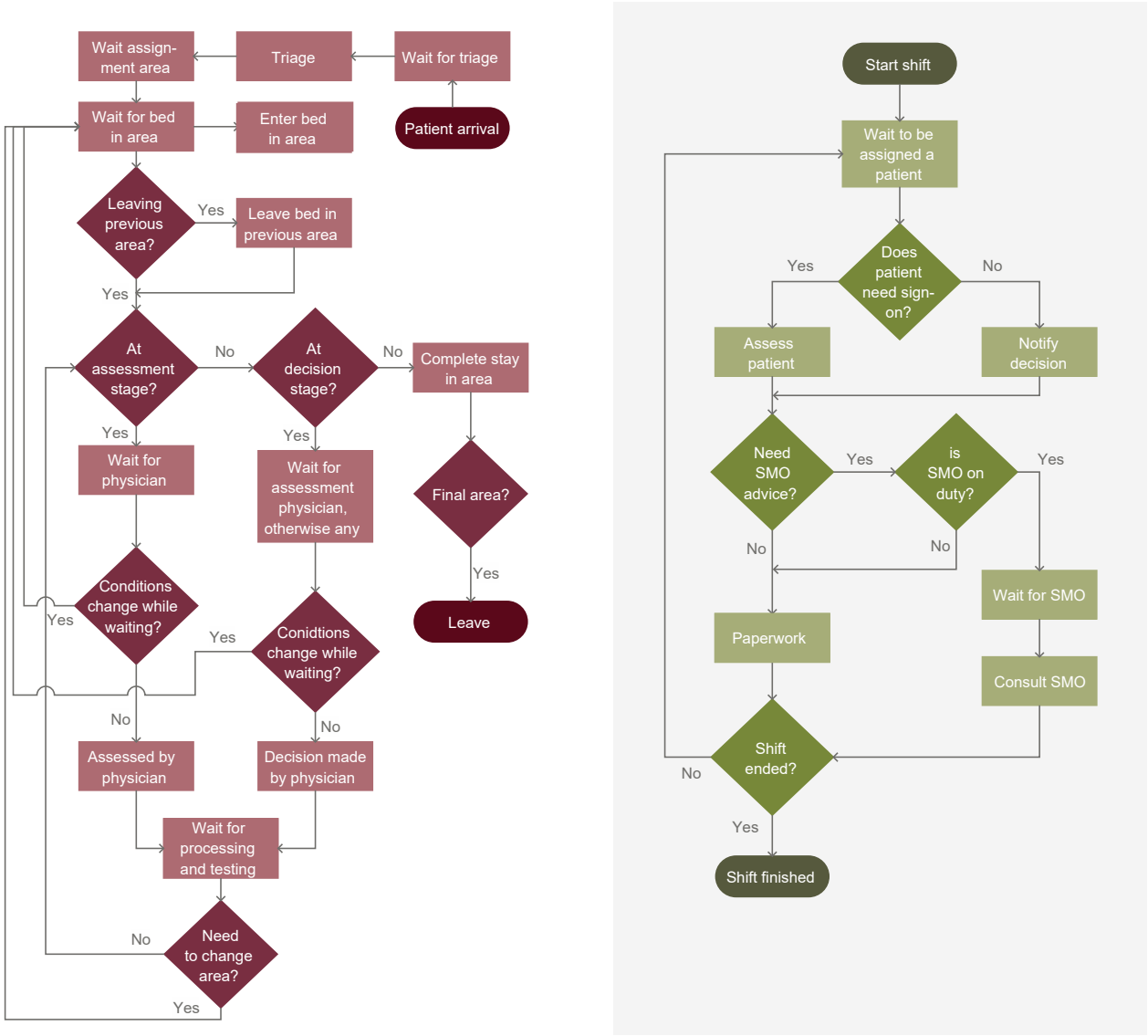

Figure 1.5: Behavioral cycle diagram for ED patients (left) and physicians (right)

Choice of Patient

The most important decision in the whole simulation is how to choose the next patient to be treated and how to dispatch physicians to patients. Figure 1.6 shows the logic diagram behind it. To start

with, every time a physician becomes free, a new patient arrives or the patient's results are ready all the patients waiting in the ED are checked. This means that all of them are sorted in a list according to their triage code (most urgent patients on top). If two or more patients have the same triage code, they are sorted according to the longest waiting time since they started waiting for the last time, not since they were triaged. Then, starting from the top of the list, a compatible physician is dispatched to each patient. If there is not a compatible physician for a patient, the dispatching goes on with the next patient until all free physicians are dispatched or there are no more patients waiting. A compatible physician means someone that is:

- On duty
- Not busy
- Enough skills for that patient
- If decision is needed, the physician needs to be the same one (if on duty) that assessed the patient

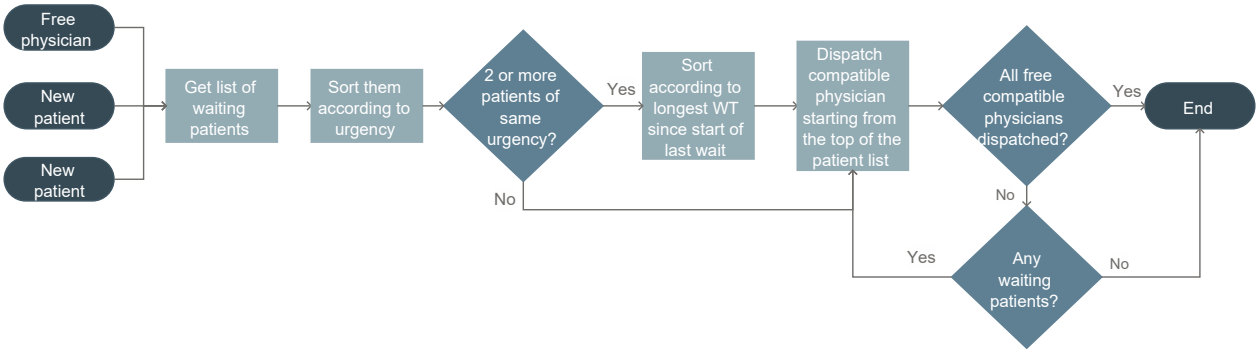

Figure 1.6: Logic Diagram for choosing the next patient to attend

1.3 Application of the Simulation Model to the ED

The conceptual model is then translated into a computer model. A screenshot of the main frame of the simulation model implemented in JaamSim can be seen in Figure A.1 in Appendix A. In this section, we go through the UoA reserachers investigation. We explain how they processed and analyzed the historical data and the results they obtained. We show the KPIs they defined and the results of the simulation model. We finalize with their suggestions for future research.

1.3.1 Data Pre-processing

The ED provided a data file containing information about the movements of 70,000 patients within the ED for the year 2017. Each row corresponded to a patient's stay in a room in the ED, thus several rows could belong to one patient. The information recorded in each row was: event ID, admit time, triage time, gender, age, room number, room enter and leave time, ED discharge time and other time stamps and descriptive statistics. A lot of this information was continuously repeated, making it difficult to process. The researchers combined the rows of the data that corresponded to

the same event ID to get a complete picture of a patient's stay in the ED. The data was simplified by matching every room to a service, and combining room stays within a single service into a single service stay. Also, visits to the service "Other" before visits to any other service were removed as it was assumed that these visits are actually patients waiting to be assigned to another service. Moreover, event times were adjusted to fit the order shown in Figure 1.3, since it is assumed that records where sign-on occurs after a decision is made or before an area is entered are errors in the data.

Moreover, some further processing was needed. Because only the sign-on time, decision time and departure times were available in the original data -and not the actual time physicians spent with the patients- several assumptions needed to be made. They assumed that all the time a patient spends with a physician can be represented in two blocks, one occurring after the sign-on and one after the decision. The total duration of the two blocks follows a triangular distribution with parameters estimated by the ED physicians. This total time is partitioned randomly between the two blocks with at least 30% in each block. Any remaining time between sign-on and decision, or decision and departure, is assumed to be time patients spend having tests, scans or other activities that are not a "wait".

With all this processing, the data could be analyzed and the input file for the simulation model generated.

1.3.2 KPI Definition

The Key Performance Indicators defined by the researchers reflect the performance targets introduced in Section 1.1.2. The KPIs are:

- *Mean Triage to Sign-On Time*: average time between the assignment of the triage code to the patient and the time the physician does the assessment. Ideally, this is less than one hour.
- 95^{th} *percentile of the total time*: value of the total time spent in the ED below which 95% of the patients fall. This value should be minimized but at most 6 hours (360 minutes).
- *Fraction of patients leaving the ED within 6 hours*: it indicates the fraction of patients that have spent at most 6 hours in the ED. Ideally, this percentage is at least 95%.

1.3.3 Analysis of Historical Data

The researchers examined the time from admission to four other time stamps: triage, sign-on, decision, and departure. They observed that, as the triage code increased (less urgent) the mean throughput time in the ED decreased. Another noticeable difference between the triage codes was that for codes 1 and 2 there was very little time between triage and sign-on, and a much longer time between sign-on and decision. For the higher codes the time between triage and sign-on is about the same as between sign-on and decision.

They also analyzed the main KPIs, shown in Table 1.2. As it can be seen, all the KPIs comply with the performance targets, except for Acutes and Ambulatory Care that slightly breach the assessment-within-one-hour target. In all the services more than 95% of the patients stay in the

ED less than 6 hours, which also means that the 95^{th} percentile of the total time is lower than 360 minutes.

Table 1.2: Key performance indicators for the hisotrical data

	Mean triage to sign-on time (min)	95^{th} percentile of the total time (min)	Fraction of patients leaving within 6h
Acutes	79.36	352.00	0.96
Ambulatory Care	73.59	316.00	0.99
Monitored	19.14	336.65	0.98
Resus	3.42	322.00	0.99

They also investigated the most common sequences of events - the paths - that occur in the data. The 20 most frequent ones make up 48% of all the patients that go through the ED.

1.3.4 Experiments and Results

The researchers performed several experiments using the simulation model. The different scenarios represented the no pods and pods systems, some with extra personnel, different rosters and different managerial roles for the senior physicians.

As mentioned in section 1.1.2, junior physicians need advice from senior physicians (SMO). In order to give attention to this advise need, in some of the experiments SMOs were only assigned a managerial role, thus, they would not see patients. This was tested in the no pods system and the reduction in terms of staff available to see patients was significant. The outcome was that low acuity patients needed to wait very long as they were understaffed and, hence, high acuity patients were seen first. Adding extra staff when SMOs were only assigned the managerial role also improved the KPIs. They also tried assigning 50% of the SMOs' time to managerial role and 50% to seeing patients. This significantly improved the metrics due to the boost of staff.

For the pods system, the results obtained showed that it enables targeting specific areas and, thus, to significantly improve them. However, other areas suffer a significant performance drop. This happens because patients are only seen by physicians assigned to the area they are in, so it is possible that there are idle physicians in one area while patients are waiting in another area. They also identified that long waiting times for low acuity patients in higher acuity pods become a problem. Even though two different rosters for the pods structure were tested, they did not identify an optimal one.

1.3.5 Conclusions and Ideas for Future Research

They concluded that keeping the original configuration or to change to a pods system and adding some extra staff result in a similar performance. One of the big advantages of the pods system is that it allows to target specific areas that otherwise do not have a good performance. Moving to a pods system may confer additional advantages that are not captured by the model, such as

better communication or physicians staying in an area and treating patients with similar medical conditions. In addition, the researchers suggested two main aspects for future research:

- First was to change the way the senior physicians consulting with junior physicians is implemented. This is done in such a way that seniors are able to simultaneously give advice to a junior and to see a patient, which is not possible.
- Second was to implement priority accumulation for the patients as they wait. Currently a lower priority patient will have to wait for all the higher priority patients to be seen before he/she is seen. With priority accumulation it would be possible for a low priority patient to be seen earlier than a higher priority one if they have been waiting for a longer time and, thus, they have built up some priority.

Chapter 2

Research Plan

Now that the previous research is clear, we can focus on the starting point of the thesis. In this chapter we present the research plan. We define the research goals in Section 2.1, followed by the scope in Section 2.2. We continue with the translation of the goals into a research question in Section 2.3 and we finish by explaining the research approach and thesis structure in Section 2.4.

2.1 Research Goals

The objective of this research is two-fold. On the one hand, we improve the simulation model and on the other hand, we focus on optimizing the ED from a logistics point of view. Regarding the model improvement, verification of the simulation model needs to be done and the number of replications needs to be determined. When it comes to optimizing, we explore the implementation of priority accumulation, the effects of the no pods and pods systems and, lastly, how these two behave when adding extra physicians.

2.2 Research Scope

First, regarding the model improvement we focus on carrying out the necessary changes for the model verification and validation, as well as to implement certain features that will allow for future research. We consider out of scope modeling human behavior, which requires a complex and big model extension. Second, with respect to the optimization and implementation of pods, we limit ourselves to the re-allocation of the available staff into pods, regardless of whether this staff amount is sufficient or not to face the ED demand. In addition, we do not consider inside our scope the possibility of changing start and end times of the rosters. Lastly, we assess our interventions based on the ED's bed capacity in 2017, and do not take into account the bed capacity extension that the ED underwent in 2018.

2.3 Research Questions

The research goal is translated into the following main research question:

How can priority accumulation, the pods system and an increase in medical staff help improve the waiting and throughput times of the Emergency Department in Auckland City Hospital?

2.4 Research Approach and Thesis Structure

Once having understood the previous research (Chapter 1) and having a clear starting point (Chapter 2), the reminder of the thesis is structured as follows. In Chapter 3 we address the simulation model improvement by carrying out the necessary changes to verify it and validate it. We also contribute to the model by adding features that will allow for future research. In Chapter 4 we study the state-of-the-art regarding priority accumulation, we explain how we implement it in the simulation model and we discuss its effects on the ED. Moving on to Chapter 5, we do a literature review about the pods system where we learn about its benefits and we see the effects it has had in other hospitals. With the knowledge gathered from the literature, we design the solution approach where we use the improved simulation model and, besides, we use a staffing model to allocate the physicians into pods. In this chapter we also present the staffing model formulation, the results and its contributions and limitations. In Chapter 6, we show the results of running four experiments with the simulation model. These experiments allow us to understand the effects of implementing the no pods and pods systems and how their performance is affected when extra physicians are added in the ED. Lastly, we conclude this research with Chapter 7, where we summarize our findings and give an answer to the research question. We also make recommendations to the ED, present the limitations of this study, suggest future research lines and close the chapter with the project contributions to practice and science.

Chapter 3

Improvement of the Simulation Model

In this chapter we present the modifications we made to the model in order to improve it, the changes in the data processing and we also explain how we verified and validated the model. Section 3.1 presents the model extensions, followed by Section 3.2, where we calculate the number of replications needed. Section 3.3 explains the necessary changes in the data processing. Section 3.4 shows the verification and validation of the model. We close this chapter with Section 6.3 where we summarize the taken steps for the model improvement.

3.1 Model Extensions

3.1.1 Senior/Junior Advice and Paperwork Process

The first and most important modification in the model is to change the way that senior physicians give advice to the junior ones. This is done together with a more detailed implementation of the paperwork process.

Paperwork Process

In order to understand the main change, we first need to explain the extension of the paperwork process. As shown in Figure 1.4 in section 1.2.2, the main physician's activities are the assessment of the patient and notifying the decision to the patient, both activities followed by paperwork. Figure 3.1 shows the extended activity diagram of the physicians. The original paperwork process is now split into two parts: make decisions and do paperwork. Moreover, between the assessment and notification of the decision, we add another step: to analyze the patient's results. This new process is a better representation of reality.

Figure 3.1: Extended activity cycle diagram for ED physicians

Senior/Junior Advice

Regarding the advice process, in the initial model, the availability of the senior physician to give advice to the junior one was controlled by a time series that switched on and off throughout the day. The downside of the time series is that it caused the senior physicians to simultaneously see patients and give advice to juniors, which is an impossible scenario. In the new approach we get rid of this time series and introduce a controller that is triggered whenever a physician becomes free, a physician finishes an assessment or patients' results are ready. If the controller finds juniors waiting, it will send a senior physician -if available- to give advice. Both, the senior and junior will analyze results and/or make decisions together (activities filled with brighter green in Figure 3.1). Thus, the paperwork is no longer done under the senior's supervision. Moreover, we prevent the senior physicians from becoming bottlenecks by letting the juniors make their own decisions if they have been waiting for x amount of time. Figure 3.2 shows the behavior that physicians have in the ED with the aforementioned extensions included. Figure 3.3 displays the behavioral cycle diagram specifically for SMOs.

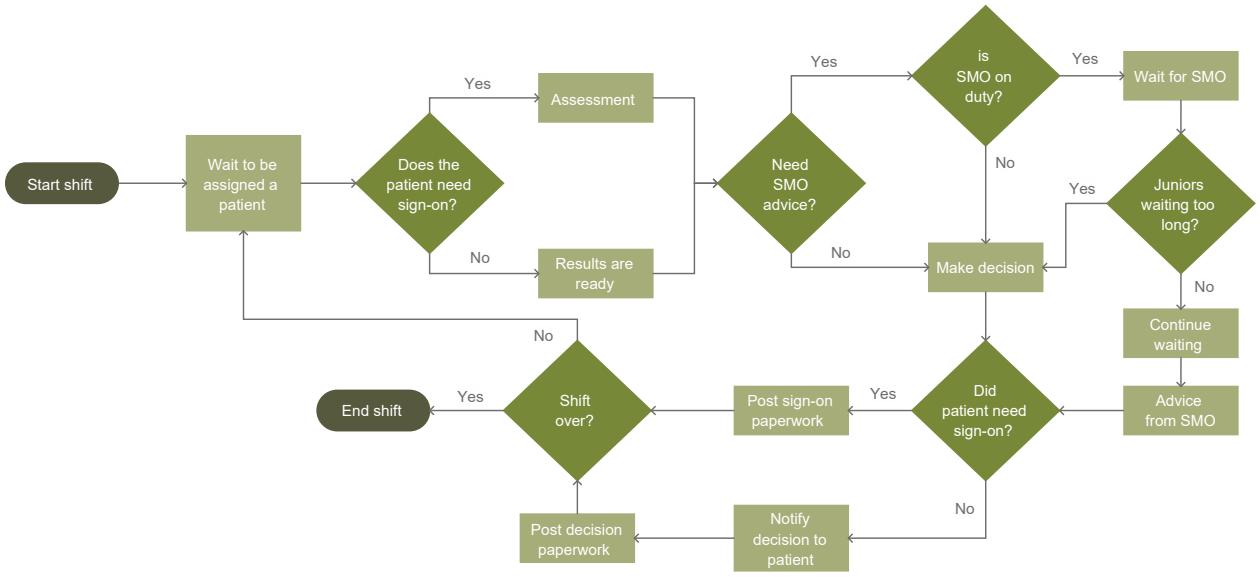

Figure 3.2: Extended behavioral cycle diagram for ED physicians

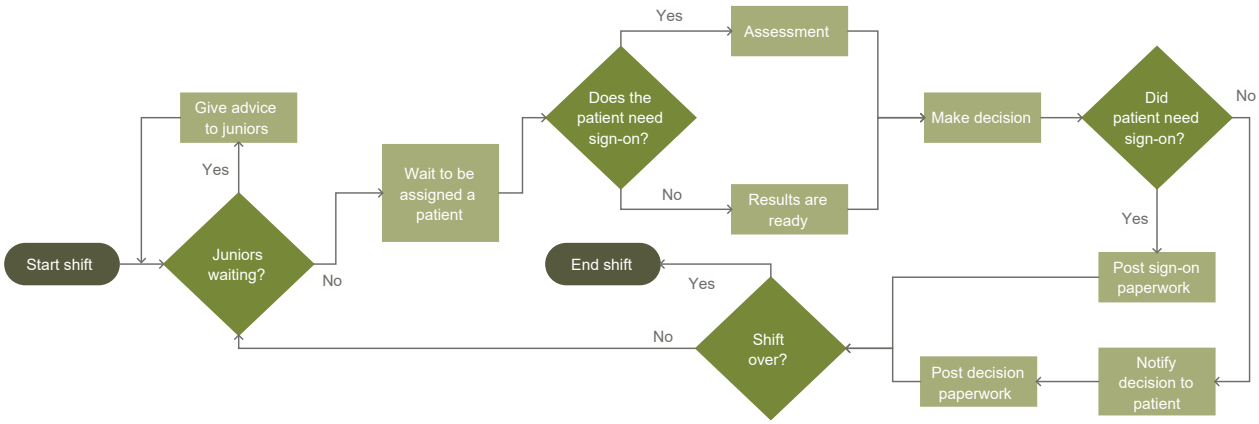

Figure 3.3: Extended behavioral cycle diagram for SMOs

3.1.2 Patient Sorting

The second main modification offers the end-user the possibility of sorting patients in six different ways, shown in Table 3.1. To understand them, we need to introduce the factors that influence the sorting order. These are:

- Triage Code (TC): patients with lower triage codes are given priority.
- Waiting Time (WT): this can either be the waiting time since the patient was last attended to (WT1) or since the triage (WT2). In both cases, longer waiting times are prioritized.
- Service (S): patients get priority depending on the service they belong to.
- Sign-on flag (SF): patients are flagged, i.e., given priority, when they are about the breach the one-hour-sign-on-target or when they have spent more than 4.5 hours in the ED. The first flag appears x minutes before breaching the one-hour-sign-on-target. This amount of time is specified by the user and can vary among the different triage codes.
- Decision flag (DF): patients that need decision and have spent more than 4.5 hours in the ED are prioritized.

3.1.3 Physician Queues and States

The third improvement is done with the aim to facilitate the verification of the model. In the initial model, all the physicians would wait in a queue, whether they were off-duty or on duty but idle. This made it difficult to follow them in the simulation and to check that the flow was the correct one. We have changed this by making two different queues, one for off-duty physicians and one for on-duty idle physicians. Furthermore, we have defined states for the physicians so as to be able to track their work and to see what fraction of their time they spend in each state. These stare are: idle, make decisions and paperwork, giving or receiving advice, waiting for advice and, last, seeing patients.

Table 3.1: Patient sorting methods

Method	Factors order	Description
1	TC	Patients are sorted according to the TC only.
2	TC >WT1	If two patients have the same TC, they are sorted according to their WT1.
3	SF >TC>WT1	The patients that are flagged appear on top of the list and the ones that are not appear on the bottom. Within each of these two groups, patients are sorted according to the TC. If the TC is the same, they are sorted according to the WT1.
4	DF>TC>WT1	The same as the previous method but with the decision flag instead of the sign-on flag.
5	DF>S>TC>WT2	After sorting the patients according to who is flagged and who is not, within each of these two groups they are sorted depending on their service. If the service is the same, they are sorted according to the TC. If this is the same, the sorting depends on the WT2.
6	WT2 & TC	This method is called Priority Accumulation. See Chapter 4 for more details.

3.1.4 Model Simplifications & Assumptions

Due to the model extensions, we need to modify or add simplifications and assumptions. On the one hand, we remove simplification 6 and 7 presented in Section 1.1.2. On the other hand, we add one simplification and one assumption. The new simplification is for juniors to make their own decisions if they have to wait for too long and to not return for advice. The new assumption is for patients needing decision to be prioritized over new patients, except if Resus or triage code 1 patients need assessment.

3.2 Number of Replications

For the simulation study, the length of a run needs to be defined and the number of replications set. This is a terminating simulation with a run length of one year. Regarding the number of replications, we need to compute how many we need in order to achieve the desired level of confidence of the model output. To do this, we follow the sequential procedure proposed by Law (2007). This states that the smallest n for which Expression 3.1 holds determines the number of replications needed.

$$\frac{t_{n-1,1-\frac{\alpha}{2}}\sqrt{S_n^2/n}}{\overline{X_n}} < \gamma' \tag{3.1}$$

In this formula, n is the replication number, $\overline{X_n}$ is the average of n replications, S_n^2 is the variance in n replications, $t_{n-1,1-\alpha/2}$ is the t-student with $(n-1)$ degrees of freedom and confidence level $(1-\alpha)$. Finally, γ' is the relative error.

To achieve a 95% confidence and a relative error of at most 5% (γ' is 0.048), we need a minimum of 3 replications (Table 3.2). Running 3 replications with a run length of 365 days implies a run-time of 14 minutes and 39 seconds.

Table 3.2: Number of replications per performance measurement

KPI	N° of replications	Relative error
Mean triage to sign-on time	3	0.0065
95th percentile of the total time	3	0.0341
Fraction of patients leaving within 6h	2	0.0285

3.3 Changes in the Data Processing

Data Pre-Processing

As explained in Section 1.1.2, the ED underwent a bed capacity expansion in 2018. Even though we do not model it, the extension actually involved the re-distribution of beds among services, resulting in services with different bed capacities compared to 2017. While increasing the bed capacities and re-distributing beds among services can easily be done in the simulation, the problem comes to how patients are distributed into the services in the simulation. Keeping the same patient classification into services while changing the services' bed capacities would result in longer queues for some services and shorter ones for others. Thus, to allow simulating the extended bed capacity situation in future work, we provide the possibility - in the Python code - to split the patients into different services. In more detail, we split Acute patients into three groups: high acutes, low acutes and ambulatory acutes. This way, in a future research we will not send the same volume of patients to services that have a different capacity.

Data Post-processing

We also change the post-processing of the data in order to take into account several replications. This implies a big change in the Python code since the output file generated by the JaamSim, which includes the replications, is complex and troublesome. The data post-processing outputs show the point estimates of each KPI for each main service per replication. It also shows the average for all the replications per service, together with the confidence intervals.

3.4 Model Verification and Validation

Verification

One of the objectives of this study was to verify the model developed by the UoA researchers. We have verified the model using three different techniques (Law, 2007). The first one consists in developing and testing the model in different modules, thereby writing and debugging the main parts of the model and successively adding and debugging more levels of detail. This way we have gradually made the model more complex. Second, we have compared the state of the simulated system (event list contents, state variables, statistical counters,...) with hand calculations to make sure that the simulation worked as intended. Last but not least, during the whole process we have observed the animation of the simulation to ensure a correct flow. Using these strategies we found several bugs that we corrected.

Validation

In this study we address the model validation by means of comparing the real system and the model outputs and by interacting with the ED medical staff. Not surprisingly, when comparing the simulation results with the historical data we found discrepancies. These differences suggested how to improve the model. In short, these changes are:

1. First dispatch physicians from the Resus pool and then from the other pool(s). Otherwise, Resus patients wait too long for assessment.
2. Junior physicians are allowed to make decisions on their own if they have been waiting for too long. Otherwise, the SMOs become the bottlenecks of the simulation generating long waiting times and queues.
3. Give priority to patients that need decision and have been in the ED for more than 4.5 hours. Otherwise, beds are not freed and patients needing assessment wait for too long.
4. Give different priorities to the services within the general pool. Otherwise monitored patients wait too long for assessment.

When applying these changes we generated the patient sorting method 5, introduced in Table 3.1. This method resulted to be the one reflecting reality the best. Despite these modifications, the model was still not valid. There were discrepancies that needed to be solved by calibrating the model, i.e., tweaking parameters by trial and error. These included the parameters of the normal distributions of the paperwork and decision times and scaling factors among others.

Figure 3.4 and Table 3.3 show the results of the model validation. Notice that the simulation tails are longer than the ones of the historical data. This explains why the 95th percentile of the total time is significantly higher in the simulation KPIs. We can also observe that the mean is not the best estimator for the triage to sign-on time, since, for example, the distribution of Ambulatory Care is quite similar but the tails are longer, and this biases the average. This is why to validate the model we look at the KPIs and time distributions. We now have a model that reassembles reality and it is validated through expert judgments (UoA researchers and ED management). We did not

statistically validate the model as it was out of our scope, however it should be part of the future work.

Table 3.3: Comparison of historical and simulation KPIs for validation

	Mean triage to sign-on time (min)		95^{th} percentile of the total time (min)		Fraction of patients leaving within 6h	
	Hist	Sim	Hist	Sim	Hist	Sim
Acutes	79.36	68.13	352.00	453.52	0.96	0.86
Ambulatory Care	73.59	111.81	316.00	549.73	0.99	0.85
Monitored	19.14	40.04	336.65	413.35	0.98	0.90
Resus	3.42	16.44	322.00	307.89	0.99	0.97

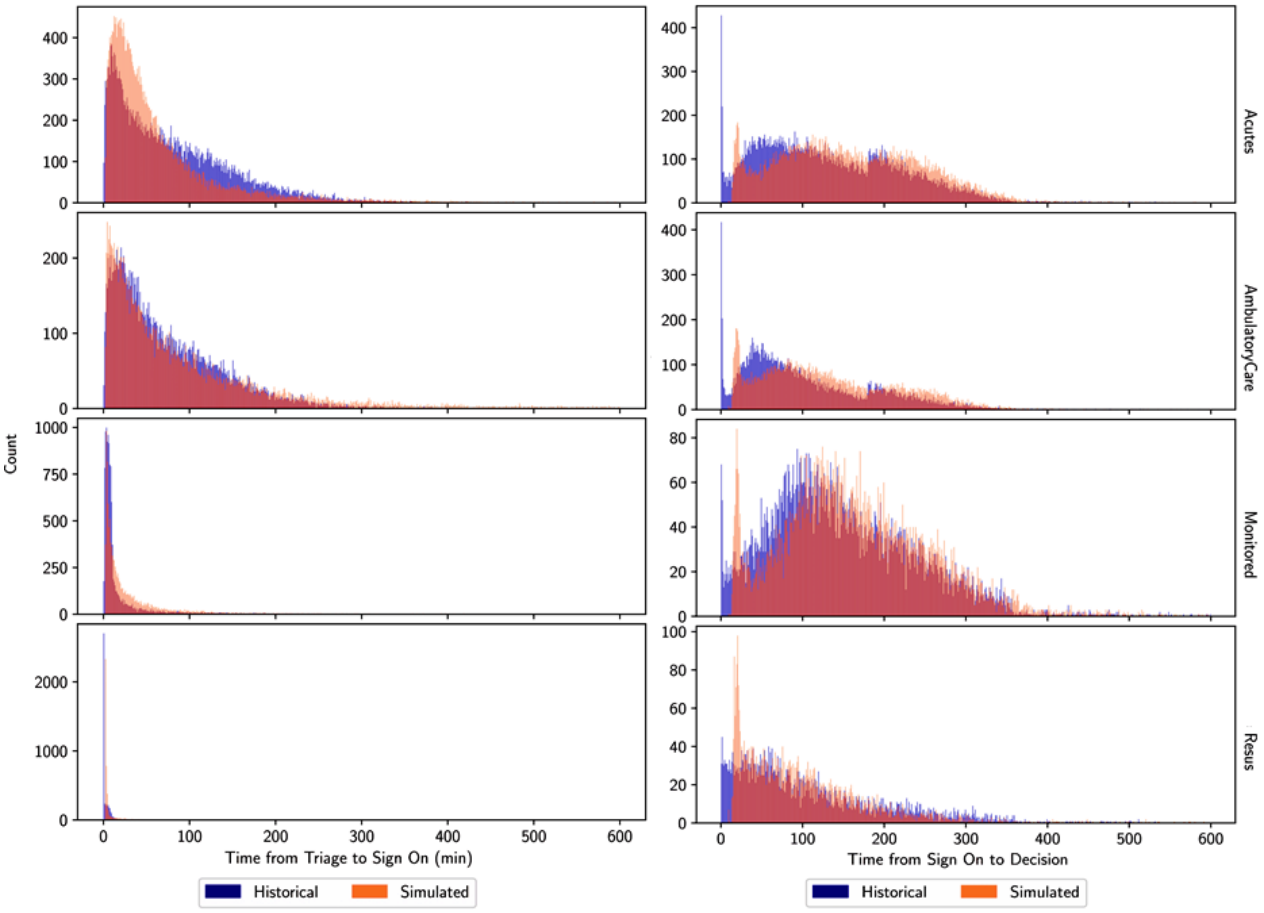

Figure 3.4: Comparison of historical data and simulation results. Left: Mean triage to sign-on time distribution (min). Right: Sign-on to decision time distribution (min). Blue: historical data. Orange: simulation results.

3.5 Conclusions

In this chapter we have explained how we extended and improved the simulation model so as to obtain a valid model. We began explaining the model extensions, which namely involve modifications in the junior/senior advice and paperwork process as well as the sorting of patients. We also determined that three replications were necessary to obtain a 95% confidence with a relative error of at most 5%. We closed the chapter with the verification and validation. We verified the model by using three different techniques. To validate the model we had to introduce some main changes in the logic of the simulation and afterwards we had to fine-tune some parameters. With this, we managed to obtain a valid model according to experts' opinions. Now the simulation model is ready to be used. The next chapters focus on the investigation of priority accumulation, the pods system and a workforce increase.

Chapter 4

Priority Accumulation

In this chapter we focus on researching and implementing priority accumulation, which was suggested as future work by the UoA researchers. In Section 4.1 we review the state-of-the-art regarding priority accumulation. In Section 4.2 we introduce its implementation in the simulation model. In Section 4.3 we discuss about the application of priority accumulation with our settings and, lastly, we close the chapter with Section 4.4 where we summarize our findings.

4.1 Literature Review

Waiting lines in healthcare systems are composed of various types of patients with different needs. To manage such queues, priority queuing disciplines can be used. Acuity rating systems, such as the Canadian Triage and Acuity Scale (CTAS) or the Australian Triage Scale (ATS), are similar to a classical priority queue and are an example in the healthcare emergency field (Sharif et al., 2014). These two systems classify the emergency patients into five classes according to their severity level and each class has a specified performance target. Nevertheless, using such a system leads to only selecting patients of a given priority when there are no more waiting patients of a higher category. This results in low priority patients being overtaken continuously and having to wait long, which is inappropriate as the patient's condition can deteriorate over time (Li et al., 2017). It is thus desirable to seek for a modification that also takes into account other factors so that the service requirements for all patient categories are met.

In 1964 in the context of computer processor design, Kleinrock (1964) suggested a time-dependent priority queue for a single server, where customers are allowed to accumulate priority as a linear function of their waiting time. The rate of accumulation depends on their classification: the higher the urgency of a class, the greater the rate at which a customer from that class accumulates priority. Whenever a server becomes free, the selected customer, provided that the queue is not empty, is the one with the highest accumulated priority at that point in time. Therefore, in a stable queue, a non-urgent customer will eventually accrue enough priority to enter service even in the scenario where more urgent customers are present, and at a earlier point in time than if the customer's waiting time were ignored. Kleinrock's results are a recursive set of formulae to calculate the expected waiting time for each class. He showed that by fine-tuning the accumulation rates, one

can meet the KPIs for each class that might not be met in a classical priority system. One can argue that, for example, medical staff in emergency departments implicitly make use of a similar approach whenever they take into account both the patient's acuity and waiting time in the selection of the next patient to be treated (Sharif et al., 2014).

While Kleinrock (1964) focused on the mean waiting time, Stanford et al. (2014) focuses on the tail of the waiting time distribution for each customer class, since this is usually the one determining the performance of the healthcare queuing system. Stanford et al. (2014) revisited Kleinrock's model, they renamed the time-dependent priority queue to "accumulating priority queue" (APQ) and derived the waiting time distributions for each customer class in a single server setting. Sharif et al. (2014) derived the expressions for the APQ waiting time distribution for each class in a multi-server setting with an application to healthcare. However, their model assumes Poisson arrivals for each class and a common exponential service time distribution. Therefore, this model can only be applied in situations where the treatment duration is similar among patient classes, which is not the case in emergency departments.

In 1967 Kleinrock and Finkelstein (1967) also studied an accumulating priority system in which customers' priorities accumulate as a non-linear function of their waiting time. Such a system can be beneficial in a healthcare setting to reflect that patients' treatment can become even more urgent as they wait longer. Li et al. (2017) extends the analysis of Kleinrock and Finkelstein (1967) by studying the equivalence of non-linear and linear APQs with Poisson arrivals, generally distributed service times and non-linear accumulation functions. Their results hold both for single and multi-server queues.

Healthcare modelers have used a variant of the accumulating priority mechanism for an emergency care setting. Hay et al. (2006) introduced the mechanism that they call "operating priority", which is a single number expressed in terms of an initial priority and a priority accumulation rate. In such a mechanism, patients are given an initial score reflecting their urgency and when they join the waiting queue, the priority increases as they wait. Both the initial priority and the accumulation rate are functions of the patient's class. This differs with any of the previous models by allowing patients to start with a certain nonzero priority upon arrival. They also made use of skills sets, thereby implementing a smart system in which a direct claim is not made for a particular kind of resource. Instead, the model evaluates which is the best resource to be dispatched so that they make best use of the differing skills of junior and senior physicians and balance the use of differing levels of clinical expertise. With both the operating priority and skills sets, they managed to allocate resources depending on the patient's severity, how busy the hospital is and the patient's waiting time. The authors observed that their mechanism better reflected the actual behavior of an emergency department than the classical priority mechanism.

4.2 Priority Accumulation Prototype

We have implemented PA in the simulation model as one of the patient sorting methods, following the mechanism introduced by Hay et al. (2006). To calculate the total priority of a patient at a specific point in time, we use an initial priority and we add up the priority accumulated over the period from triage until the current time. The formula used is as follows:

$$P_t = P0_x + r_x(t - a) \quad , \tag{4.1}$$

where t is the current time, x is the patient's triage code, $P0$ is the initial priority, r is the accumulation rate and a is the triage time.

Both the initial priority ($P0_x$) and the accumulation rates (r_x) are parameters for which the user has to give a numerical input in order to run the simulation. Both of them depend on the triage code, therefore the user needs to specify five values for each parameter. However, deciding on good values for these parameters is an optimization problem on its own. An approach would be to perform several simulations with different values for each parameter, compare the results to the KPIs and choose the values that result in the best KPIs.

The model updates the total priority for each patient every time there is a new patient arrival, a physician becomes free and a patient's results are ready. Even though we have implemented such a method, this is just a prototype and it has not been tested yet. The reason behind it is introduced in the next section.

4.3 Discussion

First of all, let us go back to the priority accumulation motivation. UoA researchers studied the effects of pods and no pods and used a time series to control the advice process between seniors and juniors. This allowed either no SMOs visiting patients or seeing patients and giving advice at the same time, which is not possible in real life. With SMOs not seeing patients at all, or just for a small portion of their time, the system showed an unstable behavior, both for pods and no pods, caused by the server utilization being close to one. Due to the priority system this effect is felt mainly by the low priority patients, whose waiting times appeared to grow exponentially. At first sight, it seemed reasonable to improve the long waiting times of low priority patients by means of applying PA and, therefore, we suggested its implementation.

Note that PA will not change the throughput of the ED, but only the order in which patients are seen. Therefore, if there is a lack of staff to meet demand, and in consequence the system is unstable, PA will not be the solution to the problem, as it can only be applied to stable systems (Li et al., 2017). By using a different controller than the time series to manage the senior/junior advice process, the waiting times for low priority patients do not appear to grow exponentially anymore in the no pods system. Nevertheless, as we will see in Chapter 5, the pods system - with the same amount of staff as no pods - shows a similar unstable behavior to the previous research due to the unpooling of resources. Therefore, PA will not be a solution now either.

In order to achieve the desired KPIs, we need to find the optimal values for the initial priorities and the accumulation rates. These values depend on the amount of physicians available in each pod. Taking as an example two pods with the same patient mix, patient volume and bed capacity but different amount of physicians, low priority patients will be overtaken more often in the pod with less physicians. In order to avoid this, it is necessary to adjust the initial priorities and accumulation rates, thereby resulting in different values for each pod. We believe that having such a sensitive approach is troublesome and, instead, it is preferable to implement the sorting methods introduced

in Chapter 3, as these are likely to be more robust to changes in patient mix or allocation of resources.

Because of the aforementioned reasons, the reminder of this thesis focuses on having a proper allocation of the available resources and investigates the effects of unpooling them, rather than applying PA.

4.4 Conclusions

We started this chapter reviewing the state-of-the-art regarding priority accumulation. We saw the work of several authors on implementing PA in healthcare in order to prevent low priority patients from being overtaken. Next, we developed a PA prototype for our simulation model, which sorts patients based on a priority. This priority depends on the initial patients' acuity and the amount of time they wait since triage. Nevertheless, we decided to not carry on the research on PA due to two reasons. First, our pods system with the original amount of physicians was showing an unstable behavior and, therefore, we cannot apply PA as it is only suitable for stable systems. Second, to the best of our knowledge, the initial priorities and accumulation rates are contingent on the available physicians in each pod. Hence, it is more desirable to use other patients sorting methods that are more robust.

Chapter 5

Staffing Model for a Pods System

This chapter deals with the pods system and a staffing model to allocate physicians to pods. We start with a review of the state-of-the art work on pods and team-based work in Section 5.1. We continue with the description of the solution approach in Section 5.2, where we use a simulation model and a Mixed Integer Linear Program (MILP). In Section 5.3 we introduce the staffing model (MILP model), show its formulation, explain the results and we finish describing its contributions and limitations.

As a reminder, in this thesis we consider pods to be specific geographic areas that accommodate specific groups of patients whose care is provided by dedicated teams of physicians and nurses.

5.1 Literature Review

This literature review focuses on the state-of-the-art analysis of pods, team-based work and unpooling of resources.

To begin with, "Pod" is an ambiguous concept and is referred to by several names such as: pod, pod/s system, podular system or team-based pod system. Some literature regarding ED layout makes use of the pod concept as cluster of rooms physically separated from the rest (Pati et al., 2014), while other scholars describe it as a geographic division of the rooms, without remarking any physical separation (Torrence Memorial 2014, Melton III et al. 2016, Gavin and Peterson 2017, Morgareidge et al. 2014). All of them agree that each pod should be staffed with a team composed of physicians and nurses. Nevertheless, only Pati et al. (2014) and Morgareidge et al. (2014) explicitly mention the use of a decentralized nurse station in each pod. Furthermore, each pod is meant to serve a focused group of patients, which can be dedicated to a specific patient population (such as pediatric) or acuity (depending on ESI levels (ESI, 2004)), and patients are assigned to a pod upon arrival. Although pods are focused on a specific patient group, Melton III et al. (2016) and Torrence Memorial (2014) emphasize the fact that each of their pods is equipped to handle any type of patients, therefore allowing for flexibility. Further, in all cases, the patient volume per hour determines the opening and closing times of the pods, the number of pods open and staff.

In an effort to improve the EDs' performance Melton III et al. (2016) and Morgareidge et al.

(2014) present major changes in the EDs, alongside a transition to a pods system. Some of these modifications consist of adding extra laboratory and radiology resources, changes in the IT, system or standardization of processes. Carrying out many changes simultaneously makes it difficult to determine the effects of implementing pods, even though Melton III et al. (2016) believe that pods are one of the key changes with one of the greatest impacts in his redesign. Both studies report significant performance improvements. While Morgareidge et al. (2014) show an average LOS decrease of 44.5%, Melton III et al. (2016) achieve an increase from 46.5% to 81.4% in the percentage of patients spending at most 3 hours in the ED. On the other hand, Torrence Memorial (2014) and Gavin and Peterson (2017) only mention the transition to a pods system. Torrence Memorial (2014) reports an average decrease in door-in-to-doctor-time from 79 to 39 minutes and a median LOS decrease of 1 hour and 18 minutes. Gavin and Peterson (2017) achieve, for all patients with ESI levels 4 and 5, a LOS under 3 hours. So far, all the results have shown performance improvements. Only Pati et al. (2014) discuss the fact that pods becomes inefficient when just a few beds/rooms are occupied in a pod, leading to idle medical staff whose help could be useful in a pod with a higher workload.

The above mentioned studies report several advantages of the pods system. First, the physicians are able to stay within closer proximity to their patients and work with a smaller and more specific team of nurses, which fosters improved communication (Gavin and Peterson 2017, Torrence Memorial 2014). Second, the assignment of a patient to a pod upon the patient's arrival, and the possibility of the pod coordinator to track the patients in the waiting room belonging to his/her pod, promotes accountability among team members. This results in a lower likelihood of leaving patients unattended (Torrence Memorial, 2014). Third, the pods configuration reduces the number of patient care transfers, which leads to a minimization of potential medical errors (Torrence Memorial, 2014). Lastly, having nurse stations in each pod improves nurse-patient visibility and reduces the walking distances between nurses and patients (Morgareidge et al., 2014).

Only the research conducted by Pati et al. (2014) analyses the disadvantages of the pods system. First, pods may obstruct the visibility across acuity zones due to walls separating different patient care areas. This presents an obstacle in emergency times when additional hands and resources are much needed and could be quickly delivered if visibility were not restricted. Second, obstructed visibility can be counter-productive for physicians. This problem becomes evident in the case of having small pods, which can lead to perceived and real isolation by physicians as well as impede them from teaming up with other ED medical staff. Third, the pod system presents a logistical challenge for support departments (pharmacy, dietary, documentation, etc.) to keep track of the patients' location in the different pods. In addition, nurse stations are typically decentralized in a pods system, whereas support departments are not, which entails additional walking distances.

The pods system implies having teams of physicians and nurses exclusively working together throughout the duration of their shifts. Other literature (Dinh et al. 2015, Lau and Leung 1997) uses the team-based concept without mentioning the implementation of pods. Before applying team-based care, patients were put into a common pool after triage and visited by any physician. This lead to (1) patients waiting long times as there was poor accountability among physicians and (2) poor job motivation as efficient work simply brings along the penalty of more work. Under team based-care, several teams of three to four physicians are formed – neither Lau and Leung (1997) nor Dinh et al. (2015) specify the amount of nurses, if any – and patients are assigned to the teams upon triage. Lau and Leung (1997) evenly assign patients to each team in terms of number and complexity, whereas each of the teams described by Dinh et al. (2015) serves a focused group of

patients.

Both, Lau and Leung (1997) and Dinh et al. (2015) evaluate the impact on the ED performance of introducing team-based care. While the results of the study conducted by Dinh et al. (2015) show a significant 17% increase on the daily ED presentations that left the ED within 4 hours, Lau and Leung (1997) significantly reduce the average triage-to-doctor time from 35 to 22 minutes. They attribute these improvements to an increase in job motivation and a clearly defined responsibility for patient care by assigning patients to teams from the time of arrival. Nevertheless, Lau and Leung (1997) mention the difficulty of managing separate queues in a common space as patients in different queues compare to each other.

In the end, teaming up medical staff, regardless of using pods, comes down to pooling and unpooling human resources. Pooling is an operations management technique suggested to reduce the negative effects of demand variability in order to enhance performance. Through analytical models, scholars have shown that pooling separate demand streams of similar customer types served by similar servers enables shorter waiting times and higher average utilization (Song et al., 2013). However, pooling customers with very different characteristics can lead to inefficiencies (Song et al., 2015). Song et al. (2013) discuss that pooling can become appropriate in non-discretionary work settings such as factories, where highly specified routines and tasks are involved. However, EDs are discretionary work settings, where workers have control on work content, pace and resources.

Song et al. (2013) evaluate how pooling tasks (having patients in a same pool) and/or resources (all physicians working with all nurses, technicians, machines, etc.) affects the throughput time. Song et al. (2015) extend this research by adding a fairness constraint and also investigating the pooling effects on the waiting times. Both of their results suggest that (1) assigning a patient to a physician upon the patient's arrival (unpooled tasks) and (2) having each physician working with a dedicated team of two nurses throughout the shift (unpooled resources) yield shorter throughput and waiting times when compared to pooling. These results differ from the ones obtained with queuing theory in non-discretionary work settings, which suggest that pooling yields shorter throughput times.

The reduction of throughput time with dedicated physicians (unpooled tasks) comes from fostering an ownership physician behavior. Song et al. (2015) mention some of the different practice patterns: (a) physicians proactively pull for results instead of waiting for information to be pushed, (b) they work together with nurses to better coordinate tasks, (c) they start sooner the discharge process of those patients who are ready to leave and (d) they make sure that patients in the waiting room are placed in a bed as soon as one of their beds becomes free. Moreover, the shorter throughput time with dedicated resources is due to a distributed utilization of shared resources as this impedes fast-working physicians to over-utilize shared resources (Song et al., 2013). Finally, Song et al. (2015) believe that shorter waiting times are attained with unpooled tasks and resources due to (1) physicians proactively initiating the placement of their patients in the waiting room into their beds, instead of the triage nurse placing the next patient in an open bed, and (2) due to an indirect queuing effect as patients being treated have shorter LOS, which makes beds in the ED available sooner and, in turn, upcoming patients have to wait less.

To conclude, in this section we have discussed about the state of the art regarding pods, team-based care and pooling of tasks and resources. We have seen that the pods concept is not clearly defined in the literature and there is little research about it. Nevertheless, all the studies that implemented pods show an ED performance improvement. One of the main reasons for such

an enhancement is the team-based work, which is a concept we have also seen in other literature not concerning pods. Most of the studies presented in this section conclude that working in teams engenders a staff behavior that contributes to a performance improvement as the work is clearly divided and the responsibilities are clearly defined.

5.2 Solution Approach

The solution approach of this research involves the employment of a simulation model and a Mixed Integer Linear Program (MILP) as key tools.

To begin with, we want to investigate the effects on the ED performance of implementing pods or not. In order to study what-if scenarios, we have developed and extended a simulation model, which we have already introduced in chapters 1 and 3. By setting different configurations in the ED simulation model, we will be able to draw conclusions for management suggestions. In order to properly simulate a pods system, the simulation requires the establishment of physician teams. For this purpose, we develop a staffing model that allocates as efficiently as possible the available ED workforce into pools for each pod.

In short, the staffing model's output is an input to the simulation model. It allocates physicians into pools in the following manner: with the no pods system, the model allocates physicians into two pools; with pods, into three. As a result, we obtain a matrix that tells us which doctor is working where at what time. We then translate this matrix into JaamSim time series, which specify the location and working hours, and input it to the simulation model.

5.3 Staffing Model

5.3.1 Introduction to the Model

The staffing model allocates the ED physician shifts into different pools corresponding to each existing pod and it does so through two phases that complement each other. The objective of this allocation is to best distribute a given number of physicians among pods. We do not focus on determining whether this given total number of physicians is optimal to face the ED's demand or whether it will cause the ED to be over or understaffed. This staffing model is a Mixed Integer Linear Program.

To begin with, let us explain what "ED physician shift allocation" means. The ED provided a list that specifies the working times of each physician role (SMO, Registrar, HO...). Moreover, there are several repetitions of each role and they can work either in the morning, afternoon or at night. This list is defined for each day of the week for an entire week, and all weeks are assumed to be the same throughout the year. For instance, 6 Registrars work from Monday to Sunday, from which 2 work in the morning, 2 in the afternoon and 2 at night. Each of these Registrars does not represent a physical person, as the same person cannot work 7 days in a row, but it represents what we call a "shift". Therefore, the staffing model allocates physician shifts to pools instead of physical people. Nevertheless, for the sake of simplicity we will refer to these "shifts" simply as ED physicians. See

Appendix B (subsection B.1.1) for more details.

As mentioned before, the model does the allocation in two phases:

- Phase 1: in this phase we focus on improving the pods' bottlenecks in terms of staff available by reallocating – if possible – physicians at the hour in each pod where the lack of staff is the most severe. By doing so we improve the worst-case scenario, even though we do not pay attention to how we allocate the staff in the remaining non-bottleneck hours. This is where phase 2 plays an important role.
- Phase 2: it focuses on allocating as efficiently as possible the ED physicians into the pools in order to balance the workload among pods. Moreover, phase 2 uses as input the improved bottlenecks resulting from phase 1, thereby making its solution equal or better than phase 1.

The allocation of physicians depends on the system being used. When using the no pods system, the model allocates them into two pools, whereas with the pods system it allocates them into three pools. Remember that we refer to pods and pools indistinctively as each pod owns a pool of physicians. Furthermore, in reality physicians stay in the same pod, unless stated otherwise. In the model we allow them to change pods to give flexibility in order to better match demand and supply. However, we restrict them to change at most 7 times per week because we aim at having one change per day.

In the model, we fix the allocation of physicians in Resus at two physicians during day time hours and one during night time hours. These limits are both upper (UB) and lower (LB) bounds. We allocate two physicians because the ED management wants to ensure a rapid response to incoming resuscitation patients during day time hours, assuming the risk of any of these two physicians becoming idle due to low workload in Resus. Moreover, we just allocate one physician in Resus during night time hours because the management does not want to run the risk of having idle physicians in Resus at night when they are much needed in other pods. However, if there is a lack of physicians in Resus - at any time - a call will be made to physician of other pods.

Since the allocation of physicians in Resus is always the same, the remaining physicians go to General in the no pods system. Hence, it makes no sense to apply phase 2 to the no pods system, as there will be no difference between the allocation of phase 1 and phase 2. Phase 2 would be applicable for no pods in the case where the amount of physicians available were large enough to make the ratios physicians/patients similar across pods, and thus, more physicians would be allocated in both pods. On the other hand, the fixed allocation in Resus is why, with the pods system in phase 2, we only balance the workload between the General and the Ambulatory pods.

5.3.2 Input

In this subsection we describe the input that the model requires for its parameters. For further details on any of the inputs, see Appendix B, section B.1.

To begin with, the ED actually provided two lists of physicians shifts. The first one consists of the physician shifts used with the no pods system, with a total of 24 shifts (see Table B.1 in Appendix B). In this one, the pod where each physician is assigned to is indicated. We refer to this

first list as the "original roster". The second list, which has a total of 28 shifts, was suggested by the ED for the pods system, but does not assign physicians to pods (see Table B.2 in Appendix B). We refer to this second list as the "extended roster". Regardless of which system each list is for, we use them in both systems to see the differences between having more or less staff capacity.

Since we wanted to allocate the ED physicians into pods according to demand, we found it necessary to analyze the number of beds occupied per pod per hour of the week. Such analysis is fundamental as not only the amount of patients admitted is enough, but their length of stay also needs to be taken into account. For a more detailed explanation, see Appendix B. Furthermore, we needed to know the amount of patients that a physician can take care of at the same time. This, together with the demand, determines how many of the available physicians are to be allocated to each pod. For this, we contacted the ED staff, who estimated the average patient-physician ratios. These ratios differ across physician roles and pods.

Another input to the model is the pods where each physician role is allowed to be in, according to the ED rules, presented in Chapter 1. An important rule is to have an SMO in each pod, implying that SMOs are allowed in all the pods. We control this rule by already assigning each SMO to a pod in the parameters that we input. Therefore, this rule is not modeled as a mathematical constraint. Lastly, we use pod weights for the objective function of phase 1, which are proportional to each pod's patient volume.

5.3.3 Model Formulation

In this subsection we present the model generalization. We make use of P to indicate the set of pods being used, which can either refer to the no pods system or the pods system. Beware that some constraints only operate under the pods system. We also deploy D to indicate the set of physicians being used, which can either be from the original or the extended rosters.

Phase 1

Indices and sets

p	index of pods
d	index of doctors
h	index of hours
P^{NP}	set of pods = {General, Resus}
P^{P}	set of pods = {General, Resus, Ambulatory}
P	current set of pods being used, either P^{NP} or P^{P}
D^{OR}	set of doctors from the original roster
D^{ER}	set of doctors from the extended roster
D	current set of doctors being used, either D^{OR} or D^{ER}
D^{HO}	subset of House Officer (HO) doctors, $D^{HO} \subset D$
H	set of hours in a week
H^{R2}	subset of hours for which two doctors are needed in Resus, $H^{R2} \subset H$
H^{R1}	subset of hours for which only one doctor is needed in Resus, $H^{R1} = H - H^{R2}$

Parameters

$w_{h,d}$	1 if doctor d is working at hour h
$O_{h,p}$	number of beds occupied at hour h in pod p
$x_{d,p}$	number of patients that doctor d can handle at the same time in pod p
$A_{d,p}$	1 if doctor d is allowed in pod p
g_p	weight of pod p

Variables

$b_{p,h,d}$	1, if doctor d is in pod p at hour h. 0, otherwise
$C_{h,p}$	patient capacity: amount of patients that can be handled at the same time at hour h in pod p
$S_{p,h,d}$	1, if doctor d changes pods at hour h. 0, otherwise
y_p	LB of the $C_{h,p}/O_{h,p}$ ratio in pod p

Model

$$\text{maximize} \quad \sum_{p \in P} g_p \, y_{1p} \tag{5.1a}$$

subject to

$$C_{h,p} = \sum_{d \in D} b_{p,h,d} \, x_{d,p} \quad , \quad \forall h \in H, p \in P, \tag{5.1b}$$

$$y_p \leq \frac{C_{h,p}}{O_{h,p}} \quad , \quad \forall h \in H, p \in P, \tag{5.1c}$$

$$\sum_{p \in P} b_{p,h,d} = 1 \quad , \quad \forall h \in H, d \in D \mid w_{h,d} = 1, \tag{5.1d}$$

$$-S_{p,h,d} \leq b_{p,h,d} - b_{p,h-1,d} \quad , \quad \forall p \in P, h \in H, d \in D, \tag{5.1e}$$

$$S_{p,h,d} \geq b_{p,h,d} - b_{p,h-1,d} \quad , \quad \forall p \in P, h \in H, d \in D, \tag{5.1f}$$

$$\frac{1}{2} \sum_{p \in P} \sum_{h > 0 \ \in H} S_{p,h,d} \leq 7 \quad , \quad \forall d \in D, \tag{5.1g}$$

$$\sum_{d \in D} b_{p,h,d} = 2 \quad , \quad \forall h \in H^{R2}, p = \{Resus\}, \tag{5.1h}$$

$$\sum_{d \in D} b_{p,h,d} = 1 \quad , \quad \forall h \in H^{R1}, p = \{Resus\}, \tag{5.1i}$$

$$\sum_{d \in D} b_{p,h,d} \geq 1 \quad , \quad \forall h \in H, p \in P - \{Resus\}, \tag{5.1j}$$

$$\sum_{d \in D} b_{p,h,d} - \sum_{d \in D^{HO}} b_{p,h,d} \geq 1 \quad , \quad \forall h \in H, p = \{General\}, \text{ only if } P = P^P, \tag{5.1k}$$

$$b_{p,h,d} \in \{0,1\} \quad , \quad \forall p \in P, h \in H, d \in D \mid w_{h,d} = 1 \ \& \ A_{d,p} = 1, \tag{5.1l}$$

$$C_{h,p} \in \mathbb{R}^+ \quad , \quad \forall p \in P, h \in H, \tag{5.1m}$$

$$S_{p,h,d} \in \{0,1\} \quad , \quad \forall p \in P, h \in H, d \in D \mid w_{h,d} = 1 \ \& \ w_{h-1,d} = 1, \tag{5.1n}$$

$$y_p \in \mathbb{R}^+ \quad , \quad \forall p \in P \tag{5.1o}$$

The objective function (5.1a) maximizes the weighted minimum ratio of the number of patients that can be handled to the number of beds occupied, over all pods. Constraint (5.1b) defines the amount of patients that can be handled at the same time in a pod and hour. Constraint (5.1c)

ensures that the lower bound of the ratio $C_{h,p}/O_{h,p}$ is found. Constraint (5.1d) not only makes sure that physicians cannot be in more than one pod at the same time, but also forces them to be assigned to a pod if they are working. Constraint (5.1g) sets the maximum number pod changes per physician per week to 7. Constraints (5.1e) and (5.1f) allow finding the absolute value of $b_{p,h,d} - b_{p,h-1,d}$. Constraints (5.1h) and (5.1i) restrict the amount of physicians in Resus to 2 and 1 respectively. Constraint (5.1j) makes sure that there is always a physician in each pod (except for Resus as it is already restricted in the two previous constraints). Constraint (5.1k) is only used for the pods system and it ensures that HO physicians are not alone in the General pod. Constraints (5.1l)-(5.1o) specify the variables' domains. Note that variables $b_{p,h,d}$ and $S_{p,h,d}$ only exist under certain conditions defined in constraints (5.1l) and (5.1n) respectively.

Phase 2

The parameters, variables and constraints used in phase 2 are the same ones as in phase 1, except for parameter g_p, variable y_p and constraints 5.1c and 5.1o. The output value of y_p in phase 1 is input as a parameter with name z_p in phase 2. The new parameters and variables phase 2 deploys are:

Parameters

z_p	LB of the $C_{h,p}/O_{h,p}$ ratio in pod p

Variables

V_h	distance between $C_{h,p}/O_{h,p}$ in General and $C_{h,p}/O_{h,p}$ in Ambulatory Care at hour h
L_h^+	first slack variable for the absolute value of V_h
L_h^-	second slack variable for the absolute value of V_h

Model

$$\text{minimize} \quad \sum_{h \in H} L_h^+ + L_h^- \tag{5.2a}$$

subject to

$$V_h = \frac{C_{h,Ambulatory}}{O_{h,Ambulatory}} - \frac{C_{h,General}}{O_{h,General}} \quad , \quad \forall h \in H, \tag{5.2b}$$

$$\frac{C_{h,p}}{O_{h,p}} \geq z_p \quad , \quad \forall h \in H, p \in P, \tag{5.2c}$$

$$V_h \leq L_h^+ \quad , \quad \forall h \in H, \tag{5.2d}$$

$$-V_h \leq L_h^- \quad , \quad \forall h \in H, \tag{5.2e}$$

$$V_h \in \mathbb{R} \quad , \quad \forall h \in H, \tag{5.2f}$$

$$L_h^+, L_h^- \in \mathbb{R}^+ \quad , \quad \forall h \in H \tag{5.2g}$$

The objective function (5.2a) minimizes the absolute distance between the $C_{h,p}/O_{h,p}$ ratios in the Ambulatory Care and General pods and constraint (5.2b) defines such distance. Constraint (5.2c) makes sure that the bottlenecks of phase 2 are not worse than the output ones from phase 1. Constraints (5.2d) and (5.2e) allow finding the absolute value of the distance. Constraints (5.2f)

and (5.2g) define the variables' domains.

5.3.4 Results and Discussion

The MILP model is coded in AIMMS version 4.33.3.898 and solved with CPLEX 12.7. Both phases for the no pods system and phase 1 for the pods system are solved to optimality with a tolerance of 0%. Due to the large amount of variables and constraints in phase 2 for the pods system, we set the tolerance to 0.5% in order to get a solution in a reasonable amount of time. Moreover, we believe that such a small gap between the LP bound and the best solution will not have a significant effect on the simulation model later on.

Figures 5.1 and 5.2 show the results of the MILP model for each system with the original roster. The bars represent the number of beds occupied (model input), whereas the dashed red line and solid dark blue line represent the amount of beds that can be handled given the allocation of ED physicians by phase 1 and 2 respectively (model output). It is worth mentioning that the three peaks that are observable for every day of the week in the general pod show the overlaps between the night, morning and afternoon shifts. Thus, we can practically omit these peaks as they only last for one hour and their usage would be for passing information on to the next shift.

First, we focus on Figure 5.1. A first observation is how clear the general pod supply is much smaller than demand. This can either hint that the patient-physician ratios are not well estimated or that there is a clear lack of personnel. If the case is the former, the hospital estimated the number of patients a physician can simultaneously take care of; this values differ depending on the pod and the rank of the physician. Should the hospital estimates be wrong, our model would still be valid as long as the estimates are at least proportionally correct. That is, the blue and red lines would simply be shifted upwards, but the shape would not change. If the case is the latter, as we already mentioned, our goal is not to determine whether the total number of given physicians is optimal or not, but to distribute the available resources as well as possible. Another observation is how phase 1 and 2 result in the same allocation for the no pods system, as explained in the model introduction. Focusing on Figure 5.2, we can see the obvious need for phase 2 in the pods system. While phase 1 focuses on the bottlenecks, phase 2 ensures a more fair distribution of physicians by better matching supply and demand than phase 1.

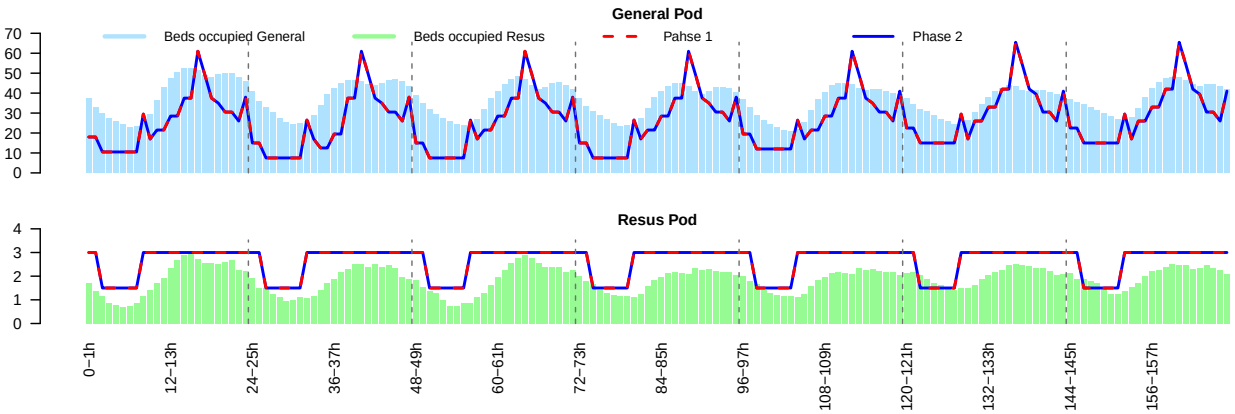

Figure 5.1: Allocation of ED physicians in the no pods system using the original roster

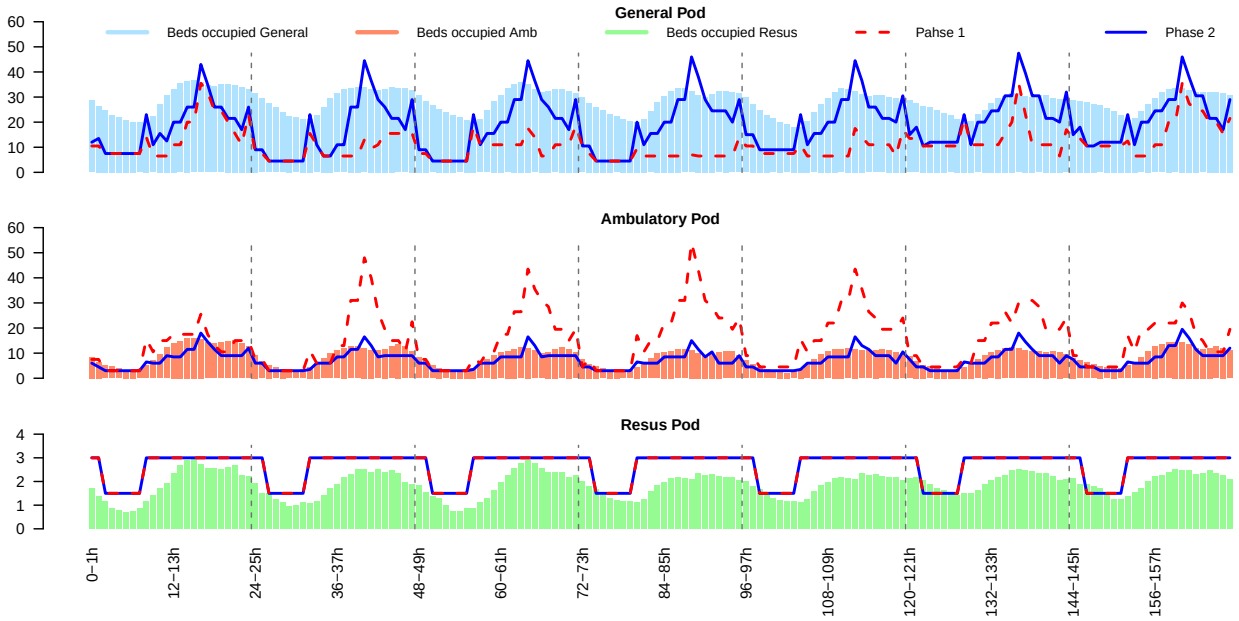

Figure 5.2: Allocation of ED physicians in the pods system using the original roster

We now move on to the results attained with the extended roster. The results from phase 1 for the pods system and the extended roster emphasize issues that we overlooked with the original roster. The LB of the C/O ratio for the Ambulatory pod results to be 0.96, a very high LB. We explain the causes behind such a high LB in Section 5.3.5. To clarify, C/O ratios below 1 indicate a lack of staff, whereas above 1 suggest a surplus, meaning that a ratio of 0.96 essentially equals supply and demand. Therefore, phase 2 is forced to allocate a significant amount of staff to the Ambulatory pod in order to keep such a high LB, thereby leaving the General pod considerably understaffed compared to how it could be with a lower Ambulatory LB. We can observe this issue in Figure 5.3. We believe that this problem, originating from phase 1, can have a serious impact on the simulation results and, in order to mitigate its effects, we have decided to use the LB of the C/O ratio for the Ambulatory pod resulting from phase 1 of the *original roster*. The value we are referring to is 0.56, which will allow a more fair distribution of staff between the General and Ambulatory pods (see Figure 5.3, dashed red line).

The allocation results are inputted to the simulation model by means of translating the output of the variable $b_{p,h,d}$ using R-studio (version 3.2.3), into a set of time series.

5.3.5 Limitations and Contributions

The first limitation concerns the fact of using the two described phases. When we first designed the model, we considered that focusing on the bottlenecks was of importance. When analyzing the results of the original roster, the need for a second phase became clear to us. However, when analyzing the results of pods and the extended roster, we observed we were not achieving our objective as well as we wanted. These results can be explained by the formulation of phase 1. It is designed in such a way that prioritizes the allocation of physicians to the easiest pod to get the highest C/O ratio, resulting in very high lower bounds for some pods and very low for others. This

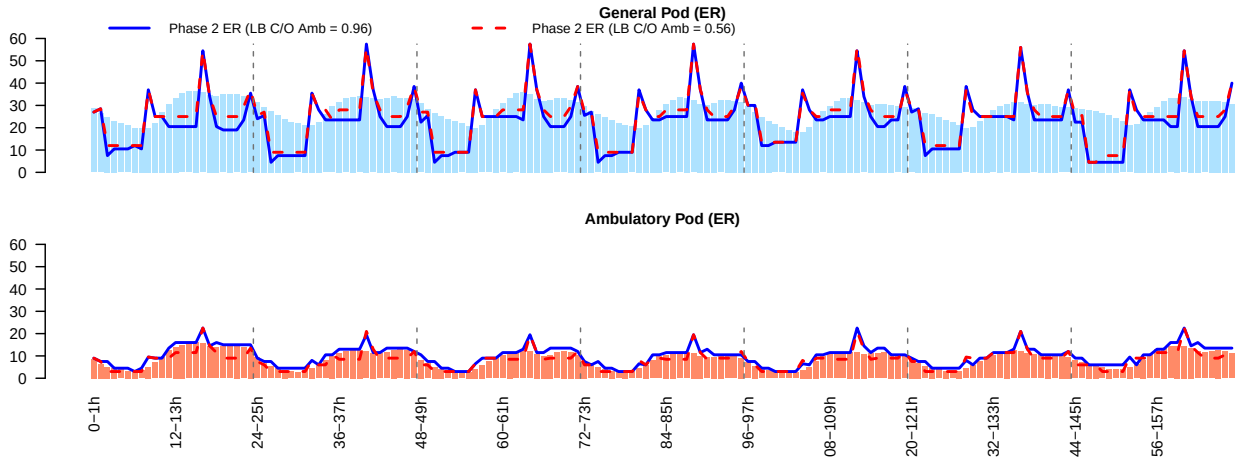

Figure 5.3: Resulting allocation from phase 2 with two different LBs for Ambulatory. Uses the extended roster (ER)

clearly constraints the results of phase 2. While phase 2 could allocate physicians such that the difference between C/O ratios among pods would be quite small, quite often it does not do so as it is restricted by the high LB determined by phase 1. For a numerical example see section B.2 in Appendix B. Two potential solution would be to penalize C/O ratios above 1 in phase 1 or to try how phase 2 works on its own, without running phase 1 first. Trying both solutions requires additional experiments and time, which we suggest as future research.

The second limitation involves one of the model constraints. We allow physicians to change pods at most 7 times per week, with the aim of having at most one change per day. First, it can happen that all 7 changes happen in one day, as we do not define a day as a time period. Second, the number of changes is an optimization problem on its own. One the one hand, more changes yield more flexibility to match demand and supply, while on the other hand, they result in an impractical allocation in real life. In our case, it happens quite often that a physician starts working in one pod, after a while changes to another pod for just one hour and, then, he or she comes back to the initial pod. We believe that this can cause inefficiency issues in real life.

Even though the model has some limitations, we should not underestimate its contribution in practice, as it can serve as a basis for new allocations for rosters, significantly saving human time and avoiding human errors.

5.4 Conclusions

We started this chapter reviewing the state-of-the-art concerning pods and team-based work. Most of the literature concluded that using pods with physicians and nurses working in small teams improved the efficiency. The biggest contributor to this enhancement is the physicians' increased ownership over patients and work. In order to create teams of physicians, we decided to develop a two-phase MILP model. This model allocates the available ED physicians into pods as efficiently as possible, based on the number of beds occupied in each pod. While phase 1 focuses on the bottlenecks, phase 2 focuses on balancing the workload among pods. The output of phase 1 strongly

constraints the results of phase 2, and not always for good. Therefore, we can ascertain that there is room for improvement by changing the objective of phase 1 or, maybe, completely removing this first phase in order to achieve a better allocation. Despite its limitations, this model is a strong allocation tool that can serve as a basis for future rosters, saving planners a considerable amount of time. The results of the staffing model are inputted to the simulation model, whose results are presented in the next chapter.

Chapter 6

Analysis of Results

This chapter analyzes the difference between the no pods and pods systems and how these two behave when adding extra workforce. Section 6.1 defines the design of experiments. Next, Section 6.2 presents and discusses the results as well as compares them to the literature. Last, the chapter finishes with the conclusions in Section 6.3.

6.1 Experimental Design

The experimental design of this research (Figure 6.1) consists of a total of 4 experiments that result from varying 2 experimental factors with 2 levels each:

- Number of pods: as we have been explaining, we want to see the difference between having pods or not in the ED. While in the no pods system physicians are divided into two pools (2 pods), in the pods system they are divided into three pools (3 pods).
- Amount of physician shifts: as explained in the previous chapter, the ED provided two rosters, the first one with a total of 24 physician shifts (referred to as the original workforce) and the second with a total of 28 (referred to as extended workforce).

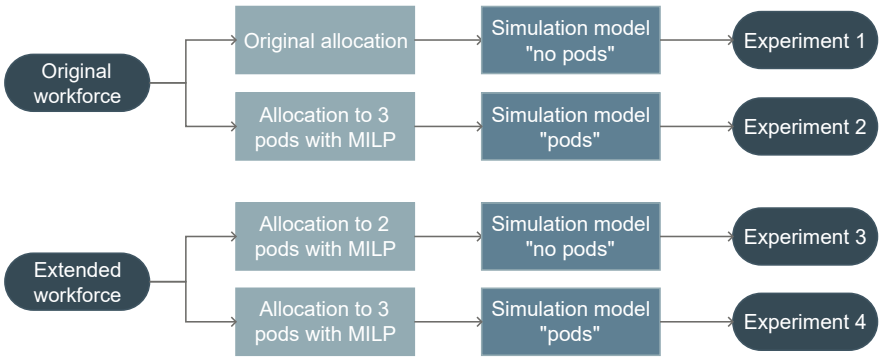

Figure 6.1: Experimental design

Note that experiment 1 uses the allocation given by the original roster (see Table B.1 in Appendix B). For the rest of the experiments, we need to apply use the staffing model to obtain an allocation of physicians to pools. Moreover, we make use of common random numbers in order to reduce the variance between the same replication of different experiments and, thus, get smaller confidence intervals when making comparisons. To summarize, the settings of each experiment are:

- Experiment 1: no pods, original workforce
- Experiment 2: pods, original workforce
- Experiment 3: no pods, additional workforce
- Experiment 4: pods, additional workforce

6.2 Results and Discussion

6.2.1 Results at and ED Level

We start by analyzing the results at an ED level. Table 6.1 shows the overall ED results for each KPI and experiment. Table 6.3 displays the Confidence Interval (CI) of the experiments' pairwise comparisons, showing that all the observed differences between configurations are statistically significant. When comparing two experiments, we can see in Table 6.1 how the three indicators simultaneously show either an improvement or deterioration. Therefore, we can easily rank the experiments from best to worst in the following order: 3, 4, 1, 2. First, this order suggests that adding extra staff results in shorter waiting and throughput times as well as a higher fraction of patients leaving the ED within 6 hours. Second, it hints that the no pods system leads to a better performance compared to the pods system.

Table 6.1: Overall ED results[a] per KPI and experiment.

KPI	Exp.1	Exp.2	Exp.3	Exp.4
Mean triage to sign-on time (min)	69.6	231.9	32.3	40.2
95^{th} percentile of the total time[b](min)	458.5	1409.3	349.2	371.8
Fraction of patients leaving the ED within 6h	0.87	0.69	0.95	0.93

[a] Each of the values is the weighted sum of the four main services' results. The used weights account for the four main services' patient volumes (see Appendix B, Figure B.5, "4 main services normalized" plot).

[b] As we use a weighted sum, we cannot claim that the 95th percentile of the total time truly represents the 95th percentile in the ED, but it serves as a guidance.

6.2.2 Results at a Service Level

We move on to analyzing the results at a service level. Table 6.2 shows the point estimates of each KPI per service and experiment. The CIs of the pairwise comparisons of these point estimates are presented in Table 6.4. It is interesting to see in Table 6.2 how the aforementioned order of experiments applies for all services but Ambulatory Care, whose order is: 3, 4, 2, 1.

Table 6.2: Point estimates of each KPI per service and experiment.

	Mean triage to sign-on (min)				95^{th} percentile total time (min)				Frac. patients leaving within 6h			
	Exp.1	Exp.2	Exp.3	Exp.4	Exp.1	Exp.2	Exp.3	Exp.4	Exp.1	Exp.2	Exp.3	Exp.4
Acutes	68.1	416.1	36.5	51.7	453.5	2449.4	368.1	407.1	0.86	0.48	0.94	0.90
Amb. Care	111.8	68.5	38.2	38.9	549.7	411.6	327.3	330.6	0.85	0.92	0.97	0.97
Monitored	40.0	82.3	22.1	24.5	413.4	603.3	358.1	373.8	0.90	0.79	0.95	0.94
Resus	16.4	18.6	14.0	14.7	307.9	345.9	285.3	288.2	0.97	0.96	0.98	0.98

First, we start comparing the transition from no pods to pods while keeping the original workforce (Exp. 1 & 2). While such a transition leads to a noteworthy performance enhancement in Ambulatory Care, the Acutes and Monitored services experience a notable performance deterioration. Checking Table 6.4 (columns 3, 6 and 9; rows 3, 6, 9 and 12), Ambulatory Care experiences an average decrease in the triage to sign-on time and in the 95th percentile of the total time of 43 minutes and 138 minutes (2h 18 min) respectively, and a significant 0.07 average increase in the fraction of patients that leave within 6 hours. On the contrary, Acutes experiences an average increment of approximately 348 minutes (5h 48 min) and 1996 minutes (33h 16 min) in the first two KPIs and an average reduction of 0.38 in the third KPI. The triage to sign-on time for Resus worsens by an average time of 2 minutes, whereas the other metrics do not show a statistically significant difference. In our opinion, Ambulatory Care shows such a notable performance enhancement due to having its own dedicated pool of physicians. This dedicated pool prevents them from being overtaken by higher priority patients like Acutes and Monitored. Nevertheless, the downside is that these last two services strongly notice the effects of having a reduced pool of physicians.

Second, we continue by comparing the transition from no pods to pods while adding extra workforce in both systems (Exp. 3 & 4). While we expected a similar behavior as in the previous transition, to our surprise, this one results in a worse performance for Acutes, Monitored and also Ambulatory Care. Nevertheless, the differences are much smaller than when comparing experiments 1 and 2. Checking Table 6.4 (columns 5, 8 and 11; rows 5, 8, 11 and 14), the triage to sign on time approximately increases, on average, by 15 minutes in Acutes and by at most 2 minutes in the rest of the services. The 95th percentile of the total time increases on average by 39 minutes in Acutes and by 16 minutes in Monitored; for Ambulatory Care and Resus the differences are not statistically significant. The fraction of patients that leave within 6 hours remains very stable between both experiments and it only shows a remarkable difference in Acutes, decreasing from 0.94 to 0.9.

It is our belief that with no pods and extra staff (Exp. 3), Ambulatory Care patients are still overtaken by Monitored and Acutes, leading the Ambulatory Care mean triage to sign-on time to be longer than in the other services. With pods and extra capacity (Exp. 4), Ambulatory Care patients are not overtaken anymore, and thus, we expect an improvement compared to pods, but it does not happen. To the best of our knowledge, this is caused by the Ambulatory Care pod suffering from the same effect as the other pods, the idle-physician/busy-pod problem. This does not happen as often when having less physicians (Exp. 2) as the workload per physician is higher and they are less likely to be idle. Presumably, in experiment 4, patients can be waiting in one pod while physicians are idle in another, causing all the pods to worsen compared to not having pods.

Lastly, we also compare the transition from having the original amount of staff to adding extra staff, while keeping the same system. Both with no pods (Exp. 1 & 3) and pods (Exp. 2 & 4),

Table 6.3: Confidence intervals[1] for the pairwise comparisons of the point estimates shown in Table 6.1. * denotes a NOT statistically significant difference.

	Mean triage to sign-on time (min)			95^{th} percentile of the total time (min)			Fraction of patients leaving the ED within 6h		
	Exp.2	Exp.3	Exp.4	Exp.2	Exp.3	Exp.4	Exp.2	Exp.3	Exp.4
Exp.1	-162.3 ± 18.06	37.29 ± 2.71	29.38 ± 2.89	-950.83 ± 81.24	109.22 ± 13.31	86.64 ± 10.17	0.18 ± 0.005	-0.08 ± 0.005	-0.06 ± 0.002
Exp.2		199.59 ± 17.38	191.67 ± 17.33		1060.05 ± 77.27	1037.47 ± 77.65		-0.27 ± 0.01	-0.24 ± 0.007
Exp.3			-7.92 ± 0.18			-22.58 ± 3.17			0.02 ± 0.004

Table 6.4: Confidence intervals[1] for the pairwise comparisons of the point estimates show in Table 6.2. * denotes a NOT statistically significant difference.

		Mean triage to sign-on time (min)			95^{th} percentile of the total time (min)			Fraction of patients leaving the ED within 6h		
		Exp.2	Exp.3	Exp.4	Exp.2	Exp.3	Exp.4	Exp.2	Exp.3	Exp.4
Acutes	Exp.1	-347.98 ± 39.07	31.68 ± 3.06	16.39 ± 3.3	-1995.92 ± 187.21	85.42 ± 187.21	46.38 ± 24.95	0.38 ± 0.006	-0.09 ± 0.019	-0.05 ± 0.015
	Exp.2		379.66 ± 40.3	364.38 ± 40.19		2081.34 ± 194.12	2042.3 ± 194.9		-0.46 ± 0.015	-0.42 ± 0.01
	Exp.3			-15.28 ± 0.34			-39.04 ± 1.25			0.04 ± 0.005
A. Care	Exp.1	43.36 ± 6.21	73.6 ± 6.73	72.94 ± 7.22	138.09 ± 23.69	222.38 ± 15.28	219.13 ± 11.75	-0.07 ± 0.005	-0.07 ± 0.005	-0.12 ± 0.004
	Exp.2		30.24 ± 2.21	29.58 ± 2.42		84.29 ± 26.36	81.04 ± 19.86		-0.05 ± 0.005	-0.05 ± 0.003
	Exp.3			-0.66 ± 0.49			-3.25 ± 6.85*			0.004 ± 0.002
Monit.	Exp.1	-42.24 ± 11.95	17.95 ± 0.87	15.59 ± 0.43	-189.91 ± 78.49	55.23 ± 16.11	39.59 ± 18.83	0.12 ± 0.03	-0.05 ± 0.02	-0.04 ± 0.03
	Exp.2		60.19 ± 12.66	57.83 ± 12.01		245.14 ± 94.57	229.51 ± 96.55		-0.17 ± 0.01	-0.15 ± 0.01
	Exp.3			-2.35 ± 0.65			-15.63 ± 7.33			0.013 ± 0.009
Resus	Exp.1	-2.13 ± 1.63	2.43 ± 0.77	1.79 ± 1	-38.09 ± 45.21*	22.61 ± 31.23*	19.66 ± 22.81*	0.02 ± 0.03*	-0.01 ± 0.01*	-0.004 ± 0.01*
	Exp.2		4.56 ± 1.04	3.92 ± 1.18		60.7 ± 14.92	57.75 ± 22.58		-0.02 ± 0.03*	-0.02 ± 0.02*
	Exp.3			-0.64 ± 0.33			-2.95 ± 8.53*			0.002 ± 0.015*

[1]Each individual CI is built with an approximate 99% confidence in order to achieve an approximate 90% overall confidence for each KPI in Table 6.3 and for each KPI per service in Table 6.4. We must say "approximate" as with only 3 replications we cannot ensure a normal distribution, but we rely on the central limit theorem. These CIs are calculated following the method "All Pairwise Comparisons" described by Law (2007) in chapter " Comparing Alternative System Configurations", section "Confidence Intervals For Comparing More Than Two Systems"

all KPIs in all services show a noteworthy improvement when increasing the workforce. Besides, it seems that for pods to obtain a similar performance to no pods, the former needs to have more physicians than no pods. Also, experiment 3 is the ones that gets the closest to the desired values of the KPIs.

6.2.3 Comparison to the Literature

In this subsection we discuss the differences yielded by the pods system in our research and in the literature.

While the pods system does not show an outstanding performance in our research, the studies in the literature report a performance improvement when using pods. The researched papers are retrospective studies that use before and after intervention data, unlike our research. Hence, the observable differences between our and their pods system might be due to human behavior, which we have not captured in our simulation model. As Song et al. (2013) describe, EDs are discretionary work settings where workers have control over several factors. One of these factors is the work pace, as physicians might work faster when the ED is busy and slow down when it is quiet. Another factor is responsibility. The literature describes that physicians have more accountability over patients when using pods (or team-based work), as patients are assigned to a pod (or physician) upon arrival. This accountability engenders a behavior that triggers a performance improvement. Even though we also assign patients to pods upon the patients' arrival, we do it in a way that patients are being pushed to beds by the system instead of being pulled by physicians and, thus, such accountability is not modeled.

We think this difference may also be caused due to the comparison of the no pods and pods systems with different physician capacities in the literature. It is not clear to us whether they just re-allocate the available staff into pods or they need to hire additional physicians, but we believe they actually add extra medical staff. To some extent, in our case this would be similar to comparing experiments 1 and 4. Therefore, the remarkable effects of the pods system described in the literature might also be the result of an increased workforce. In fact, only Dinh et al. (2015) mention the possibility of obtaining amplified positive effects in their re-design due to additional physicians.

6.3 Conclusions

In this chapter, we have assessed the effects of the no pods and pods systems on three KPIs. In addition, we have also analyzed how adding extra workforce affects both systems. Overall, our results seem to indicate that the no pods system outperforms the pods one as the former yields shorter waiting and throughput times and a higher fraction of patients leaving the ED on time. Furthermore, adding extra physicians leads to an improved performance in both systems. Our results also suggest that the pods system has a different effect on the KPIs depending on the number of physician present. With the original workforce, the pods system manages to significantly improve the metrics for the group of patients who performed the worst with no pods. Nevertheless, this happens at the expense of deteriorating the KPIs for the other pods. When increasing the workforce in the pods system, this effect is not present anymore but, on the contrary, the

idle-physician/busy-pod problem arises. The aforementioned results are from a logistics point of view, thereby showing the extent of the logistical implications of both systems. Meanwhile, the literature reports an outstanding performance of the pods system, which means that the human side makes the pods system advantageous.

Chapter 7

Conclusions and Recommendations

In this chapter we present the conclusions of our research by answering the research question in Section 7.1, followed by the recommendations in Section 7.2. Then, we introduce the limitations of this thesis in Section 7.3 and suggest future lines for research in Section 7.4. We conclude the chapter with the project contributions to practice and science in Section 7.5.

7.1 Summary of Findings

In the framework of completing the master thesis of Industrial Engineering and Management, the author performed research at the University of Auckland in collaboration with the Emergency Department of Auckland City Hospital. The research focused on studying how the implementation of a pods system, increased workforce and priority accumulation affect the ED's performance. This led to the following research question:

"How can priority accumulation, the pods system and an increase in medical staff help improve the waiting and throughput times of the Emergency Department in Auckland City Hospital?"

The answer to this research question is the following. First, we think that priority accumulation can only be implemented when the ED staff capacity is increased, as this increase will ensure a stable system. Priority accumulation can help reduce the waiting and throughput times of low priority patients, but at the expense of increasing them for other patient groups. Nevertheless, we believe that priority accumulation is sensitive to the variation in physician capacity and hence, we suggest the use of other more robust sorting methods like the ones already implemented in the simulation model. Second, we found that, overall, from a logistics perspective the no pods system outperforms the pods system as it yields shorter waiting and throughput times. Despite this generalization, the pods system allows targeting the Ambulatory Care pod and manages to improve its metrics; yet, to the noteworthy detriment of other services in other pods. Lastly, increasing the workforce affects both systems positively. However, our results suggest that the effect of the pods system on the KPIs is contingent to the number of physicians present. While the pods system with the original workforce achieved a significant improvement in the Ambulatory Care metrics, adding extra physicians results in an overall improvement, but it does not impact Ambulatory Care as before. To the best of our knowledge, this is due to some pods being busy and others having idle physicians.

7.2 Recommendations

Based on our results, we think that moving from a no pods to a pods system is not beneficial for the ED from a logistics point of view. Nevertheless, as seen in the literature, the human side -which we have not assessed- plays an important role, thereby making the pods system advantageous. With our research we do not have enough evidence to make a complete recommendation to the ED. Hence, we leave it up to the ED management to decide whether it is worth trying the pods system. In any case, we are certain that an increase in medical staff will result in shorter waiting and throughput times, as it seems they were understaffed in the original situation. Besides, the demand increases yearly, therefore, for these two reasons we suggest the addition of extra physicians.

7.3 Limitations

One of the limitations this project presents is that the simulation model cannot run for more than one year. This is because it is designed to only use the data from 2017, from which the patient characteristics, paths, consultation times, etc. are taken and inputted to the model as deterministic parameters. Hence, this means that no distributions are derived. However, the advantage this method provides is the variance reduction in results.

Another limitation is the use of the same number of physicians throughout the year. EDs are known to have seasonal effects and the data provided show these effects in patient arrivals. We know the ED does increase the workforce in busy period of the year, such as the first two weeks of January. Yet, we have not included this addition in our model, which might have resulted in worse KPIs compared to the historical data.

Lastly, the other limitations are the ones regarding the staffing model, which have already been presented in Chapter 5. Summarizing, phase 1 can strongly restrict the allocation results of phase 2, in a way in which pods have an unequal workload, leading to one pod being notably understaffed whereas the others are not. In addition, the allocation resulting from the staffing model cannot be implemented in practice straight away as it needs fine-tuning.

7.4 Future Research

In this section, we present several future lines for research that take into account all the steps involved in this thesis.

First is to improve the simulation model even further. We suggest changing how junior physicians wait for senior advice. In the current model, juniors are sent to make their own decisions - about their current patient - if they have been waiting in the advice queue for more than a certain amount of time. We propose they move on to see other patients and come back later for advice regarding the patient they could not consult about. Moreover, we recommend changing the way Resus patients are signed-on. In the current model, a Resus patient needs to wait upon arrival if no physicians are free at all, regardless of the pool they belong to. Alternatively, we suggest that a physician stops with his/her current work and takes care of this incoming Resus patient. After stabilizing the

patient, this physician can resume his/her previous task.

Second, it can also be interesting to see how the four studied experiments perform when simulating human behavior. An option is to change the physician speed depending on how busy the ED is. Nevertheless, this requires changing the physician-patient consultation time, which is actually an input parameter to our model, taken from the historical data. To address this issue, a solution is to multiply the given consultation time by a scaling factor, depending on how busy the ED is just before starting the consultation. Another option is to model the physician accountability described in the literature by implementing a pull system. Currently, patients are assigned to a pool and pushed to a bed where they wait for a physician to see them. Instead, if they were assigned to a pool and to a physician, the physician could pull the patient from the waiting room and deal with all the patients that would have been assigned to him/her in a way that improves the patient flow. Moreover, we also suggest changing the model such that it dispatches physicians taking into account skill sets in order to make best use of the differing skills between junior and senior physicians and to balance the use of various levels of clinical expertise. This modification is based on the research conducted by Hay et al. (2006). However, this would require the introduction of new parameters that should be discussed with the ED.

Third, it is necessary to confirm whether the idle-physician/busy-pod problem exists when using the pods system and extended roster. Checking the fraction of time the physicians are idle can be done through the physician states which we have already defined in the simulation model. However, to be able to use these states and get information out of them, they need to be set as an output of the simulation model and post-processed in the Python code. Besides, we find two non-mutually exclusive ways to mitigate the idle-physician/busy-pod problem, which most probably would improve the metrics of the pods system. The first one is the implementation of backup pools. In such case, each pod would own a pool with a base number of physicians and the remainder would be in the backup pool. With this system, each pod's physicians would not be allowed to change pods, whereas each physician in the backup pool could assists any pod needing help. Thus, physicians in the backup pool would not need to be allocated with the staffing model. This new system would also provide a more practical allocation than the one currently obtained with the staffing model. The second solution consists in making pods of similar sizes, as we believe the pod size influences the ED performance. Small pods are more likely to have idle physicians which would be prevented -to some extent- with similar pod sizes.

Furthermore, in our opinion, continuing the research on priority accumulation with our simulation model and settings can result in an interesting academic contribution. Using the already implemented priority accumulation prototype and increasing the workforce in the ED to have a stable system, one can analyze the effects of this queuing discipline on the no pods and pods system.

Lastly, the staffing model should be revised. We believe the objective of phase 1 is not properly designed for the purpose we want it. Therefore, one should investigate how it can be changed. It may even be that phase 1 is not needed and phase 2 may already perform well on its own. However, phase 2 will take a longer running time as it will not be bounded by phase 1. Hence, we suggest the introduction of an initial solution upon which phase 2 can build a better one.

7.5 Project Contributions

7.5.1 Value for Practice

The practical relevance of this thesis is the following. First, we have given the ED insight about average bed occupation per service and hour of the day. This is key information for determining staffing and new rosters and shifts. Second, we have provided the simulation model, which is a very powerful tool to simulate any further changes in the ED. Besides, we have presented different ways to improve this model. Finally, we have also contributed with a staffing model for physician allocation in pods. Even though it needs some adjustments, it will act as an another powerful tool that will save time and potentially avoid human errors.

7.5.2 Value for Science

With this research, we contribute to science in two main ways. The first one is with regard to the literature. While there are not many studies about the pods system and there is not a clear definition of what such system involves, our contribution has been a first state-of-the art review. Such review presents studies that have used pods in EDs as well as other studies solely using the methodologies that the pods system seems to involve, which are team-based work and unpooling of resources. From here, we make a call to scholars to further investigate this topic, to standardize and define the pod concept and to write a more extended state-of-the-art review than ours.

Furthermore, our other contribution is the comparison of two discrete-event-simulation software programs: JaamSim and Tecnomatrix Plant Simulation. On the one hand, we present JaamSim, an open source DES software tool that runs in Java, which was developed by the consultancy company Ausenco. On the other hand, we present Tecnomatrix Plant Simulation, a commercial software tool developed by Siemens PLM Software, used for modeling, simulating, analyzing and optimizing. Both are for decision support in Operations Research and Management. We compare both tools from different points of view:

- The first main aspect that comes to light is the difference in cost of both tools. While JaamSim is a free open source tool that offers daily updates and maintains a forum with several topics, Plant Simulation involves very high costs, making small and medium companies to be reluctant to adopt it. Nevertheless, Plan Simulation offers free academic license to students, with the condition of not using it for paid services or other commercial purposes.
- The basic objects provided by JaamSim were too limited for the needs of our project. Therefore, we had to develop customized objects. This had two implications. First, it was a difficult task for non-advanced users, and second, when using customized objects JaamSim needs to be launched by another software program such as Eclipse, which hindered the whole process. Contrarily, Plant Simulation offers many objects with many characteristics.
- The previous point can also be seen from another perspective. Advanced users of JaamSim can create a repository of components that can be used by third parties. This gives the possibility to share and reuse customized objects with other simulation programs, whereas Plan Simulation comes as one monolithic product where customization of objects is difficult as there is a lack of access to the source code.

- Another main difference between these two tools is how the logic of the simulation is implemented. JaamSim allows to do so through several objects, or by writing conditions in other objects. To us, this was a big complication as it is difficult to implement complex logic in such a way. Contrarily, Plant Simulation offers objects called "Methods", which enable the modeler to program custom logic into the model.
- As opposite to JaamSim, obtaining KPIs with Plant Simulation is easy as one can program several calculations in these "Methods" and present the results in customized tables. With our knowledge, this process was not even possible in JaamSim and, instead, the outputs for each entity had to be recorded in a text file and processed in Python.
- Furthermore, JaamSim does not provide a debugging tool for the simulation model. In case of having customized objects, only debugging of the Java code can be done. Contrarily, Plant Simulation offers a useful and complete debugging tool that links the visual simulation and the methods.
- Lastly, we found the fact that JaamSim does not allow to simultaneously select and move multiple objects an inconvenience. This was unhandy when extending the model.

Summarizing, we think that Plant Simulation or other DES software packaged provide simulation tools with which the development of our simulation model and the manipulation of outputs would have been much easier. The downside is that the hospital would require a costly license.

For an interesting comparison of available open source DES programs and their advantages and disadvantages compared to commercial DES programs, we recommend reading the review conducted by Dagkakis and Heavey (2016).

Bibliography

Dagkakis, G. and Heavey, C. (2016). A review of open source discrete event simulation software for operations research. *Journal of Simulation*, 10(3):193–206.

Dinh, M. M., Green, T. C., Bein, K. J., Lo, S., Jones, A., and Johnson, T. (2015). Emergency department clinical redesign, team-based care and improvements in hospital performance: A time series analysis. *Emergency Medicine Australasia*, 27(4):317–322.

ESI (2004). Emergency severity index triage algorithm. `https://www.esitriage.com/esi-algorithm`. [Online; accessed 7-August-2019].

Gavin, N. and Peterson, K. (2017). Team-based pod system reduces lengths of stay for treat-and-release patients. *ED management: the monthly update on emergency department management*, 29(6):67–69.

Hay, A. M., Valentin, E. C., and Bijlsma, R. A. (2006). Modeling emergency care in hospitals: a paradox-the patient should not drive the process. In *Proceedings of the 2006 Winter Simulation Conference*, pages 439–445. IEEE.

Kleinrock, L. (1964). A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11(3-4):329–341.

Kleinrock, L. and Finkelstein, R. P. (1967). Time dependent priority queues. *Operations Research*, 15(1):104–116.

Lau, F. and Leung, K. (1997). Waiting time in an urban accident and emergency department–a way to improve it. *Emergency Medicine Journal*, 14(5):299–303.

Law, A. M. (2007). *Simulation modeling and analysis*. McGraw-Hill New York, 4th edition.

Li, N., Stanford, D. A., Taylor, P., and Ziedins, I. (2017). Nonlinear accumulating priority queues with equivalent linear proxies. *Operations Research*, 65(6):1712–1721.

Melton III, J. D., Blind, F., Hall, A. B., Leckie, M., and Novotny, A. (2016). Impact of a hospitalwide quality improvement initiative on emergency department throughput and crowding measures. *The Joint Commission Journal on Quality and Patient Safety*, 42(12):533–542.

Morgareidge, D., Hui, C., and Jun, J. (2014). Performance-driven design with the support of digital tools: Applying discrete event simulation and space syntax on the design of the emergency department. *Frontiers of Architectural Research*, 3(3):250–264.

Pati, D., Harvey Jr, T. E., and Pati, S. (2014). Physical design correlates of efficiency and safety in emergency departments: a qualitative examination. *Critical care nursing quarterly*, 37(3):299–316.

Robinson, S. (2015). A tutorial on conceptual modeling for simulation. In *Proceedings of the 2015 Winter Simulation Conference*, pages 1820–1834. IEEE Press.

Sharif, A. B., Stanford, D. A., Taylor, P., and Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79.

Song, H., Tucker, A. L., and Murrell, K. L. (2013). The impact of pooling on throughput time in discretionary work settings: An empirical investigation of emergency department length of stay.

Song, H., Tucker, A. L., and Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053.

Stanford, D. A., Taylor, P., and Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3):297–330.

Torrence Memorial (2014). Easing the wait: New pod system reduces time spent in the ed. `https://www.torrancememorial.org/News_Center/2014/May/Easing_the_Wait_New_Pod_System_Reduces_Time_Spen.aspx`. [Online; accessed 15-April-2019].

Appendix A

Simulation Model Main Frame

The simulation model is divided in six main areas (Figure A.1).

- Arrival and triage: patients are generated in this area and assigned their attributes. Then they are triaged, assigned a physician pool and sent to the corresponding service.
- Service and doctor consultations: these area of the simulation represents the 8 different services of the ED. After being triaged, patients are placed in a bed or wait in queue until one becomes free. They wait for a physician to see them for assessment or for the decision to be notified. The time patients spend doing tests, scans and waiting for the results is simulated in this area as well.
- Discharge: before discharging the patients, their time stamps are recorded in order to obtain the simulation output.
- Doctor generators and rosters: each of the physicians in this area represents a physician shift (presented in Table B.1 and B.2). We can also observe the physician generators and the associated time series, that define their working times.
- Junior/SMO advice and admin work: this area of the frame is where senior physicians give advice to the junior ones. Also, it is where physicians come to fill in information to the system, fill in forms, analyze patients' results and make decisions.
- Sorting and dispatching logic: this area is where the vast majority of the model logic is programmed. The main decisions are 1) which patient should be treated next and 2) which physician is to be dispatched to which patient.

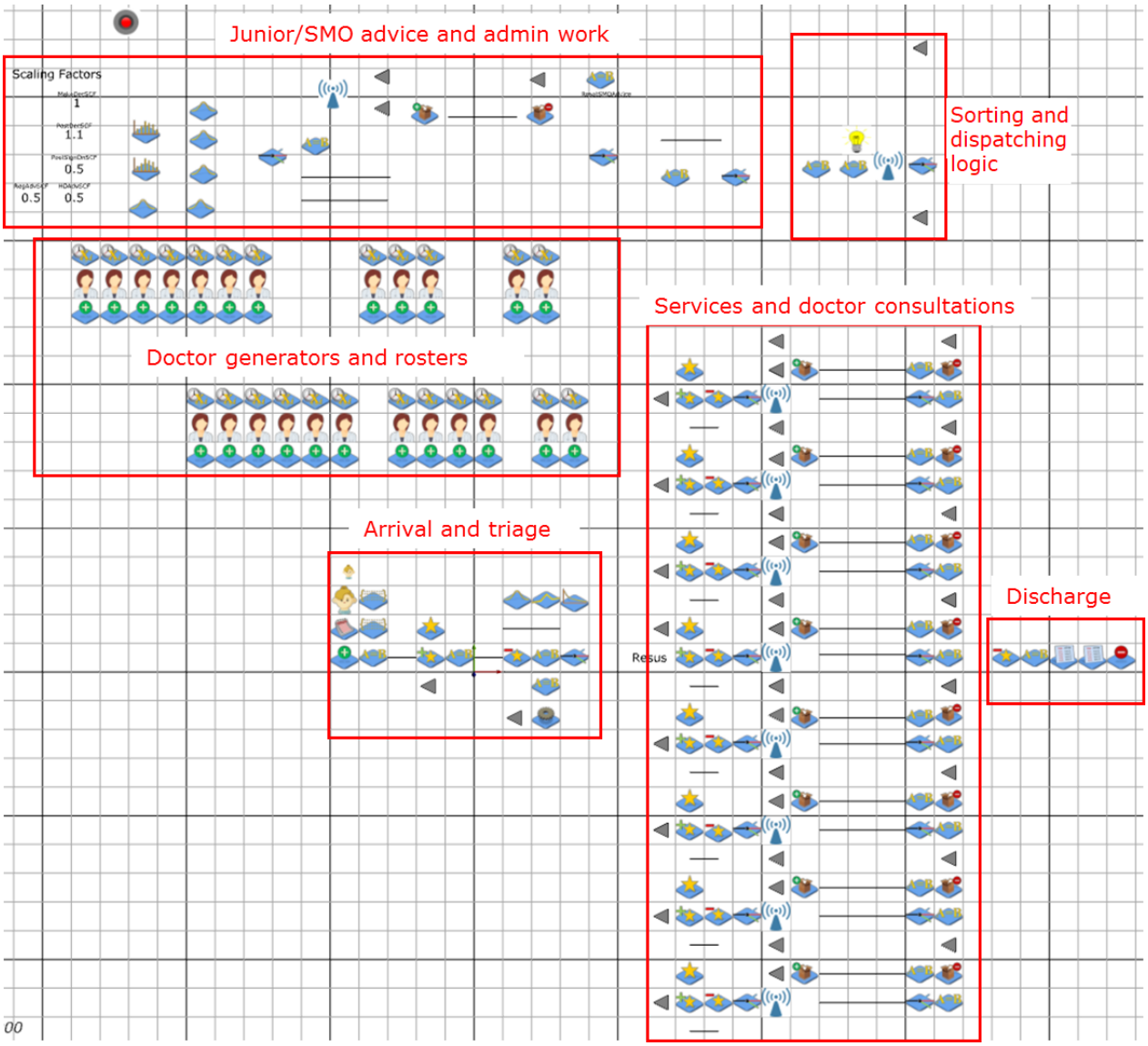

Figure A.1: Main frame of the JaamSim model

Appendix B

Staffing Model

B.1 MILP input

B.1.1 Physician Shifts

As introduced in chapter 5, section 5.3.1, the ED provided the no pods roster they were using before the implementation of the pods system. This can be seen in Table B.1. It specifies the physicians' role (according to level of knowledge), whether they work in the morning, afternoon or at night and the start and end times. Moreover, in this case it also specifies the pools in which the physicians work under the no pods system. Nevertheless, we do not pay attention to them when using the MILP model as our objective is to find a new allocation. But we do use them for the first experiment presented in chapter 5.

As explained in section 5.3.1, we refer to each of the rows of Table B.1 as a "shift" since they do not represent a specific physical person but a period of time that needs to be filled by someone of the specified role. These shifts are repeated every day of the week, except for 3 cases, which they are only present during weekends. For sake of simplicity, it is easier for us to refer to each of these shifts as a physician, even though they are not people per se. In addition, the ED also provided another roster for the pods system, however, without any allocation. This can be seen in Table B.2.

B.1.2 Beds Occupied

The calculation of number of beds occupied is essential to get to know the demand for ED services. This will be fundamental when staffing pods or defining new rosters in a future. For now, both actions are beyond scope.

To carry out such calculation we take into account the patients' admissions as well as their length of stay. Using R-studio, we calculate the amount of beds occupied at the same time for every hour of the week throughout the year, and we then take the average in order to have a unique week representing the whole year. For simplicity, we do not take into account seasonal effects. We perform these computations for every ED service, which enables us to extract some interesting

information. For example, Figure B.1 shows a daily boxplot for the beds occupied in Ambulatory Care. We can observe that Monday is, by far, the busiest day and, on the contrary, Thursday is the least busy. From the boxplot we can interpret that the distribution curve of the beds occupied is right-skewed in all cases. Furthermore, we are also interested in the occupation of beds per hour. Figure B.2 shows the beds occupied per hour per day of the week in the Ambulatory Care service. We can see how the demand decreases during the night, it rises in the morning reaching its peak in the afternoon. Even though the variation of demand in this service between day and night hours is considerable, this pattern is stable throughout the week. All the services present a similar pattern.

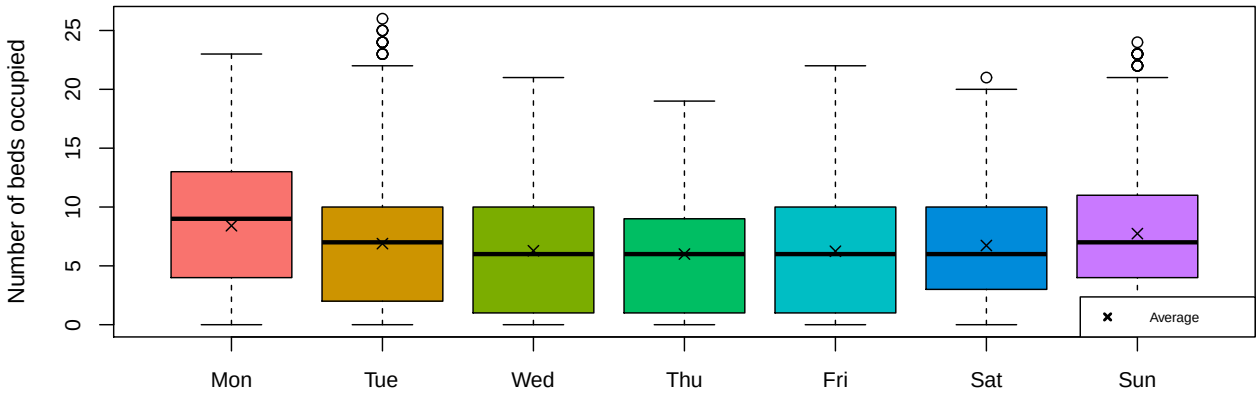

Figure B.1: Beds occupied per day of the week in the Ambulatory Care service

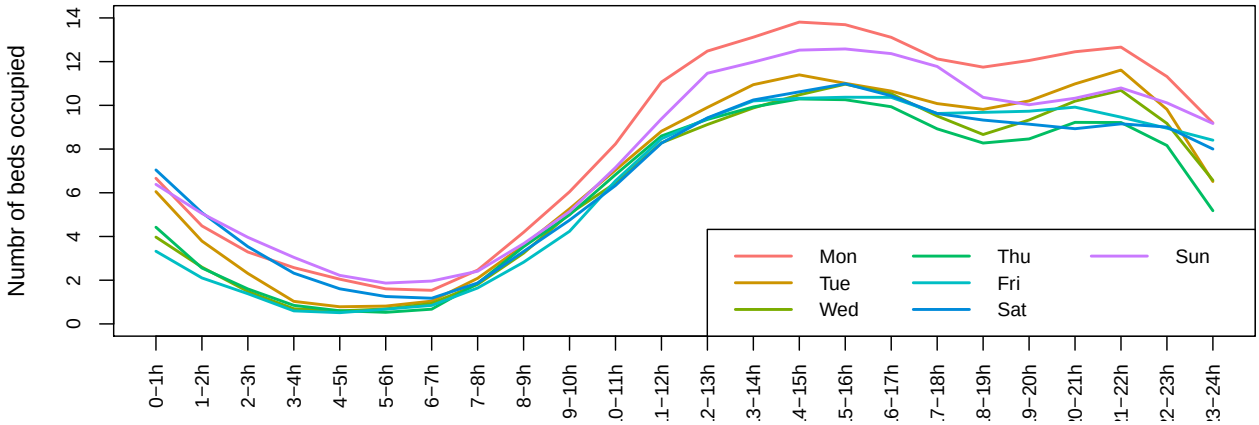

Figure B.2: Comparison among beds occupied per hour per day of the week in the Ambulatory Care service

Having done this analysis for each service (for a matter of space the rest of the figures are not included), we can group the data and calculate the beds occupied for each pod. Figures B.3 and B.4 show the beds occupied per hour of the week per pod for the no pods and pods systems respectively. The numbers behind these two figures are actually the input to the MILP model. The dashed lines simply separate the days.

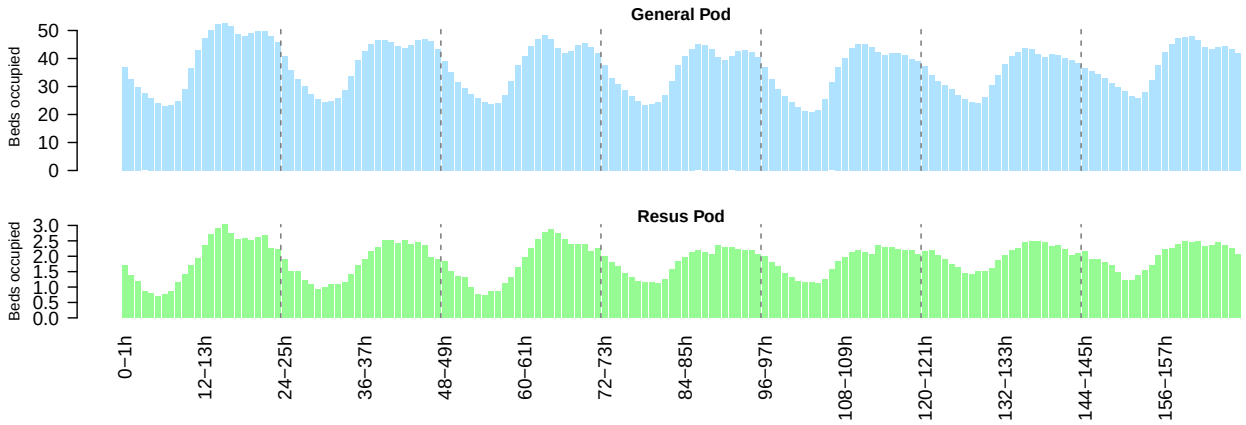

Figure B.3: Beds occupied per pod in the no pods system

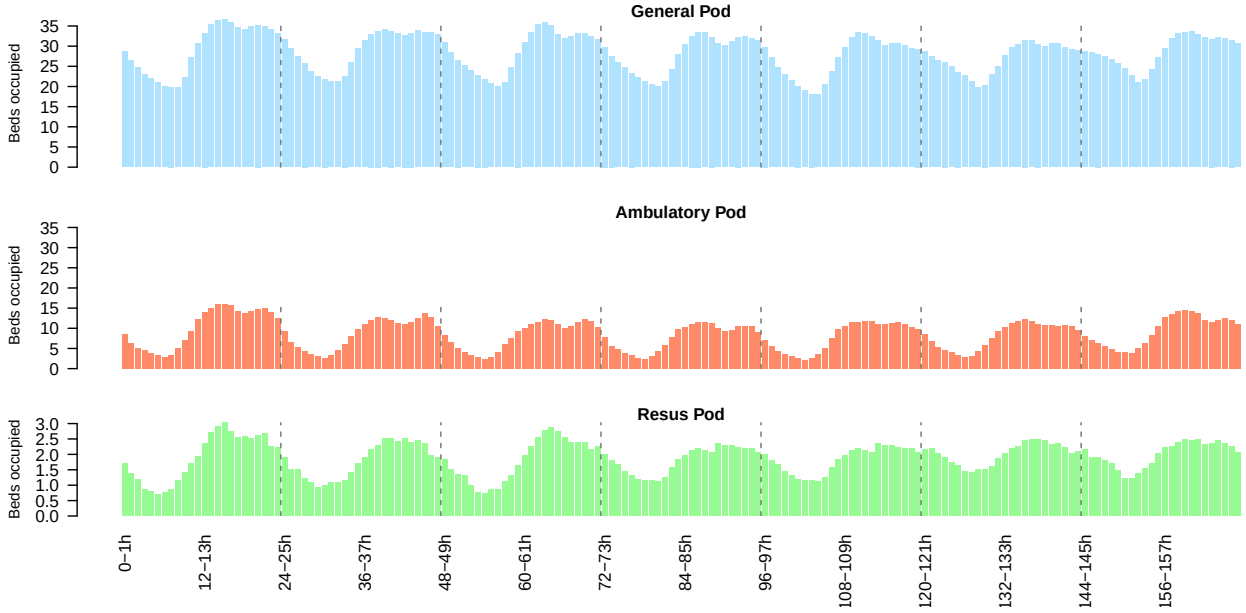

Figure B.4: Beds occupied per pod in the pods system

B.1.3 Weights

It is also of interest to know the volume of patients in each service, area and pod. Figure B.5 shows the percentage of volume of patients per service and area. Check Table X in chapter 1 to see which services are included in which areas. 41.7% of the patients pass through Acutes, 22.2% through Ambulatory Care, 17% through Monitored and 7% through Resus. The other services handle a very small percentage of patients. When grouping these services in areas, the Acutes area handles 46.2% of the patient volume, followed by Ambulatory Care, whose percentage has increased to 29.8%. Monitored and Resus remain the same as before since these areas only include the Monitored and Resus services. Moreover, we have also normalized the percentage for only the four main areas as these are used in chapter 5. From this information we can conclude that the overall ED performance

will be mainly affected by the performance in the Acutes service.

In a same manner, the overall ED performance will be affected more by one pod than another. To reflect this, we have also calculated the volume of patients in each pod and we have used such volumes to define the pod weights used in the MILP model. Figure B.6 shows the weights of each pod in the no pods and pods systems.

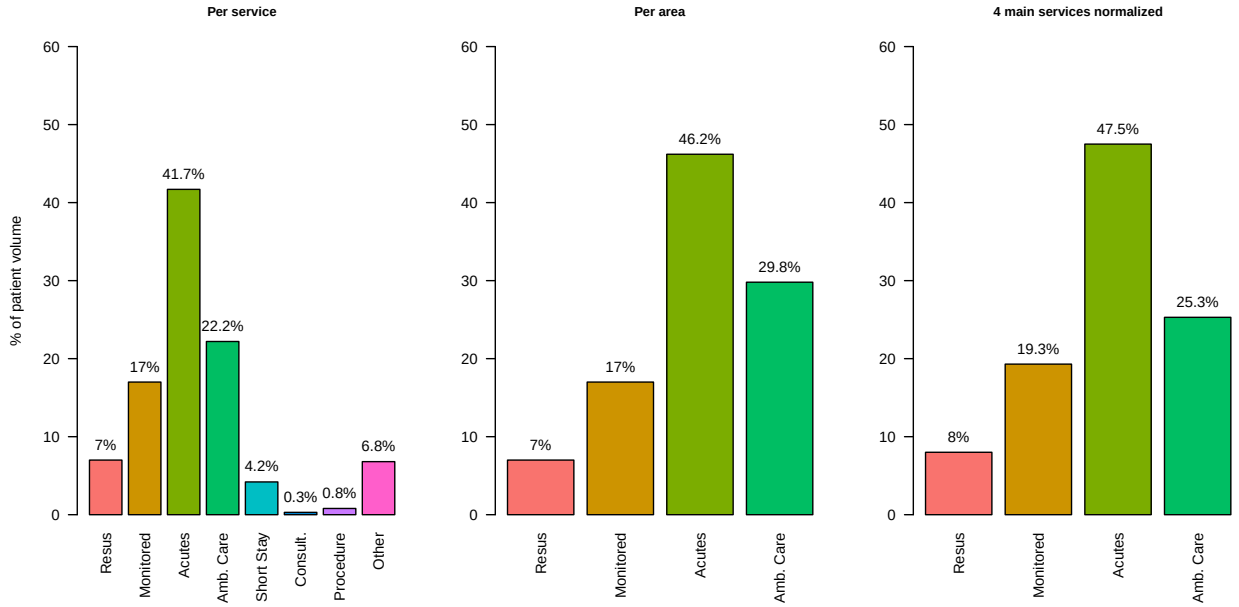

Figure B.5: Percentages of patient volumes per service and area

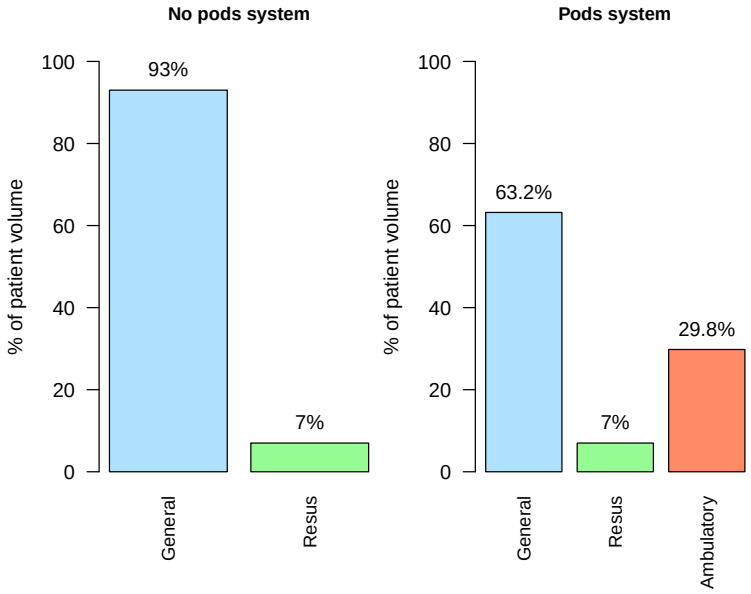

Figure B.6: Pod weights for each system

Table B.1: Original roster. (M: Morning, A: Afternoon, N: Night). Tue* means that M1 and M2 Registrars are absent on Tuesdays from 10:00 to 14:00.

Role	Shift	Start Time	End Time	Duration	Area
SMO	M1	8:00	17:00	9	Resus
SMO	M2	8:00	17:00	9	General
SMO	M3	8:00	17:00	9	General
SMO	A1	16:00	0:00	8	Resus
SMO	A2	16:00	0:00	8	General
SMO	A3	16:00	0:00	8	General
SMO	N1 (Fri-Sun)	0:00	8:00	8	Resus
Fellow	10	10:00	20:00	10	General
Fellow	10 (Sat-Sun)	10:00	20:00	10	General
Fellow	14	14:00	0:00	10	Resus (14:00-19:00), General
MOSS	12	12:00	22:00	10	General
MOSS	14	14:00	0:00	10	General
Registrar	M1 (Tue*)	8:00	18:00	10	Resus (8:00-14:00), General
Registrar	M2 (Tue*)	8:00	18:00	10	General
Registrar	A1	16:00	2:00	10	Resus (19:00-24:00), General
Registrar	A2	16:00	2:00	10	General
Registrar	N1	22:30	8:30	10	Resus (24:00-8:00), General
Registrar	N2	22:30	8:30	10	General
HO	M	8:00	18:00	10	General (Not Monitored)
HO	A	16:00	2:00	10	General (Not Monitored)
HO	N	22:30	8:30	10	General (Not Monitored)
HO	N (Fri-Sun)	22:30	8:30	10	General (Not Monitored)
CNS/NP	M	9:00	19:00	10	General (Not Monitored or Acutes)
CNS/NP	A	12:00	0:00	12	General (Not Monitored or Acutes)

Table B.2: Extended staff roster. (M: Morning, A: Afternoon, N: Night). In red, the added shifts compared to the original roster.

Role	Shift	Start Time	End Time	Duration	Area
SMO	M1	8:00	17:00	9	
SMO	M2	8:00	17:00	9	
SMO	M3	8:00	17:00	9	
SMO	M4	8:00	17:00	9	
SMO	A1	16:00	0:00	8	
SMO	A2	16:00	0:00	8	
SMO	A3	16:00	0:00	8	
SMO	A4	16:00	0:00	8	
SMO	N1 (Fri-Sat)	0:00	8:00	8	
Fellow/MOSS	M1	8:00	17:00	9	
Fellow/MOSS	M2	8:00	17:00	9	
Fellow/MOSS	M3	8:00	17:00	9	
Fellow/MOSS	A1	16:00	2:00	9	
Fellow/MOSS	A2	16:00	2:00	9	
Fellow/MOSS	A3	16:00	2:00	9	
Fellow	N1 (Sun-Thu)	22:30	8:00	10	
Registrar	M1	8:00	18:00	10	
Registrar	M2	8:00	18:00	10	
Registrar	A1	16:00	2:00	10	
Registrar	A2	16:00	2:00	10	
Registrar	N1	22:30	8:30	10	
Registrar	N2	22:30	8:30	10	
HO	M	8:00	18:00	10	
HO	A	16:00	2:00	10	
HO	N	22:30	8:30	10	
HO	N (Fri-Sun)	22:30	8:30	10	
CNS/NP	M	9:00	19:00	10	
CNS/NP	A	12:00	0:00	12	

B.2 How Phase 1 Affects Phase 2 in the Staffing Model

In this section we give an example to reflect the limitations mentioned in chapter 4.

As we explained in the MILP limitations in chapter 4, phase 1 is designed in such a way that prioritizes the allocation of physicians to the easiest pod to get the highest C/O ratio, resulting in very high lower bounds for some pods and very low for others. Let us give an example using the results shown in Table B.3. These are the results of phase 1 with the extended roster and pods system at hour 146 (2am to 3am), which corresponds to the bottleneck hour for both pods. While phase 1 assigns 1 Registrar to the General pod and 2 HOs to Ambulatory, it could have also assigned 1 of the HOs to General and leave only 1 HO in Ambulatory. With the first solution, the objective results in a value of 0.320 (not taking into account Resus), whereas with the second solution the resulting value is 0.299. Even though the second solution makes more sense because the General pod has more beds occupied, and thus a higher need for staff, the model chooses the first one. The reason behind this choice is that having two physicians in Ambulatory results in a higher C/O ratio than having them in General, which multiplied by the pod weight, compensates and overtakes the fact that the Ambulatory pod has a lower weight.

Table B.3: Output phase 1, pods system, extended roster, at h =146. Phys = physicians

	General pod (w=0.63)					Ambulatory (w=0.3)					Objective Phase 1
	Phys.	C	O	C/O	$C/O * w$	Phys.	C	O	C/O	$C/O * w$	$\sum(C/O * w)$
Model Decision	1 Reg	4.5	28.1	0.160	0.101	2 HO	6	6.28	0.956	0.287	0.388
Move HO	1 Reg 1 HO	7.5	28.1	0.267	0.168	1 HO	3	6.28	0.478	0.143	0.311

The LBs of phase 1 clearly have an impact on the results of phase 2. We show it with a new example in Table B.4. These are the results of phase 2 for the no pods system, extended roster at hour 39 (3pm to 4 pm). Phase 2 allocates the physicians as shown in the third row. This results in an absolute distance of C/O ratios between pods of 0.345. If instead, we move one Registrar from Ambulatory to General, the distance is reduced to 0.145, meaning that the stress among pods is more similar. Nevertheless, the second solution cannot be chosen because the C/O of Ambulatory pod is lower than the LB defined by phase 1, which is 0.956.

Table B.4: Output phase 2, pods system, extended roster, at h =39. Phys = physicians

	General pod				Ambulatory				Objective Phase 2
	Phys.	C	O	C/O	Phys.	C	O	C/O	$\lvert C/O$ distance$\rvert$
Model Decision	X	23.5	34.1	0.689	Y	13	12.5	1.040	0.32
Move Registrar	X+1Reg	28	34.1	0.821	Y- 1 Reg	8.5	12.5	0.680	0.141

To conclude, the objective of phase 1 is not well designed, which affects the results of phase 2.

Phase 1 should be designed such that the minimum ratios are maximized but difference between them is not as big. However, it is yet to be determined whether phase 1 is needed at all.