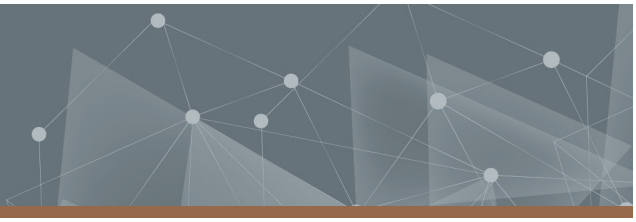
CHALMERS
UNIVERSITY OF TECHNOLOGY



Discrete-Event Simulation of an Ear, Nose and Throat Emergency Flow

Master's thesis in Management and Economics of Innovation, and Quality and Operations Management

Sören Lambrecht
Johan Skoglund

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISION OF SUPPLY AND OPERATIONS MANAGEMENT

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

www.chalmers.se
E2020:003

This page has intentionally been left blank.

Report No. E2020:003

Discrete-Event Simulation of an Ear, Nose and Throat Emergency Flow

Sören Lambrecht
Johan Skoglund

Department of Technology Management and Economics
Division of Supply and Operations Management
Chalmers University of Technology
Gothenburg, Sweden 2020

Discrete-Event Simulation of an Ear, Nose and Throat Emergency Flow

Auhors: Sören Lambrecht and Johan Skoglund

Supervisors: Peter Almström and Per Medbo, Department of Technology Management and Economics, Division of Supply and Operations Management

Examiner: Peter Almström, Department of Technology Management and Economics, Division of Supply and Operations Management

Report No. E2020:003
Department of Technology Management and Economics
Division of Supply and Operations Management
Chalmers University of Technology
SE-412 96 Gothenburg
Telefon +46 (0)31 772 1000

Gothenburg, Sweden 2020

Abstract

The purpose of this thesis was to map and simulate the current state of the emergency flow of the Ear, Nose and Throat (ENT) clinic at a regional hospital in the Swedish Västra Götaland region, further referred to as the Hospital, as well as suggest an alteration to the flow with the aim of freeing up resources that could be scheduled for the vocational care flow.

To understand the current state of the system, theory on operations management and the organisational aspects of healthcare was researched to understand the context of the simulation study. In addition, theory on simulation as a method was studied to provide a roadmap for the simulation modelling phase, executed in the simulation software SIMUL8.

The method leveraged interviews and observations as sources of primary data. The interviews provided background knowledge about the system, assisted in the conceptual mapping and helped validate the system throughout the modelling and results collection process. Observations provided process times and an understanding of the individual processes. Secondary data was also provided from the Hospital in the form of information about all patients admitted during the last three years. The data was analysed and aggregated manually and with the assistance of computer software, assigned to distribution functions and tested for goodness-of-fit. Due to the COVID-19 pandemic, the data collection phase was interrupted early and access to the clinic restricted, leading to limited statistical validation, and, therefore, extra validation of logical flows and process times.

By understanding, describing and modelling the physical state of the emergency flow, it was understood that different resources affect the throughput time of the system differently. Physicians and nurses working directly with physicians are essential to the system and can not be relocated or reduced. However, nurses that are not working directly with physicians are not highly utilised, and thus able to be rescheduled in order to increase resource availability in the vocational care flow.

A desired future state model is presented in this thesis and allows for task sharing between two of the nurses in the system, allowing for shorter average throughput times in the entire emergency flow. The freed up resource can be relocated to the vocational car flow of the ENT clinic and decrease throughput time as well as contribute to increase accessibility of the entire ENT clinic. The authors recommend the Hospital to pilot the solution on a smaller scale before a full-scale implementation.

In terms of generalisability and suggestions for further research, this thesis indicates the occurrence of sub-optimised systems within healthcare and that simulation modeling could be used as a tool in healthcare to solve complex problems in healthcare operations. Simulation can be used to visualise and explain complex systems such as patient flows, leading to better understanding and communication.

Acknowledgements

The work entailed in this master's thesis was carried out between January and June 2020 as the final project of the authors' master's degrees in Industrial Engineering and Management at Chalmers University of Technology, Gothenburg. A number of people have been involved in the work that led to this report, and we would like to express our gratefulness to everyone who has been involved.

Since it was decided that the hospital where the study was carried out was to be kept anonymous, we will not name the people who have been involved from the hospital's side. However, we would like to extend our sincere gratitude to all of you who were involved in answering our questions, letting us observe the processes and providing feedback to our model and suggestions. Without you, this study could now have been carried out.

We would also like to thank Peter Almström and Per Medbo, who assisted with supervision and guidance when we were faced with complications or if we needed some quick input to the report and model.

Sören Lambrecht

Sören Lambrecht

Johan Skoglund

Gothenburg, June 2020

Contents

This page has intentionally been left blank.

1

Introduction

1.1 Background

Swedish healthcare is known to have a very high degree of clinical excellence and ranks among the top 10 nations in Europe (Björnberg & Phang, 2019). However, Sweden scores low in the accessibility category due to long queues and waiting times, even though the government has spent a considerable amount of money to rectify the issues. Partly due to this, Sweden has one of the highest healthcare spend per capita in Europe.

This study is carried out at one of Sweden's largest hospitals, situated in one of Sweden's largest regions, where the demand for care is high and waiting times for care in many cases are very long.

As a response to the accessibility and cost problems experienced in Swedish healthcare, a pilot study of a new production system within otorhinolaryngology in the Västra Götaland region is being carried out with the help of Chalmers University of Technology researchers. The aim of the pilot study is to increase the throughput rate without adding additional resources. To achieve this, there is a need to create a general operations mindset and appropriate controls, as well as aiding in the design of efficient operations and processes. The pilot study also aims to develop tools that can be used for more effective planning and decision making. This thesis is part of the pilot study.

The otorhinolaryngology clinic at the Hospital deals with patients seeking medical treatment for conditions related to the ears, nose or throat (ENT). The clinic receives both elective and emergency patients through referrals, time bookings and in some cases by patients walking in and asking for care. Paired with the department having several flows for different patient groups and carrying out both simple and complex surgeries, the planning is complicated further. As a result of this, the patient flows quickly become complex. To tackle these issues, this thesis aims to leverage discrete-event simulation (simulation modelling) to find potential bottle necks and experiment with possible flow improvements.

1.2 Problem description

As mentioned in chapter 1.1, Swedish healthcare is wrestling with long waiting times and capacity issues leading to low accessibility scores. The government has historically tried to solve the problem by increasing spending, but the efforts have not yielded any significant improvements (Björnberg & Phang, 2019).

According to Kaplan and Porter (2011), there is a general lack of understanding in healthcare about how much delivery of patient care actually costs. Measuring healthcare cost is challenging due to the complex nature of healthcare delivery and the wide array of resources used. There is also a difficulty in taking a cost cutting approach, as cost reductions need to be backed by accurate data on resource utilisation and process efficiencies to avoid jeopardising the quality of the service delivered.

Healthcare process improvement aims to improve both the quality of care and cost effectiveness (Colldén & Hellström, 2018). It is driven by high-level societal objectives and financial incentives to deliver higher value to patients by utilising and managing resources effectively (Conway, 2019). Historically, this has not been the case as many different healthcare functions have operated in silos with limited communication and cross-functional collaboration, leading to higher overall costs in healthcare delivery.

Queues are not an issue related only to healthcare but to all kinds of operations such as lines at the supermarket, in airports and manufacturing operations. According to Slack et al. (2013), queues appear due to e.g. limited capacity, task-time variation and variability in demand or supply. By utilising operations management methods and tools for planning and control, long queues can be reduced and lead to an increase in operational efficiency.

Simulation is a well-tested method to analyse how new or altered production systems will act in reality (Chung, 2004). Simulation is a cost effective tool to plan the design and scale of a proposed change, as simulation runs, made on a computer, can be made in a fraction of the time demanded in the real system, with reduced analytic requirements required compared to traditional mathematics and operations research. Furthermore, animating a simulation model can act as a tool for visualising, demonstrating and communicating a process, instead of relying solely on text and numeric descriptions.

Based on the identified problems, it is believed that operations management theory paired with discrete-event simulation can be used effectively to map and understand the ENT emergency flow. Furthermore, this thesis aims to discern how and if the emergency flow can be altered to free up resources without negative implications to the clinic.

1.3 Aim and research questions

The aim of this thesis is to, with the help of discrete-event simulation, map and model the patient flow at the ENT clinic within the Hospital. The simulation model and the

thesis's conclusions will, as a part of the pilot study, act as a base for analysis of the system and propose potential improvements.

Based in the aim and problem analysis, two research questions have been formulated that will be answered by utilising discrete-event simulation:

- *RQ 1: What is the current state of the Hospital's ENT emergency flow?*
- *RQ 2: How can the emergency flow be altered to free up resources?*

1.4 Limitations and delimitations

This thesis will be delimited to the ENT clinic at the Hospital, mapping the process and simulating the emergency patient flow. This is because the thesis is a part of the above mentioned wider pilot study, which focuses on the ENT clinic.

The simulation model will be limited to the above mentioned unit, starting with the patient being received at the clinic and ending with the patient leaving the clinic to proceed to another unit or exit the flow entirely.

As the thesis takes an operations management and simulation perspective, any tasks that are considered to need specific medicinal expertise will be considered outside the scope. The focus will instead be placed on the administrative and logistical aspects connected to the patient flow.

Another limitation experienced during the thesis work was the COVID-19 pandemic, limiting access to the clinic where the study was conducted. As the pandemic took force during the study, the clinic was forced to restrict access to non-essential visitors. This impacted the data collection work, which in turn impacted the rest of the study and simulation model, forcing adjustments to the method as the work progressed.

2

Theoretical framework

This chapter introduces the academic theory used during the study and for subsequent analysis and discussion. The chapter covers three topics connected to the research question: *Operations planning and control, Healthcare organisation* and *Discrete-event simulation*.

2.1 Operations planning and control

According to McLaughlin and Olson (2017), quality improvement is progressing at a slow rate in healthcare generally, and a strong focus on operations can help resolve this issue by increasing safety, improving clinical outcomes, reducing costs and helping organisations become more competitive. Furthermore, Chan and Green (2013) suggest that operations research is useful for dealing with a high rate of variability and navigating through periods of high congestion.

2.1.1 The 4 V's of operations management

The 4 V's of operations management is a way to describe an operation's characteristics in its dimensions of *volume* and *variety* of output, *variation* in demand and the degree of *visibility* that customers have into the operation (Slack et al., 2013). These V's can further be used to describe how an operation transform its inputs into outputs, helping categorise and distinguish between different operations in both similar and different sectors.

Volume

The volume dimension of an operation describes how high volumes an operation produces, e.g. the number of patients treated in a hospital (Jacobsson, 2012; Slack et al., 2013). An operation with a high volume would typically enjoy benefits from specialised process technology and high systematisation and repetition, leading to lower unit costs and greater efficiency than an operation producing low volumes (Slack & Lewis, 2017).

However, there are not only benefits to producing in high volumes. An operation providing a luxury service, such as a hotel, might want to limit the number of guests and tailor its service to each individual customer, thus providing a higher variety (Slack et al., 2013).

Variety

Variety implicates the range of different activities offered by the operation, for example a taxi service being able to pick up customers from many different locations, or the number of different diseases treated in a hospital (Jacobsson, 2012; Slack & Lewis, 2017). An operation with a high variety requires a higher degree of flexibility than an operation with a more standardised offering, like a bus service that can plan its routes in advance and determine pickup locations and schedules in advance (Slack et al., 2013).

Variation

Variation describes how demand fluctuates as a function of changing customer demand, e.g. for healthcare services or degree of patient contact (Jacobsson, 2012; Slack et al., 2013). An operation experiencing high variety, normally has to readjust its resource base more frequently to deal with fluctuating demand and seasonality. This requires a higher ability to adjust its capacity in terms of e.g. inventory or extra resources, often leading to a higher unit cost (Slack & Lewis, 2017).

Visibility

Visibility is the dimension which describes how 'exposed' an operation's activities are to its customers (Slack et al., 2013). In general, an operation which processes customers has a higher degree of visibility, for example a physical store or a healthcare process in comparison to an operation which handles material or information. Due to this higher visibility, customers will typically have a shorter waiting tolerance, and may choose to walk out if they do not get served in what they perceive to be a reasonable time. Thus, a high-visibility operation generally requires personnel with better customer service skills.

The degree of visibility can, even in customer processing operations, be altered to better suit the operation (Slack et al., 2013). For example, a retailer can choose to be web-based instead of using physical stores. By doing so, the operation can more resemble a factory with more standardised packing and dispatching, as well more centralised warehouses. Furthermore, the customer's waiting tolerance typically becomes higher, making it possible for the operation to schedule its activities to a greater extent, leading to higher resource utilisation and lower costs, compared to a physical store.

Implications of the 4 V's of Operations Management

Generally, all 4 V's have implications for how costly a product or service will be to deliver (Slack et al., 2013). Briefly summarised, a high volume, low variety, variation and visibility help maintain processing costs on a low level, and vice versa. The 4 V's are illustrated in figure 2.1, with high cost factors being located to the left in the figure.

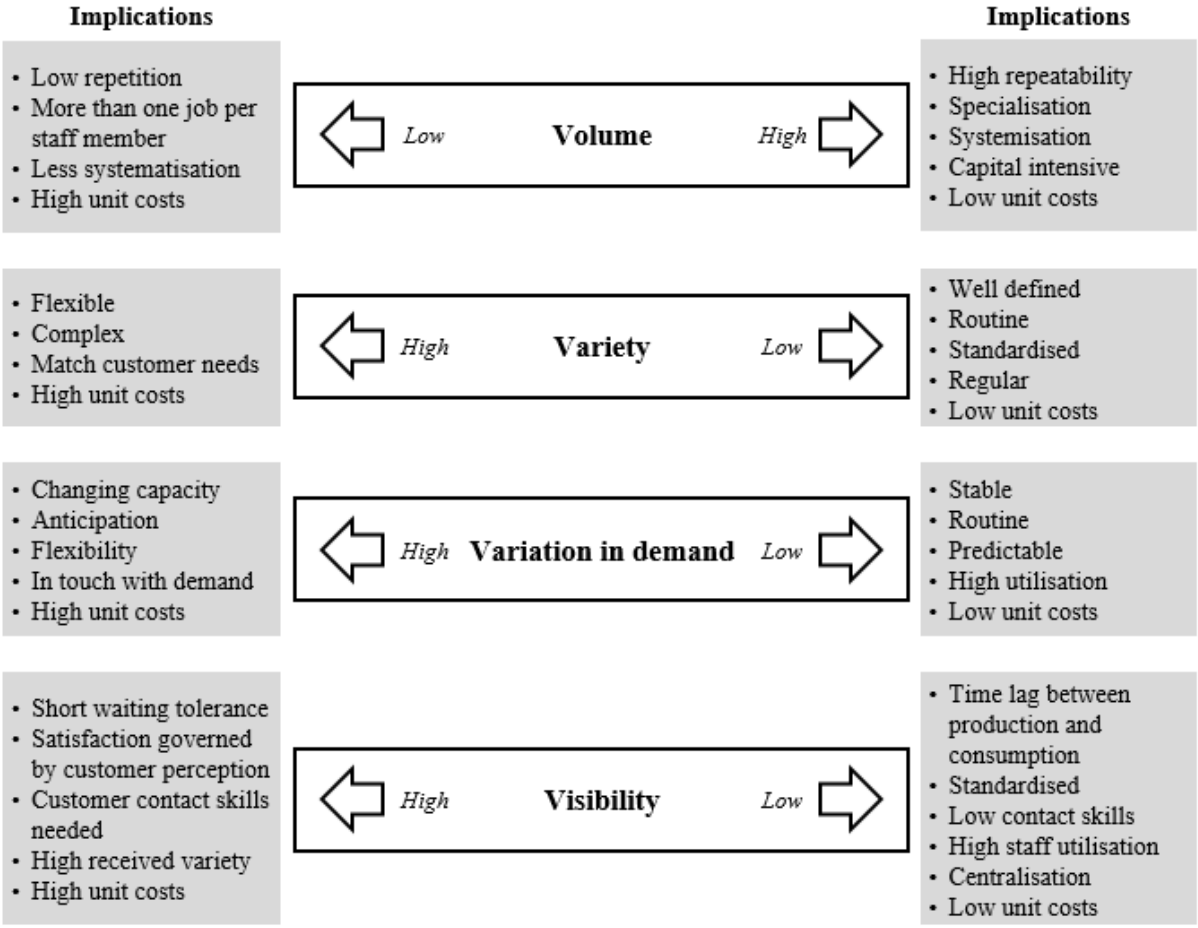


Figure 2.1: Overview of the 4 V's (Slack et al., 2013)

2.1.2 The performance objectives of healthcare operations

A common way for an operation to measure its performance is according to the basic performance objectives *quality, speed, dependability, flexibility* and *cost* (Slack et al., 2013). These generic objectives apply to any type of operation, and can be used to guide the organisation's focus and improvement initiatives. Of course, each operation will have to interpret the performance objectives to its own context and goals, and healthcare is no exception.

Quality

Quality is a wide concept that briefly can be regarded as the practice of 'doing things right' (Slack et al., 2013). Delivering high-quality products and services to customers help bring higher external customer satisfaction, and bring benefits to the internal operations by reducing cost for correcting mistakes and increasing the dependability of deliveries.

In healthcare, quality is important because it concerns giving patients appropriate treatments and carrying them out in a correct manner (Slack et al., 2013). Thus, healthcare customers can become more satisfied with their care and the tax payer can be sure to receive value for his or her money (Ross, 2014). Furthermore, emphasising high-quality care can limit the need for revisits (Modig & Åhlström, 2012), and ensure that patients are protected from mistakes of well-meaning personnel (Ross, 2014).

Speed

Speed can be described as the time elapsed between a customer request and the receipt of those products or services (Slack et al., 2013). Speedy deliveries mean that customer receives products or services quicker, increasing the likelihood that they are willing to pay, or increasing the benefit that they receive.

Healthcare operations are very dependent on speed (Belvedere, 2014). This fact is particularly true in emergency services, where increasing speed by a matter of seconds can determine whether a patient lives or dies (Slack et al., 2013). Thus, being able to deliver the correct care within minutes is a great advantage within healthcare. Moreover, speed can also be an advantage within non-emergency operations to minimise waiting times, e.g. when waiting for test results or x-ray pictures to come back.

Dependability

Dependability is doing things in time, such as delivering products or services when they are needed or promised (Slack et al., 2013). Dependability is important because delivering too early could mean that a product needs to be placed in inventory, running the risk of becoming old, or that a service is carried out before it is needed. Delivering too late runs the risk of supplying the product or service when it is no longer needed.

Dependability in healthcare could mean for example that appointments are carried out in time and not cancelled, or that test results are returned when they are promised (Slack et al., 2013). Moreover, achieving a high dependability can help save time and money through utilising resources more effectively, and keep the operation running with a higher degree of stability.

Flexibility

Flexibility is described as being able to change the operations and can be defined according to four general categories: (1) product/service flexibility, meaning how an operation can introduce new/modified products or services, (2) mix flexibility, the ability to produce a wide range or different kinds of products and services, (3) volume flexibility, governing how an operation can scale up or down its activity or output volume, and (4) delivery flexibility, being able to change when deliveries are carried out (Slack et al., 2013).

In healthcare, flexibility might mean that new types of treatments can be introduced (product/service), that a wide range of treatments can be supplied (mix), that the volume of treated patients can be adjusted (volume), and that the operation is easily able to reschedule appointments (delivery) (Slack et al., 2013). Moreover, a flexible healthcare operation can quickly transfer staff and equipment between departments, adapt quickly to changing patient conditions, and take care of sudden spikes in patient inflow without disrupting the rest of the operations.

Cost

Cost is a central performance objective for many operations, and is heavily affected by the choices being made in regards to the other performance objectives (Slack et al., 2013). A low internal cost structure can enable a company to successfully compete on price, or keep profit margins high to free up capital for new investments or dividends to shareholders.

Healthcare is a costly sector, requiring substantial expenses for technology, materials, facilities and staff (Slack et al., 2013). Moreover, a healthcare operation must be cost efficient (Belvedere, 2014). By reducing costs where possible, funds can be transferred to departments that require more money and in turn, the provided care can offer a higher degree of quality (Slack et al., 2013).

Trade-offs between performance objectives

Naturally, an organisation cannot be good at everything. Skinner (1969) exemplifies the concept of trade-offs by stating an airplane design analogy: *"For instance, no one today can design a 500-passenger plane that can land on a carrier and also break the sonic barrier"* (Skinner, 1969, p.140). In the same way, there might exist trade-offs between operations performance objectives (Slack et al., 2013). An easy way to illustrate how an operation performs is to draw the objectives in a polar diagram, illustrated in figure 2.2 where a taxi and a bus service are compared.

However, Slack et al. (2013) states that there are two general ways of viewing trade-offs. The first way is to regard an increase in one performance objective to come at the expense of another, also called 'repositioning'. The other way is to increase the organisation's

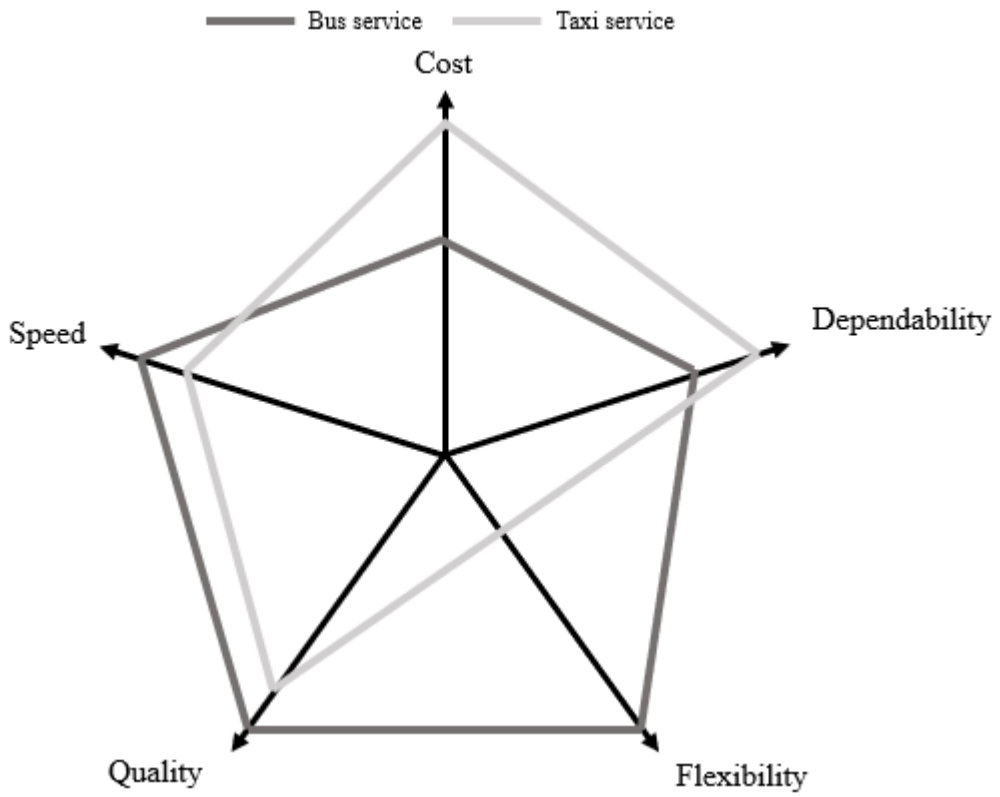


Figure 2.2: Polar diagram of performance objectives (Slack et al., 2013)

effectiveness and overcome trade-offs in a sense that several performance objectives can be improved simultaneously. Generally, an operation deals with both approaches.

2.1.3 Capacity management

A fundamental decision area in an operation is capacity, dictating the ability to provide sufficient supply to satisfy demand (Slack & Lewis, 2017). However, it is not as straightforward as simply providing a steady state of capacity. As demand fluctuates, an operation needs to adjust accordingly, leading to important decisions having to be made in order to get its capacity right and not incur losses in its competitive abilities. For example, too much capacity leads to underutilisation of resources and higher costs, while too little capacity fails to supply customer demand and impacts the operation's revenues and drives up cost here as well.

In healthcare operations, capacity management is difficult due to the processing and scheduling of patients, paired with issues regarding high resource utilisation due to capacity bottlenecks moving across the system (Boaden et al., 2008). Furthermore, hospitals tend to focus more on long-term aggregate capacity because of trends in demand and budgetary requirements. Consequently, it is difficult for managers in healthcare to respond to small changes in volume or variety.

However, healthcare operations can, like other operations, deal with varying supply

and demand, and there are generic strategies for doing so. Slack et al. (2013) outlines three generic approaches for coping with demand fluctuations: *level capacity plan, chase demand plan*, and *demand management*.

Level capacity plan

A level capacity plan attempts to ignore forecast demand fluctuations and deal with demand by setting processing capacity at a constant level during the planning period (Slack et al., 2013). Because services cannot be stored, a level capacity plan would imply that a service operation would run with slack resources during periods of low demand, usually leading to wasted staff resources and low productivity. Generally, a lower capacity utilisation can be accepted in high contact service operations, like healthcare, to be able to meet peak demand (Chase, 1978).

In healthcare, a level capacity plan can help contain costs at the expense of longer queues, but this implies that patients are willing to wait for the services (Boaden et al., 2008).

Chase demand plan

The chase demand plan is opposite to the level capacity plan and implies that company attempts to match its capacity to varying forecast demand, and is usually leveraged by operations producing perishable goods or services (Slack et al., 2013). Ways to adjust the capacity are e.g. adding or removing shifts, using part-time staff and subcontracting. In healthcare, the chase demand plan focuses on service quality and helps maintain low waiting times (Boaden et al., 2008). For example, the NHS (United Kingdom National Health Service) chose to start increasing its capacity during wintertime to cope with the higher demand.

Demand management

The third strategy for coping with demand is to try to adjust the demand itself (Slack et al., 2013). Two general approaches to this are price adjustments and providing alternative products and services. Since healthcare services rarely can be stored for later use, demand management can be an effective way to manage processes related to patient admissions and surgical schedules, based on the expected time a patient will spend at a hospital, and can lead to better quality and cost performance (Boaden et al., 2008).

An important part of healthcare demand management is the extent of 'failure demand' (Boaden et al., 2008). This is when services have to be provided again due to the customer not being satisfied the first time, or if an alternative service is not available at the time when a customer needs it. The reason for failure demand can be because of several system failures, e.g. when a system cannot provide a service or if a customers do not understand what the system expects from them. According to Kjøller and Westergaard

(2017), reducing failure demand helps an organisation shift its efforts towards solving the customer's needs, instead of focusing on the demand that the organisation created itself (failure demand).

2.1.4 Efficiency and effectiveness

When measuring a process, it is important to consider several measures to capture the process' performance (Holweg et al., 2018). Process performance can be divided into two basic categories: *effectiveness* and *efficiency*. Generally speaking, effectiveness concerns *doing the right things*, while efficiency concerns *doing the things right* (Alänge, 1994). In essence, there is no use working towards an efficient process if it is not effective in the first place.

Effectiveness it the external view of performance, describing if the process delivers what is expected by the customer (Holweg et al., 2018). With this definition, the effectiveness can only be determined by the customer alone.

Efficiency regards how good a process is at converting resources into outcomes (Holweg et al., 2018). Generally, it considers the notion of productivity, e.g. labour, energy, machinery or materials efficiency. A measure that is more complete but more difficult to assess it the total factor productivity, which consists of the ratio of the sum of all outputs to the sum of all inputs, illustrated by equation 2.1. As this is measure is a ratio, productivity can be increased by either increasing the output or reducing the inputs for a given output.

$$(Total\ factor)\ Productivity = \frac{\sum outputs}{\sum inputs} \tag{2.1}$$

In relation to services, internal effectiveness has a strong impact on performance objectives and cost (Slack et al., 2013). For example, high-quality operations rarely need to waste effort on redoing things. Fast operations maintain low in-process inventories and reduce administrative overhead costs. Dependable operations can be relied on to deliver what is planned and on time, eliminating wasteful disruptions. Flexible operations can quickly adapt to changing circumstances without impacting the other pars of the operations, while being able to switch between tasks quickly.

Some examples on how to increase efficiency are given by Sasser (1976). During peak hours, employees can be instructed only to perform tasks that are essential for service delivery, backed up by managers who can perform supporting tasks during slack times. Managers can also examine tasks performed during peak time and find out whether certain skills are inefficiently used or lacking, e.g. through the use of paramedics. A final way to manage peak capacity constraints is through cross-training of employees. If they learn to perform several tasks, employees in the system can switch from underused to strained stations and thus increase its capability to handle bottlenecks.

2.1.5 Variation and queues

Holweg et al. (2018) describe 10 principles in their book about process theory. The first principle states that all operations are composed of processes. The next step is about variation and how it inherited by different processes. Variation can occur in quantity, quality and timing. Additionally, variation can be buffered by the following three means: time, inventory, and capacity.

Buffering refers to decrease the impact of blocking and starving parts of an operation (Holweg et al., 2018). High variation impacts an operation and its processes in many ways, among these, it impacts the capacity and the queues. A way to represent the relation between capacity utilisation, variation and the length of queue is the 'Kingman Formula'. If the variation of a system increases, the utilisation rate decreases and the queuing time increases. A graphical representation of the Kingman Formula is shown in figure 2.3, illustrating the relationship between capacity utilisation (ρ) and queue length (lead time).

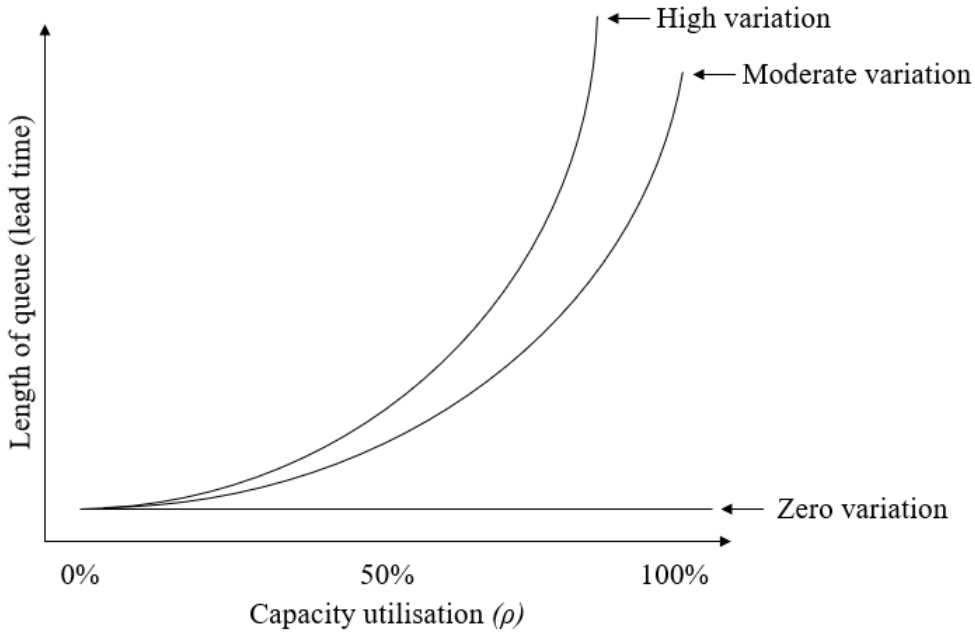


Figure 2.3: Graph of the Kingman Formula (Holweg et al., 2018)

Another measure used to judge processes is the throughput rate, which is the speed of a process (Holweg et al., 2018). It is the time it takes to journey from the beginning of a process to its end. Throughput rate is often targeted in operations management. Since being able to increase the throughput rate with the same resources means an increase in output with the equal input, this implies a higher profit. There are several ways of decreasing throughput time, some of them are: breaking up bottle necks, removing non-value adding steps and changing the physical layout of the system.

Modig and Åhlström (2012) describes the efficiency paradox and how it affects improvement work. From a organisation's perspective, maximising resource utilisation is usually desirable, however from a customer's perspective it is not. Maximal utilisation of resources implies no capacity for variation in demand and increased queuing time, therefore

it is not always desirable to maximise the utilisation. Especially in healthcare, increased queues can lead to inefficiency from a customer perspective in various ways. For example, when waiting too long for surgery, a patient has to leave work for a longer period. The authors mean that all needs create new ones, hence from a customer and society perspective healthcare needs to decrease throughput time.

2.2 Healthcare organisation

The healthcare organisation is a complex subject, and applying quality and process improvement approaches in healthcare is not entirely straightforward, partly attributed the healthcare sector's difference from other industries (Boaden et al., 2008). The stakeholder perspective is also discussed by Glouberman and Mintzberg (2001), and is discussed later in this section.

2.2.1 What differs healthcare from other sectors?

Since the healthcare sector is different from other industries, Boaden et al. (2008) outline differences related to e.g. the professional nature of healthcare, the complex internal and external structure of the healthcare system, and difficulties in succeeding in change efforts.

The professional nature of healthcare

Due to the professional nature of healthcare services, it is characterised by a high level of professional autonomy (Boaden et al., 2008). Illustrative of this, doctors have considerable discretion and autonomy due to the breadth and depth of knowledge in solving complex problems. They are also loyal to their profession, tend to work independently and usually have to be persuaded to support change initiatives. Currently, doctors tend to act as skilled craftpersons managing their own patient waiting lists, clinics and operations inside a mass production-focused general hospital (Jones, 2006).

The complex structure of healthcare

Healthcare is more complex compared to other industries working with quality improvement approaches according to Boaden et al. (2008). A wide variety of inputs are transformed to produce conceptually complex outputs, and the methodologies used to measure these are measured in a wide number of dimensions. There is a general lack of top-down, whole-system view, leading to difficulties in defining both the output and who the customers are.

Another factor adding complexity is that the healthcare system is financed by public tax money (Druckler, 2011). Healthcare institutions need to be independent from politicisation and governed according to the institution's own values. At the same time, healthcare policy making is one of the political areas with the highest magnitude, and governments are held accountable for the performance of the healthcare system. This gives politicians a high degree of responsibility, but a low degree of control over how the system performs.

The difficulty and complexity of change efforts

Boaden et al. (2008) describe several factors as to why changing healthcare is difficult. Since the system consists of several organisations with potentially conflicting missions, the assessment would require tackling issues from several dimensions.

The workforce in healthcare is made up of multiple different professions with different educations and training, e.g. doctors, nurses and engineers (Boaden et al., 2008). Due to this, these professional groups tend to be influenced by their background, leading to the groups identifying with their professions and not necessarily their organisations. Göransson et al. (2018) state that doctors and nurses tend to work for different managers, leading to difficulties regarding social affiliation and identification.

Lastly, both the internal and external environments are complex with multiple stakeholders such as patients, their families, governments and professional associates (Boaden et al., 2008). Druckler (2011) emphasises the importance of considering external factors, as a complex and chaotic system cannot exclude these factors. As the environment is complex, pressure to standardise clinical work while allowing for flexibility, paired with emerging evidence that supports or discredits current ways of working and variation in performance between units build upon the challenges of implementing change in healthcare (Boaden et al., 2008).

2.2.2 Healthcare stakeholders

Glouberman and Mintzberg (2001) explain the complexity of healthcare due to the different stakeholders and their underlying interests. These interests do not necessary overlap, hence the difficulty in governance. The authors start by describing the internal interest at a hospital and continue with describing the different stakeholders involved in hospital governance from macro perspective. In figure 2.4 managing is divide into four different directions: *out, up, in* and *down*. Managing out refers to people involved in the organisation but technically independent, up refers to managing towards people in control, in, towards units and people under clear control of the institution and down, into the operations that focuses on treatment of patients.

As figure 2.4 shows, doctors are relatively independent and focused on treating patients. Nurses are also concerned with taking care of patients, but in contrast to doctors, they are under clear control and cannot enjoy the same institutional freedom. The difference

between trustees and managers is the same as between doctors and nurse, with the exception that both are managing towards people in control and not patients.

Glouberman and Mintzberg (2001) take the framework in the left of figure 2.4 (General Hospital) and expand it into the framework represented on the right (Society). This new framework includes external stakeholders impacting the operations and leadership within a hospital. The Four Worlds of Society follows the same logic as in the General Hospital, and gives an explanation to why healthcare is one of the most complex systems known to contemporary society. The larger the difference is between different groups, the more their independent goals differ. Since the stakeholders are presented as very different, there is a need for extensive integration.

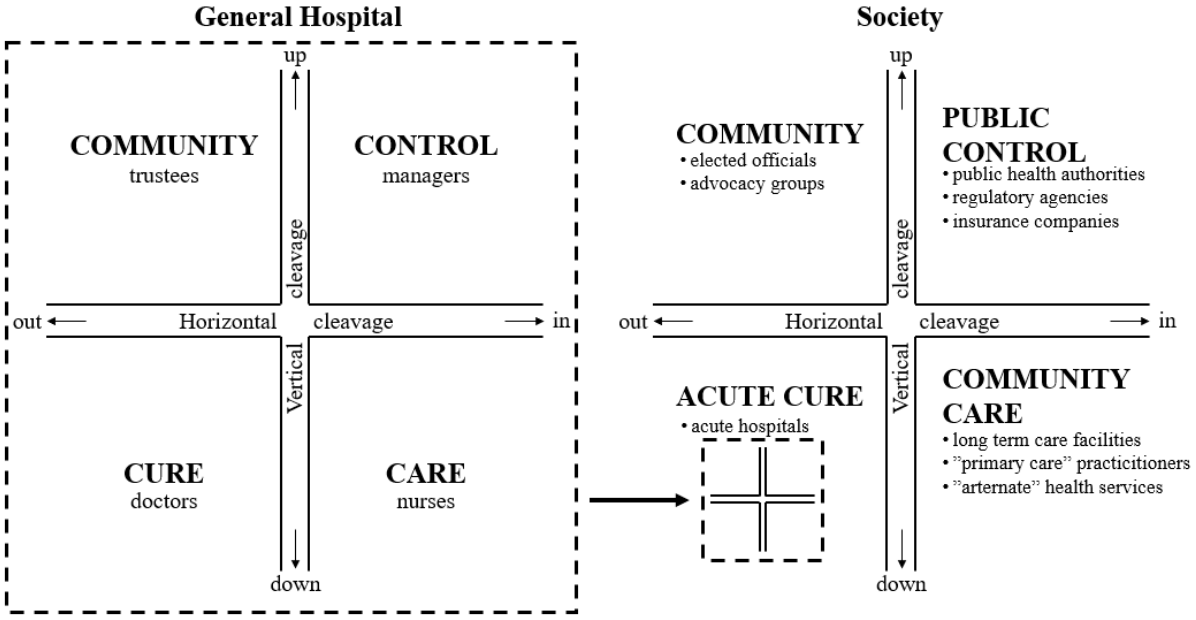


Figure 2.4: Four Worlds of the General Hospital and Society (Glouberman & Mintzberg, 2001)

2.3 Discrete-event simulation

Chung (2004) describes simulation modelling as mathematically modelling a physical system. Simulation is a comprehensive way to test hypothesises by experimenting with a model. It is also a way of gaining insights into a physical system without actually disturbing it. Different physical systems can be modelled, for example manufacturing systems and service systems like warehouses, hospitals and medical clinics.

Jingshan et al. (2017) describe several use cases of simulation in healthcare. The authors believe that simulation modelling is well suited for some of the complex systems within healthcare, and that a more scientific strenuous approach towards solving these problems is a trend that will continue.

2.3.1 Simulation methodology

Chung (2004) describes the practical agenda of simulation modelling in nine steps: problem formulation, project planning, system definition, input data collection, model translation, verification, validation, experimental design, analysis and presenting results and conclusion. Out of these nine steps six, slightly modified, steps are presented more in depth below.

All steps are a important part of a successful simulation modelling project, however the earlier a mistake is made the more difficult it becomes to rectify it. Hence, the system definition and data collection are crucial to provide a successful model translation. Since the later steps still are possible to conduct even if mistakes are made in previous steps, verification and validation cannot be emphasised enough (Chung, 2004).

System definition

The system definition sets the boundaries for the future model (Chung, 2004). In this step the system is classified, i.e. decided on how much of the physical system that should be included, what events to include and what data to collect. It is important to understand whether the system is of discrete, continuous or combined nature. Furthermore, understanding whether it is terminating or non-terminating.

The conceptual model is a high-level flow chart representing the physical model. This model sets the foundation for further modelling. It illustrates where input and output enter and exit the system, it also demonstrates what kind of data that needs to be collected. The practitioner should endeavour to construct a well representative conceptual model, yet not a too detailed one (Chung, 2004).

Empirical data and analysis

To be able to translate the conceptual model into an actually functioning simulation model data needs to be collected (Chung, 2004). The data represents the different activities in the conceptual model. For example, to model a process, the process needs to be assigned, for example, a probability distribution in order to be simulated. If not, the model does not know how long a certain entity lasts in the process. All activities, described as processes in the model, need to be described by underlying mathematical functions. Therefore, data needs to be gathered for every activity.

After the data is gathered it needs to be fitted (Chung, 2004). Random samples cannot be used, however they can be utilised to create distributions. To fit data, different software can be used, but using Excel can be fully sufficient. To generate appropriate distributions it is important to have enough data points. If not, the distributions can contribute to a misleading model.

Model translation

After understanding the physical flow and fitting data, the system (conceptual model) can be translated into a computer based model (Chung, 2004). If the prior steps are not conducted, the practitioner does not know what to include. In this step it is important to choose a simulation software that the practitioner is comfortable with. When starting to translate the model, it is important to follow a strategy that prohibits building a too complex, possibly erroneous system. By starting in a certain section and building in an add-on manner, also called a divide-and-conquer approach, it becomes easier to test the system during the build up. This is one of many strategies.

Verification

To enable that the simulation model works properly, it is important to be certain that the model is a good representation of the physical system (Chung, 2004). Verifying the model is not a one time check, but rather a continuous approach. Therefore, verification is included in the model translation. By including representations of queues and processes, in terms of pictures, it becomes easier to understand the model and make sure that it represents the physical system.

Validation

Validation is understanding whether the model actually performs as intended (Chung, 2004). There are two types of validation: face validity and statistical validity. Validation can tell the practitioner if the model performs correctly however it does not necessary tell what the defect is.

Validation is an important step after verification, since verification does not test for assumptions, simplifications, oversights, etc. (Chung, 2004). Face validity is achieved by the assistance of a domain expert. Meaning, if the expert does not see a difference between the physical system and the model, face validity is reached.

Statistical validity is reached by statistical comparison between output data from the physical model and the simulation model (Chung, 2004). Statistical validity can be conducted on individual entity data or entire system data. If statistical validity cannot be reached, the model needs to be examined for flaws. If not, the model can generally not be used for experimental design.

Experiment design and analysis

The experimental design looks slightly different if the system modelled actually exists or is non-existent (Chung, 2004). If the system is non-existent, a reference system needs to be created. Given existing or non-existing systems, the experimental design can be

of one-factor, two-factor or multiple factor design. An increase in factors contributes to a more complex analysis. To clarify, factors can for example be number of clerks or queue policies. To counteract an increasing complexity the levels of each factor should be limited.

After the experimental design, an analysis of the experiment usually follows (Chung, 2004). Depending on the experiment and the predetermined confidence level, different statistical tests should be conducted. Further, if the model is non-terminating, the steady state needs to be determined, often by using linear regression. In non-terminating systems auto-correlation could become a problem. This could occur due to subsequent runs depending on previous runs. To cancel auto-correlation, a non-terminating system can be run for a longer time and be divided into batches.

2.3.2 Pros and cons of simulation

Chung (2004) mentions three distinctive advantages with simulation modelling: experimentation in compressed time, reduced analytic requirements and easily demonstrated models. The simulation model is modelled on a computer which means that once the model is constructed experiments can be made in compressed time. With the software aiding both modelling and experimenting the analytical requirements of the practitioner are no longer as high as they once were. The modern software also includes animation making it easier to demonstrate the model. In addition to this, the model, as mentioned above, does not disturb the physical system.

Chung (2004) further explains disadvantages with simulation modelling: simulation cannot give accurate results when input data is inaccurate, simulation cannot provide easy answers to complex problems and simulation cannot solve problems by itself. If the input data is inaccurate no model can generate a just foundation for implications, also the model does not solve any problems itself, a well functioning simulation model can only be of help if the practitioner knows how to conduct relevant experiments and convert these into implications. These implications in turn can only improve the physical system if they are implemented.

3

Methodology

This chapter describes the method used to investigate the research questions. It aims to give the reader a comprehensive understanding of the work carried out for deeper understanding and potential additional research.

3.1 Introduction

Chung (2004) describes simulation as a technique for mathematical modelling of physical systems. Regarding the research questions of this thesis, simulation was regarded as an appropriate tool to find possible solutions based on theory and experiments, which then in turn helped answer the questions. Simulation provided both the ability to mapping the operation in detail, as well as the ability to experiment with the system by adjusting parameters.

In accordance with Wallén (1993), this is a descriptive study with normative tendencies. Due to mapping out the operations of the ENT clinic and modelling the flow of patients, this study utilises a descriptive method. Experimenting with different scenarios and running them in the model also included normative elements.

Given that this study considered existing problems at the ENT clinic by analysing collected data and applying an iteratively developed theoretical framework, it also includes some inductive elements. Wallén (1993) describes that inductive studies draw conclusions upon insights in data rather than by using abduction. One of the implications of induction is that empirical data directs the extent of the literature review.

3.2 Theoretical framework

The theoretical framework was created to enable a stringent analysis of the empirical data. It was later leveraged to fully analyse and understand the findings presented in the study. For example, results that were found during data collection and analysis were compared with literature from the framework in the discussion chapter. By doing so, the

authors could enable a more robust analysis and a more reliable result. The approach is well aligned with what Wallén (1993) describes as part of an inductive approach.

Recommended literature by expert guidance and literature from Chalmers University's library, Google Scholar and other internet sources were used when creating the theoretical framework. These sources helped in creating three subsections in the framework: *operations planning and control, the performance objectives of healthcare operations*, and *simulation modelling.*

3.3 Data collection

Data was collected throughout the project with the highest intensity during the early stages. To be able to model the ENT clinic's emergency flow, it first had to be mapped out and then be thoroughly examined to clearly model all the different steps. To enable this, different data sources were used. Interviews and observations provided primary data and secondary data was extracted from the ENT clinic's database.

Due to external reasons that could not be controlled, namely the COVID-19 pandemic, precautionary measures were taken by the hospital halfway through the project, restricting access to the clinic and limiting the ability to collect quantitative data. As a result of this, more assumptions had to be made, e.g. around logical sequences and process times, which were later verified with interviews.

3.3.1 Interviews

Interviews were mainly used to understand the system and verify that the mapping was conducted in a rightful manner. The interviews were not structured but adjusted according to what the interviewed candidate had to say. This kind of unstructured interviews are supported by Bryman et al. (2005). Furthermore, the authors claim that rich answers are preferred, with the options for candidates to complement these in additional future interviews.

The initial interviews were conducted face-to-face at the clinic with both physicians and nurses, and acted as a base for the conceptual model. Questions were asked about the patient flow layout, the work in each process and rough estimates of process times. As access was restricted due to precautionary measures concerning COVID-19, later interviews had to be conducted over telephone. These later interviews focused on clarifying assumptions made in the model, collecting missing data about process times and logical flows, and validating output from the simulation model. In total, six people were interviewed about the system, presented in table 3.1. Some people were interviewed more than once, e.g. Nurse A (Clinic Manager) being consulted throughout the whole project from conceptual modelling to providing input to the recommendations.

Description of interviewees		
No.	**Role**	**Medium**
1	Receptionist	Face-to-face
2	Nurse A (Clinic Manager)	Face-to-face and telephone
3	Nurse B	Face-to-face
4	Nurse C	Face-to-face
5	Physician A	Face-to-face
6	Physician B	Face-to-face
7	Physician C	Telephone

Table 3.1: Description of interviewees

3.3.2 Observations

According to Wallén (1993), observations generate knowledge that does not emerge in interviews. It is knowledge that people in possession of take for granted and do not necessarily mention. This type of knowledge is also called 'tacit'. Therefore, observations are regarded as a complement to interviews.

In this case observations played an important role, since they are fundamental for work sampling. As Kumar (2011) mentions, observations are suitable when behaviour is demanded rather than perception. Furthermore, Kumar (2011) tells that there are participating observations and non-participating observations. In this case, non-participant observations were relevant due to the need for process times without disturbing the work at the ENT clinic. Observations were also used to understand the logical sequences in the emergency flow.

Work sampling is a time study, often applied to determine the specific time of various activities for an operation. Work sampling is built upon the belief that a sufficiently large random sample can predict the behaviour of the entire group (Barnes, 1956). It can provide the necessary information to calculate a certain utilisation rate or, as in this case, the required information to model the operation. To be able to create a valid conceptual model, activities have to be classified properly. The practitioner needs to be aware of pitfalls like classifying several activities with different treats as one (Chung, 2004). Subsequently, the right times for the activities were measured by timing.

The time study was initiated by collecting smaller samples from the different process to aid in the initial modelling. Processes were timed with the help of a stopwatch, with the observer placed in the same room as the process was carried out. The times for each process was measured from beginning to end, with the ENT staff informing as the process was started, stopped or put on hold. No further divisions were made, as breaking down the processes further were not deemed necessary to answer the research questions. Doing so would have made the simulation model too detailed and difficult to manage for the purpose.

As mentioned in the interviews section, as access to the clinic was restricted, observations

could no longer be carried out. In this case, interviews were used as a substitute method for observations. The data that had already been collected was then used in combination with interview results, to act as a base for the distribution functions entered in the simulation model.

3.3.3 Secondary data

Secondary data was used to the extent where it was provided by the organisation and deemed to be appropriate for the purpose of simulation. The advantage of secondary data was that it saved time in the project and provided the simulation model with large samples. An example of secondary data collected was an overview of all the ENT clinic's admitted and discharged patients during recent years.

3.4 Conceptual model

In line with the approach of Chung (2004), the first productive step was to create a conceptual model of the operation of the ENT clinic's emergency flow. To do so, the system had to be classified. The classification consisted of determining whether the system was discrete, continuous or combined and terminating or non-terminating, which resulted in a discrete, terminating system. Post hoc, the system was mapped out on a high level flow chart base. In consent with the approach of Chung (2004), the mapped out system was validated with the help of people familiar with the operations of the ENT clinic.

System validation was carried out with the help of unstructured interviews, where several nurses and doctors were shown the conceptual model and had the opportunity to provide input and amendments. The interviews also helped discern what sub-processes each process contained and weather it was feasible to divide the processes. This validation is further described in section 3.7.

3.5 Data analysis

Chung (2004) describes the next step in simulation modelling as data collection and analysis. Therefore, data was gathered by imposing work sampling principles. The data was fitted to matching probability functions, given the requirement of it being non-deterministic data. For processes that lacked data due to COVID-19 constraints, probability distributions were created with the help of estimations from the staff.

3.5.1 Primary data

The first step in handling the collected data was to judge whether the data points collected from each process were independent (random) from one another. Law (2014) described scatter plots as a suitable method for assessing a sample's independence. The observations $X_1, X_2, ..., X_n$ were plotted in pairs such as (X_1, X_{i+1}) where $i = 1, 2,...,n-1$. If the points are scattered randomly in the first quadrant of the plane made up by (X_1, X_{i+1}), the points can be expected to be independent. However, if the points are scattered around one or more sloped lines, one can expect the X_i points to be correlated.

When the collected data had been assessed for independence, the data points were plotted in histograms to check for local modes and to estimate a hypothesis of a density function corresponding to the data points' distribution, as described by Law (2014). Several parameters such as range, mean, standard deviation, variance, and coefficient of variance were calculated from the collected data points to aid the estimation of probability distribution. The nature of the processes were also taken into consideration, since certain processes are characterised by general treats and can often be described by a limited number of probability distributions, leading to the practitioners not having to investigate all possible distributions.

After an initial hypothesis had been made regarding a suitable distribution function, the parameters for the given probability distribution were calculated. Depending on the probability distribution these parameters are calculated differently. Since the Gamma distribution, Weibull distribution and the Log-normal distribution are similar in behaviour and all well suited to describe process times to complete a task (Law, 2014), the Gamma distribution was preferred due to the approximation of the maximum likelihood estimation of its parameters. It is displayed in equation 3.1. To find the approximation of the parameters for the function, $\hat{\alpha}$ and $\hat{\beta}$ were approximated using equation 3.2 and the values corresponding to the calculated T value were found in the table presented in Law (2014, p.386).

$$F(x) = \begin{cases} 1 - e^{-x/\beta} \sum_{j=0}^{\alpha-1} \frac{(x/\beta)^j}{j!}, & \text{if } x > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

$$T = [ln\bar{X}(n) - \sum_{i=1}^{n} lnX_i/n]^{-1} \tag{3.2}$$

In the cases where it was determined that primary data collected through initial interviews and observations was missing or incomplete due to limited access to the ENT clinic, it was determined that further interviews had to be conducted with physicians and nurses. Although it would have been preferable to conduct more observations or rely on third-party data collection, the circumstances didn't allow for neither case, creating the need to continue in an alternative way. Thus, interviews were determined to be the only suitable way to collect further data.

In case of processes where data had been collected but was determined insufficient, the

Excel plugin NtRand3 was used to assist in finding suitable parameters for probability distribution functions. NtRand3 could only be used if the mean value, coefficient of variance and min value were known or could be estimated by collected data.

Since there was no difference in convenience between the probability distribution functions when estimating parameters using NtRand3, paired with the Weibull distribution's offset rate being included in SIMUL8 by default, the Weibull distribution was chosen. If the cut-off rate, also known as the min value, would not have been included by default in SIMUL8, the practitioner would have to merge two functions to describe the process time, making the model more difficult to understand for people less familiar with simulation modelling. The cumulative density function for the Weibull distribution is described in 3.3, presented by Devore (2012).

$$F(x, \alpha, \beta) = \begin{cases} 1 - e^{-(x/\beta)^{\alpha}}, & \text{if x} \geq 0. \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

In cases were a very limited amount of data was collected, the interview subjects were asked to provide minimum and average process times, as well as a lower 10% boundary to act as a base for calculating parameters for Log-normal distribution functions. The Log-normal function was used for its simplicity given the circumstances. The standard deviation could be derived with a lower, a min value, and an upper bound of the estimated 10% that process times would lie within as well as the average value. Using a Z-table and the standard deviation σ, adjusting for the min value, could be calculated using equation 3.4 presented by Devore (2012, p.174). The parameters were then assigned to a Log-normal distribution in SIMUL8.

$$F(x; \mu, \sigma) = P(Z \leq \frac{ln(x) - (\mu)}{\sigma}) = \Phi(\frac{ln(x) - (\mu)}{\sigma}) \tag{3.4}$$

Goodness-of-fit

To check for goodness-of-fit, Law (2014) outlined several methods. The method used to compare the distribution functions with the histograms was the chi-square test. By comparing the actual observed frequency (O_i) of measured values with the expected frequency (E_i) within specified intervals, it could be assessed whether the distribution function was valid or if the hypothesis had to be rejected. The X^2 was referenced to critical values for Chi-Square distributions, regarding k-1 degrees of freedom and α, the significance level, to determine whether to reject the hypothesis or not. A low X^2 value would indicate that the distribution function was a likely estimation. The formula for assessing the goodness-of-fit is shown in equation 3.5.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

Further assessment of goodness-of-fit was not conducted due to the lack of data and the limitations accompanying this. If this would not have been the case, different types of plots could have been suitable, for example a Q-Q plot or P-P plot as advocated by Law (2014). It would also have been reasonable to use both, since they amplify different differences between the model distribution function and the empirical distribution function. However, if there is limited data, the graphical comparison becomes less accurate, and therefore it was not conducted.

3.5.2 Secondary data

The secondary data provided from the Hospital was analysed mainly through the use of manual calculations and Microsoft Excel. A large data file consisting of all the patients admitted to the clinic since January 2017 up until March 2020 was provided by the hospital and acted as a foundation for the analysis.

Since the data file consisted of many database keys and acronyms, two expert interviews were conducted with representatives from the hospital to provide an understanding of what to search for in order to extract the data that was relevant to the ENT emergency flow specifically. Key figures used were the total number of patients treated each day and splitting the patient numbers into two-hour time intervals representing time slots when the clinic was more or less busy. The number of patients per time slot were then translated into exponential functions, with inter-arrival times used as a parameter, for each entry point. To enable this, the secondary data had to be complemented with primary data. Finally, the data was anonymised, aggregated and entered into the simulation model.

3.6 Simulation model

With the conceptual model as a base, the simulation model was built and programmed in the simulation software SIMUL8. The software provided a programming environment with predefined building blocks, and allowed for a great deal of customisation through using its visual logic environment, which was used for defining logical sequences and custom key performance indicators (KPIs).

Before collection of quantitative data had commenced, the model was built as an 'empty shell' consisting of entry points, queues and processes without assigned distribution functions. The model also consisted of some logical sequences. As data was collected, through interviews, observations and secondary data, and analysed the simulation model was revisited continuously and programmed with logical sequences and distributions for entry points and processes.

3.7 Verification and validation

This section addresses both verification and validation. Both processes are crucial when building a successful simulation model according to Chung (2004).

3.7.1 Verification

According to Chung (2004) verification often gets confused with validation. The author claims that the main difference between the two is that verification verifies that the model works as intended and validation insures that the model represents reality. Mentioned are also different methods to verify the model whilst building it. Some of them are animation, divide-and-conquer and writing to output file.

This simulation model was build by a divide-and-conquer approach, routs where created by entry and examined in order to find early mistakes. It was also continuously verified by using animations, always being able to see how different categories of patients moved throughout the model, helping prohibit for example false logical sequences. In addition to this, the output of several KPIs were examined to determine whether changes of the system were implemented correctly.

3.7.2 Face validity

Wallén (1993) mentions that validation is to determine if the measured result is the same as what was initially intended to measure. Bryman et al. (2005) agree with this notion, yet they divide between internal and external validation. Internal validation examines if the causality of a conclusion is rightful or not, and external whether the results can be generalised outside of the specified research context. In a simulation context, face validity and statistical validity are emphasised (Chung, 2004).

Face validity was assessed through interviews with people familiar with the system, i.e. physicians, nurses and the Hospital's logisticians. The interviews consisted of questions related to both the model's structure, logical sequences, input and output data. Interview subjects were walked through the model during individual interviews and asked to confirm or deny whether the flow of patients was correct and if the processes and queues prioritised patients in the right way. In regards to the input and output data, questions were asked whether e.g. the number of patients treated each day and inter-arrival times were consistent with the real-life system. This process of validation was conducted in an iterative manner as the model became more detailed and complete.

3.7.3 Statistical validity

Statistical validation was difficult to assess in the system for two main reasons. First, the hospital does not conduct measurements of throughput times, process times, queue contents or resource utilisation. This led to a lack of secondary data to statistically validate the model's results. Second, due to the aforementioned circumstances regarding access to the clinic and subsequent data collection difficulties, there was a lack of comprehensive primary data to measure against. Thus, a thorough statistical validation was not performed, leading to uncertainties about what kinds of statistically accurate conclusions can be determined with the model as a sole basis.

3.8 Future state model

One of the research questions was focused on suggesting alterations for greater efficiency in the emergency care flow. In this section, the process of designing a 'future state model' is presented along with steps taken to assess the model's sensitivity to different types of variation.

3.8.1 Factorial analysis

A main goal of experimental design in simulation is to assess which factors have the highest effect on the result with the least amount of simulating (Law, 2014). Based on this, the factorial analysis was designed with the idea to judge the effects of the base model's resources different KPIs, and thus discover which of the resources were the most critical.

In the base model case, four factors connected to the resources were investigated: *Nurse 1, Nurse 2, Nurse 3 in combination with Physician 1*, and *Physician 2*. The resources are described more in detail in the Results chapter, table 4.3. Each factor was set to one low and one high configuration (-/+), with the low configuration meaning 1 resource (base case) and high meaning 2 resources (one extra). According to Law (2014), one of the main advantages with simulation is to comparing alternative systems before implementation. Since the resources of this thesis' simulation model were altered between trials, it was judged that each trial was comprised of a unique system.

To determine the effects between different systems, a Two-Sample-t Confidence Interval can be utilised to assess whether the effect of a system can be considered to be different from another system (Law, 2014). An advantage of forming confidence intervals for differences in effects, compared to a hypothesis test, only assessing whether to reject the null hypothesis or not, is that the confidence interval not only provides information whether the effect is significantly different from zero, but also quantifies the difference between systems.

According to Chung (2004), if more than two comparisons are involved the t-test should be replaced by the analysis of variance (*ANOVA*). The *ANOVA* determines if there is a significant difference between the samples, therefore, if there is a difference a post hoc test/multiple comparison procedure needs to be carried out in order to determine which means are statistically significantly different. There are several different post hoc tests. Chung (2004) describes the *Ducan multiple-range test* and Devore (2012) *Tukey's procedure*. The authors of the thesis chose *Tukey's procedure* if there was homogeneity of variance, however, if the homogeneity was violated *Tamhane's T2 multiple comparison test* was applied.

First, the required number of runs (n) was calculated using the formula described by Chung (2004), presented in equation 3.6. The number of runs was derived by choosing a 95% confidence level as well as setting the Standard Error (SE) to an acceptable level given the characteristics of the physical system. In equation 3.6, (s) is the standard deviation of the replication means and (SE) an acceptable deviation in throughput times. Second, it had to be assessed whether the individual results were normally distributed. According to Law (2014), this should not be a problem due to the test being comprised of a large number of individual observations. Also, the Central Limit Theorem states, that a sample mean, derived from a large enough sample, will approach a normal distribution. This is valid for any population with independent random variables and a finite standard deviation (s).

$$n = (t_{1-\alpha/2,n-1} * s/SE)^2 \tag{3.6}$$

When comparing two samples, the *Welch's t-test* was performed and the *Welch Confidence Interval* calculated using the method described by Law (2014). First, the trials' mean values ($\bar{X}$) and standard deviation (S) was calculated using equations 3.7 and 3.8.

$$\bar{X}_i(n_i) = \frac{\sum_{j=1}^{n_i} \bar{X_{ij}}}{n_i} \tag{3.7}$$

$$S_i^2(n_i) = \frac{\sum_{j=1}^{n_i} [X_{ij} - \bar{X}_i(n_i)]^2}{n_i - 1} \tag{3.8}$$

for i = 1,2.

Second, the trials' estimated degrees of freedom $\hat{f}$ were calculated using equation 3.9.

$$\hat{f} = \frac{[S_1^2(n_1)/n_1 + S_2^2(n_2)/n_2]^2}{[S_1^2(n_1)/n_1]^2/(n_1 - 1) + [S_2^2(n_2)/n_2]^2/(n_2 - 1)} \tag{3.9}$$

Finally, the confidence interval was calculated using equation 3.10.

$$\bar{X}_1(n_1) - \bar{X}_2(n_2) \pm t_{\hat{f},1-\alpha/2} \sqrt{\frac{S_1^2(n_1)}{n_1} + \frac{S_2^2(n_2)}{n_2}} \tag{3.10}$$

When comparing more than two samples, the *ANOVA* and subsequent post hoc test where calculated in IBM SPSS Statistics. The *ANOVA* test is conducted by first calculating the means square for treatments, using equation 3.11. I represents the number of populations being compared and J the number of observations in each sample. Secondly, the mean square of errors is calculated using equation 3.12. Here, I is still the number of populations being compared and S^2 the sample variance. Finally, the F statistic, using equation 3.13 (Devore, 2012).

$$MSTr = \frac{J}{I-1}\sum_{i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \tag{3.11}$$

$$MSE = \frac{S_1^2 + S_2^2 + ... + S_I^2}{I} \tag{3.12}$$

$$F = \frac{MSTr}{MSE} \tag{3.13}$$

If the *ANOVA* resulted in differences of means, either *Tukey's procedure* or *Tamhane's T2 multiple comparison test* was carried out in IBM SPSS Statistics depending on the homogeneity of variance, as mentioned above. The logic behind these post hoc tests is similar to *Welch's t-test*, in terms of comparing means, and therefore not displayed in detail in this thesis.

Irrespective of the number of samples that were compared, when the appropriate statistical testing was conducted, each individual case, later in the report described as Trials, was compared to determine which factors had the largest impact, and which adjustments would be feasible in a real scenario. This factorial analysis was conducted for all KPIs to generate an overall picture of the impact of each resource.

3.8.2 Future state model design

After performing the factorial analysis, the impact of each resource on the physical system was assessed. Starting with the current state model in SIMUL8, alterations were made to the model's resource schedule and allocation to allow for task sharing and flexibility. The results were then compared to the base model by comparing the KPIs which are outlined in the Results chapter.

3.8.3 Sensitivity analysis

Due to a lack of statistical validation, it was determined that a sensitivity analysis had to be conducted, to see how variation in the most uncertain process times affected the measured KPIs. In agreement with one of the thesis group's supervisors, the process distributions with the least amount of underlying collected data were tested.

The sensitivity analysis was conducted by increasing the variation in the processes that were deemed to hold the highest uncertainty, i.e. the *Fax referral assessment*, *Fax referral registration*, *Further assessment*, and *Further treatment* (described in table 4.2), since these Log-normal process time distributions were approximated through interviews. Variation was doubled in all four processes, after which the total throughput time and utilisation were measured to assess whether the change in variation had a significant impact. This was statistically tested in the same way as described above in the factorial analysis section.

3.9 Reliability

Pruzan (2016) argues that reliability refers to the consistency of a measurement. Applying this to simulation modelling implies that the output of several runs of the model generate similar results. Combining validity and reliability generates a robust result. If the simulation model generates results that are well aligned with the actual output of the physical operations, and they are consistent with a low variance between runs, it implies high validity and reliability.

To determine the reliability of the model, it was run for the same number of trials as during the factorial analysis. The size of the confidence interval was qualitatively assessed, and a quantitative sensitivity analysis was performed. The sensitivity analysis indicates how sensitive a certain result is in regard to changes of certain factors, variances of processes in this case. It aids in understanding how reliable the KPI values are and if the confidence intervals of, for example, throughput times are applicable to the analysis.

3.10 Ethics

Since this study was conducted in the healthcare sector, ethics played a central part in how the research was carried out. Since some information and processes were of a sensitive nature, not following an ethical approach might have given negative implications on e.g. patient security, data integrity and the possible impact of the studies recommendations.

Confidential information pose an ethical dilemma since asking for the information might place the respondent in a difficult situation, and might also have legal implications for both the interviewer and respondent (Gillham, 2008; Kumar, 2011). However, not asking for the data might lead to the study not yielding any significant results. Therefore, information was carefully considered and questions that might have invaded privacy were carefully asked. In any case, collected information was be anonymised and handled carefully to ensure that the respondents could not be identified from this study during or after it was published.

Regarding interviews, it is important that they are carried out with respondents that are not feeling forced to participate (Berg, 2009). Even if participation is presented as

being voluntary, an employee might feel forced to participate due to e.g. pressure from a superior or peer pressure. Therefore, respondents were informed that participation was voluntary and that they could choose to end the interview at any time.

3.11 Sustainability

Sustainable development has been defined by the World Commission on Environment and Development in the Brundtland report as ”development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” (World Commission on Environment and Development, 1987, Chapter 2). The concept can be divided into three areas: (1) ecological, dealing with the environmental impacts of development, (2) social, describing well-being, happiness and human rights, and (3) economical, which can be described as both meaning economic growth as well as economic development that does not negatively impact ecological or social sustainability (KTH, 2019).

As the context of this thesis is within the healthcare sector, social and economic sustainability were of the highest relevance. The project has dealt with planning of resources, which affects how funds are distributed. By making processes more effective, more patients can be treated for less money, and resources can be distributed to other divisions or sectors in need. The ecological perspective was deemed to be less impacted by the thesis' work, as ecology is not central to the posed research questions. However, there might be indirect positive effects on ecological sustainability as resources are freed, redistributed and higher utilised, and waste is minimised.

4

Results

This chapter describes the results collected from the current state model and data analysis. Next, results from the experimental design are presented, after which a desired future state model is presented containing slight amendments to the base model.

4.1 Current state

This section contains a description of the current ENT emergency flow and its identified operations characteristics, along with a conceptual map and the results from running the simulation model.

4.1.1 System description and operations characteristics

The results from the interviews and secondary data analysis indicated that the system deals with a moderate volume of patients, averaging around 20 patients each day who receive treatment, not including the patients that are turned away in the early stages of the patient flow. Even though a considerable amount of patients enter the system, a large share is directed out of the system due to them not being in need of specialist emergency care at the ENT clinic, but should rather seek medical assistance at their local care centre. According to several of the interviewed nurses, this share of patients is responsible for a considerable amount of volume, leading to extra work having to be carried out in the beginning of the system.

The variety of conditions treated in the emergency care flow is very high, ranging from alien objects stuck in any of the ENT pathways to respiratory issues and broken noses. Secondary data covering admitted cases during recent years confirms this as there were over 2200 diagnosis codes covered during the last three years. According to interviewed nurses and physicians, variety is also deemed to be higher than in the elective care flow since the latter is divided into sub-specialities such as oto (ear), rhino (nose), tumours etc., making the emergency flow a suitable place for training and development of junior physicians. However, since the clinic provides specialist care to ENT patients, the variety is not as high as it would have been at a general emergency clinic treating a wider range

of conditions.

The general variation in the system is considered to be high in regards to patient types, arrival and process times, and ways of working. As a result of the large variety of conditions treated and the fluctuating nature of emergency cases, the variation in demand for different treatments is high, and it is thus difficult to plan ahead in terms of capacity and specialised physicians. In addition to the variation in conditions, the number of patients seeking care every day also varies, both in between days and in terms of certain time periods such as early mornings and afternoons.

Results from both interviews and observations of the processes indicate that individuals tend to work in different ways and at varying speeds. While different conditions require different actions and treatments, there are notable contrasts depending on who is staffed on a given process at a given time. As an example, some nurses tend to reject a smaller share of potential patients early in the flow, resulting in larger queues and workloads for the physicians.

The visibility in the system is high as patients experience the service provided directly. Throughout the majority of the flow, the patients are treated in processes and wait in physical queues. However, before the patients arrive at the clinic, they can be part of less visible processes such as referral handling by fax, phone queues and internal booking processes. Some discussions and consultations made by physicians are also conducted without the patient being present.

4.1.2 Conceptual map

After conducting interviews, work sampling by timekeeping and observing the distinct processes in the emergency flow, it was possible to map out the entire system. The scope of the mapped out system spans from patients entering the system in five different ways to exiting after being treated. Figure 4.1 illustrates the emergency patient flow at the ENT clinic. Figure 4.2 gives the reader a general overview of the physical layout of the clinic.

Patient types

To fully understand the logic of the system, how patients move throughout the system, it is crucial to understand the different types of patients, as a wide variety of patients enter the system every day. In the interpretation of the physical system, different patients are segmented by three types: (1) the patients that are not in need of emergency care, (2) patients that are in need of emergency care but do not need to be prioritised while in the system, and (3) patients that are in need of emergency care and need to be prioritised within the system.

The first type of patients exit the system as soon as the medical staff determines the current state of the patients and never proceeds to meet a physician. The other two

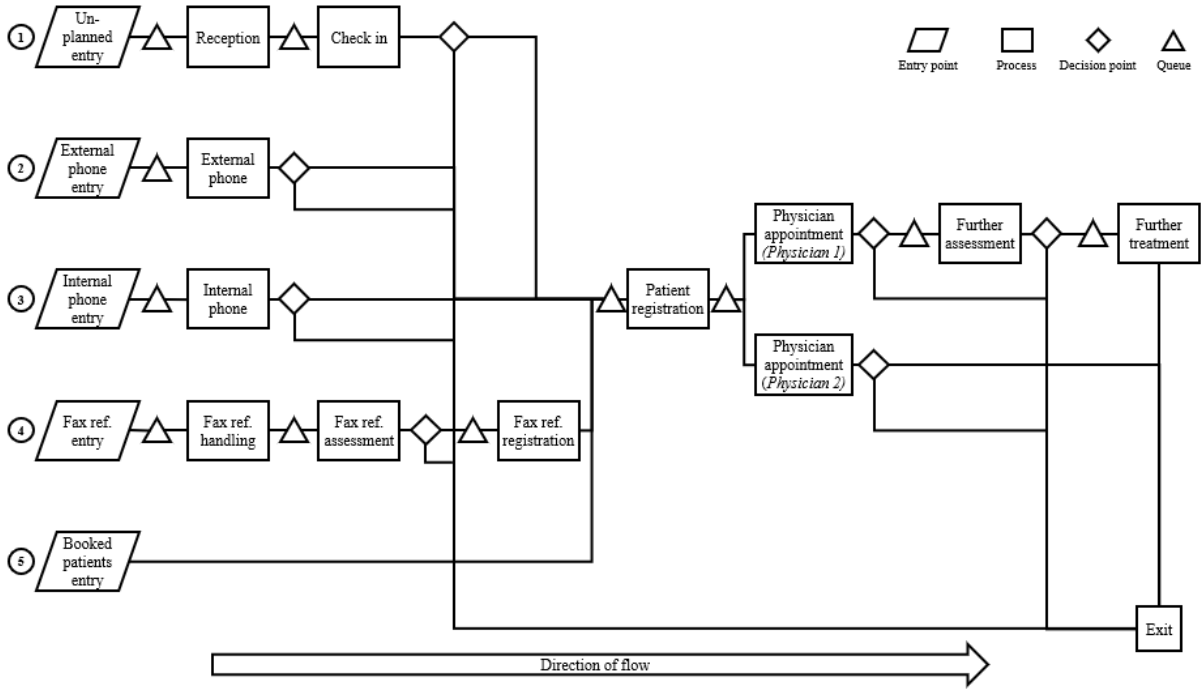


Figure 4.1: Conceptual map

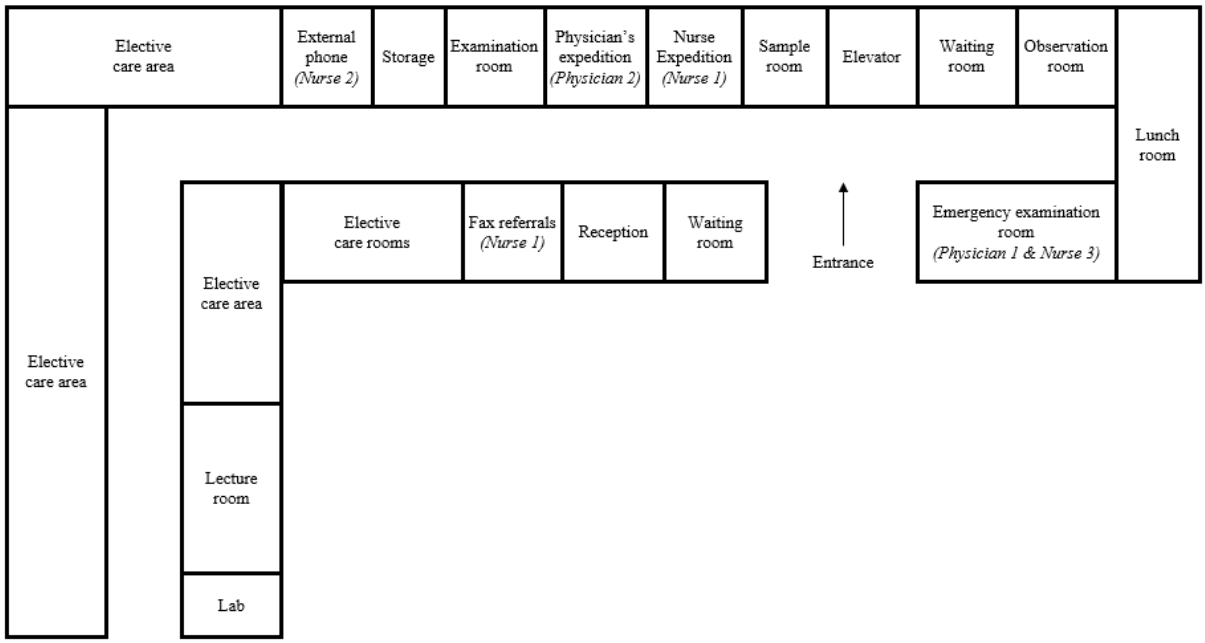


Figure 4.2: Physical layout of the ENT clinic

types both proceed through the entire system but are prioritised differently. The reason for this is that some patients entering the system do not have time to wait, otherwise they might suffer serious consequences or, in the worst case, die.

Entry points and routes

From a schematic point of view, patients move from left to right in the system, indicated by the directional arrow in Figure 4.1. The first entry point is the one most evident when visiting the clinic, it is when patients just show up to the emergency flow and request treatment. The second and third entry points are both entries for patients that call the clinic or get referred by a call from other medical staff. The split between internal and external is not entirely exclusive, nevertheless it intends to divide cases from the same hospital and external cases from outside of the hospital. The external cases could be individuals, health centres, etc. calling.

The fourth entry point describes all cases that enter the system by a referral sent by fax. Due to physicians tending to send referrals through several channels, which is not considered best practice, this implies that patients are sometimes referred multiple times. As a consequence of this, the same patient could theoretically enter the system during the same day in several ways. An example of this is when an external health care provider sends a fax and calls the clinic's external phone as well.

The fifth entry point describes all patients that are scheduled from previous days. Patients entering through entry point 5 are patients that could not be treated the previous day or patients from the elective care flow that cannot wait until their scheduled time. These patients are mostly of type 2.

Every entry point is described in table 4.1. The underlying data for the table is both secondary data as well as empirical data. The ENT clinic measures the number of patients entering the system everyday, but does not make a distinction between different entry points, hence the need for empirical data to achieve a complete view.

On a very high level, patients enter the system in a certain way with the purpose of seeing a physician. From the patients' perspective this is true for every patient, but from the medical staff's perspective this is only valid for patient type 2 and 3. This is because a patient of type 1 is not in need of emergency care, but still seeks it. In the physical system, the entry of patients takes various shapes depending on the way patients enter and their condition. They encounter different processes and wait in different queues depending on type and entry. After visiting a physician, patients are sent home in a majority of the cases. However, some patients need to be kept in the system for testing or further treatment. The reason why most patients do not stay in the system is because it is an emergency clinic and once patients are no longer in an emergency state, they can be sent home or transferred to another division.

Processes and queues

Every process describes a step in which a patient interacts with staff from the emergency flow. The various processes include interaction between the patient and a receptionist, between the patient and a nurse and between the patient and a physicians. The interactions can be different in nature, such as calling and being treated at distance

Detailed information about entry points		
Name	**Distribution of patient types**	**Inter-arrival times (Exponential distribution)**
Unplanned	Label 1: 34% Label 2: 46% Label 3: 20%	08:30-09:45: 67.8 min 09:45-11:45: 39.2 min 11:45-13:45: 52.4 min 13:45-16:00: 41.5 min
External phone entry	Label 1: 90% Label 2: 8% Label 3: 2%	Exception* 07:45: 50 patients arrive
Internal phone entry	Label 1: 75% Label 2: 20% Label 3: 5%	08:30-09:45: 96.9 min 09:45-11:45: 56 min 11:45-13:45: 74.9 min 13:34-16:00: 59.3 min
Fax referral entry	Label 1: 75% Label 2: 25% Label 3: 0%	08:30-09:45: 96.9 min 09:45-11:45: 56 min 11:45-13:45: 74.9 min 13:45-16:00: 59.3 min
Booked patients	Label 1: 0% Label 2: 100% Label 3: 0%	08:30-09:45: 135.6 min 09:45-11:45: 78.4 min 11:45-13:45: 104.8 min 13:45-16:00: 83 min

Table 4.1: Table of entry points

or being examined in a physical room. The queues, on the other hand, are rather homogeneous as they differ only in the amount of time spent in them and the order patients enter and exit. Every queue implies that the patients are waiting before entering the next process. In table 4.2 all processes that patients encounter are described. It is a detailed description including the probability distributions that best describe the process times, along with the parameters, which are combinations of α, β, average, standard deviation (σ) and offset (minimum value). However, the probability functions are not part of the original conceptual model but they are included when the conceptual model is translated to a simulation model in SIMUL8.

All processes are described from a patient flow perspective and the average of the probability distributions describe the expected time it takes for the staff to process a patient in a given process. The different processes in table 4.2 are not necessary in different physical locations but are divided into different processes because they imply different steps of treating patients. The physical location of the different processes is not essential to understanding the system and will not be discussed in detail, but figure 4.2 is provided for reference. For processes located far away from each other, travel times have been taken into consideration during the translation of the conceptual model into SIMUL8.

Description of processes			
Name	**Description**	**Probability distribution**	**Parameters**
Reception	Physically present patients get registered	Weibull	$\alpha = 1.2$ $\beta = 0.73$ Offset = 0.92 min
Check in	Physically present patients are initially assessed and triaged	Gamma	$\alpha = 0.58$ $\beta = 4.42$ Offset = 3.27 min
External phone	External patients are assessed over the phone	Weibull	$\alpha = 1.29$ $\beta = 2.83$ Offset = 2.38 min
Internal phone	Internal patients in need of transfer are assessed over the phone	Weibull	$\alpha = 0.74$ $\beta = 1.23$ Offset = 0.38 min
Fax referral handling	Fax referrals are handled and passed on to the assessment	Weibull	$\alpha = 1.2$ $\beta = 2.79$ Offset = 1.78 min
Fax referral assessment	The referral is assessed and a decision is made on whether the patient needs emergency care	Log-normal	Average = 2 min $\sigma = 1.56$ min Offset = 1 min
Fax referral registration	Patients in need of emergency care are scheduled	Log-normal	Average = 5.58 min $\sigma = 3.58$ min Offset = 1 min
Patient registration	Patients are registered and pay for their visit	Weibull	$\alpha = 1.51$ $\beta = 2.17$ Offset = 0.85 min
Physician appointment (Physician 1)	Patients are examined and treated by a physician	Weibull	$\alpha = 1.08$ $\beta = 6.03$ Offset = 9.67 min
Physician appointment (Physician 2)	Patients are examined and treated by a physician	Weibull	$\alpha = 0.84$ $\beta = 5.01$ Offset = 11 min
Further assessment	Patients in need of blood samples, x-ray, etc.	Log-normal	Average = 17 min σ=12.74 min Offset = 3 min
Further treatment	Patients are further treated by a physician	Log-normal	Average = 7 min $\sigma = 4.93$ min Offset = 3 min

Table 4.2: Table of processes

Resources

The conceptual map described in figure 4.1 is not displaying any resources. Resources in a simulation context, as well as in this case, often refer to personnel working within

the physical system. Several processes are often dependent on resources being available in order to process, in this case, patients. In the actual emergency flow, no process can function without personnel, but in the simulation model context resources are of interest if they are, or can be, shared by several processes. Furthermore, the resources' schedules can be adjusted as an easy way to vary the capacity. In this case, nurses and physicians can theoretically work all across the model. However, no receptionist was modelled since it is not a shared resource and, if relocated, is not able to aide with treating patients in the vocation care flow. Furthermore, all resources are set to have a maximum capacity of 90% to take personal time into consideration, such as going to the toilet or taking short breaks.

In table 4.3 all resources and the processes that depend om them are listed. The table can be interpreted like; if *Nurse 1* is occupied at the expedition, assessing patients, both fax referral processes need to wait until *Nurse 1* is available. The resource's utilisation depends on how many responsibilities a resource has, the time each responsibility takes and the amount of patients flowing through it.

Two types of physician resources are present in the emergency flow. *Physician 1* is the main resource that works with diagnosing and treatment and has *Nurse 3* working alongside them in the process *Physician appointment (Physician 1)*. *Physician 2* also works with diagnosing and treatment, but is a resource that is shared between all processes at the Hospital in need of specialised ENT competence. Thus, the resource spends the working days between taking referrals, consulting other caregivers by phone, and treating patients both in the emergency flow and all over the Hospital. In practice, a very low share of *Physician 2's* time is allocated to the studied emergency flow.

Resources and their responsibilities	
Resource	**Processes handled**
Nurse 1	Expedition, Fax referral handling and Fax referral registration
Nurse 2	External phone
Nurse 3	Physician appointment (Physician 1), Further assessment and Further treatment
Physician 1	Physician appointment (Physician 1), Further assessment and Further treatment
Physician 2	Fax referral assessment and Physician appointment (Physician 2)

Table 4.3: Table of resources

4.1.3 Simulated results

After determining the input data distributions and translating the model into a computer-based model, the initial state of the simulation model was reached. This model is a description of current state of the emergency flow, i.e. the physical system. The different probability distributions that are the basis of the process times in the simulation model are described in table 4.2. Distributions that concern the entry points are described in table 4.1. Together, these underlying probability distributions are the cornerstone of

the simulation model. They are what transform the model into a representation of the physical system.

As mentioned in the theoretical framework chapter, variation is inherited throughout a system and it can result in increasing queuing time. Queuing time is in direct relation with throughput rate as well as throughput time (TT). Therefore, the latter was chosen as the main key performance indicator (KPI). Since the physical system is not homogeneous, in terms of patients being assessed over the phone or via fax referral, one single throughput rate was not sufficient. Therefore, seven different KPIs regarding TT were measured.

In addition to the TT KPIs, the utilisation of four different resources were measured. Only four utilisation KPIs were measured although there are five resources in total. This was due to the *Physician 1* and *Nurse 3* working together in all their processes. The KPIs displayed in table 4.4 are extracted from the simulation base model, later referred to as Trial 1. The average out of 28 runs with a 95 percentage confidence interval are displayed. One run means running the entire simulation model for one week, Monday to Friday. The table is structured with the KPIs in the left column along with their respective units in brackets (minutes or %), with the average value of the KPI in the third column, and the lower 95% and upper 95% confidence interval limits in the second and fourth column. As an example, the average throughput time for KPI (1) is 172 minutes, with 95% of the times falling in between 170 and 175 minutes.

KPIs base system			
KPI	**-95%**	**Average**	**95%**
(1) TT Entire system [min]	170	172	175
(2) TT External phone to patient registration [min]	213	216	219
(3) TT Internal phone to patient registration [min]	9	10	11
(4) TT Fax referral to patient registration [min]	181	203	225
(5) TT from Patient registration (patient type 2) [min]	73	79	86
(6) TT from Patient registration (patient type 3) [min]	46	48	51
(7) TT Unplanned [min]	59	64	69
(8) Utilisation Nurse 1 [%]	36	37	38
(9) Utilisation Nurse 2 [%]	79	80	81
(10) Utilisation Physician 1/ Nurse 3 [%]	68	70	73
(11) Utilisation Physician 2 [%]	86	87	89

Table 4.4: Table of results for the base system

A large part of the TT consists of waiting in queues. Especially before physically entering the system patients have to wait relatively long before being assessed. For example when patients call over the external phone and happen to be put last in queue, in this scenario the waiting time before being called by *Nurse 2* could be several hours. In general, once patients physically enter the system the TT decreases, relatively to before physically entering the system, yet, it occurs that patients have to wait several hours before an appointment with a physician. One scenario where this is likely to happen is if there are a few patients of type 2 and many patients of type 3, then the patients of type 2 have to wait due to their lower priority.

Due to limitations, mentioned above in more detail in the method and discussion chapters, statistical validity was never extensively assessed or reached. Without the option of going back to the physical system and measure throughput times, it was not possible to conduct statistical tests for validation. However, by presenting the initial state of the model to employees at the ENT clinic and other stakeholders, validity was reached. Additional sensitivity analysis validated the model to an extent that the practitioners found satisfying, given the circumstances.

4.2 Factorial analysis

To better understand how resources impact the defined KPIs, a factorial analysis was used. This section describes the way in which the factorial analysis was designed, and the most important results that were extracted.

4.2.1 Factorial design

The factorial analysis was constructed and carried out in line with its description in the method chapter. Due to five resources, with the constraint of *Nurse 3* and *Physician 1* on call depending on each other, the $2k - p$ factorial design had 16 combinations. p, the degree of fractionation, was limited to two, meaning that the resources could only be increased with one unit from their base level. The base system is represented in Trial 1 in table 4.5. It has one of each resource, regarding *Nurse 3* and *Physician 1*, there is one of each.

By comparing the different trials with each other for each KPI, it was possible to get a better understanding of how different resources impact the system. Each KPI was analysed with 16 trials, where an *ANOVA* test was used to determine if the trials differed significantly. As all KPIs showed significant differences, Levene tests were conducted in SPSS to check for unequal variance between the trials. Finally, a post hoc test was conducted for each KPI. A Tukey's test was used if there was homogeneity of variances among the samples, and a Tamhane test if not. An example matrix is shown in Table 4.6, analysing the KPI *TT Entire system*.

Table 4.6 is an example of how the post hoc tests were interpreted. On the top row, the analysed KPI is noted, in this case TT Entire system, KPI (1). In the leftmost column, the trial that is being assessed is marked, in this case Trial 1. The next column states which trial is being compared to Trial 1. Column 4 contains the mean difference between the trials and column 5 the standard error associated with the mean value. Column 6 contains the significance level, stating whether there is any significant difference between Trial 1 and the compared trial. Finally, column 7 and 8 contains the upper and lower 95% confidence interval bounds. In a similar fashion, 16 matrices were constructed for each KPI.

For example, Trial 2 indicates a mean difference of 0.88 minutes and a standard error

Factorial design matrix: += High, -= Low				
Resource	**Nurse 1**	**Nurse 2**	**Nurse 3 / Physician 1**	**Physician 2**
Trial 1	-	-	-	-
Trial 2	+	-	-	-
Trial 3	+	+	-	-
Trial 4	+	-	+	-
Trial 5	+	-	-	+
Trial 6	+	+	-	+
Trial 7	+	+	+	-
Trial 8	+	+	+	+
Trial 9	-	+	-	-
Trial 10	-	+	+	-
Trial 11	-	+	-	+
Trial 12	-	+	+	+
Trial 13	-	-	+	-
Trial 14	-	-	+	+
Trial 15	-	-	-	+
Trial 16	+	-	+	+

Table 4.5: Table of factorial design matrix - factorial analysis

of 2.03, but this difference is not significantly different from Trial 1 as the significance level is above the defined alpha value of 0.05. This can be confirmed by assessing the confidence interval bounds, which contain the value 0. Hence, it is impossible to say that adding an additional *Nurse 1* to the system has any effect on KPI (1), as Trial 2 was defined as having an extra *Nurse 1 resource*.

Trial 3 indicates a larger mean difference of 65.9 minutes with a standard error of 2.17 and a significance value close to 0. Furthermore, the confidence interval does not include 0, confirming the hypothesis that the systems give different results. Hence, this gives reason to believe that adding a *Nurse 2* would reduce KPI (1) by 65.9 minutes.

In the same fashion as table 4.6 compares Trial 1, the initial state, with all other states, the 16 matrices for each KPI indicated the impact of the different trials. Choosing another trial as example, Trial 8, shows that the difference in KPI (1) is 87.8 minutes, which is greater than for Trial 3. To assess whether there is a significant difference between Trial 8 and Trial 3, a similar table was used. Here, it was shown that the systems are significantly different, and thus there is reason to believe that Trial 8 has a lower throughput time than Trial 3. This might not come as a surprise, as Trial 8 had all resource levels set to high (+). Following the same logic, each KPI was assessed.

The four utilisation KPIs were analysed qualitatively in parallel with the factorial analysis of the different throughput times, and were used in the assessment of the desired future state described in section 4.3. This was done by comparing the utilisation figures and noting if there were any large deviations from the base system.

Tamhane post hoc test - TT Entire system [min]						
Trial	**Trial**	**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**Lower 95% bound**	**Upper 95% bound**
1	2	0.88	2.03	1.0	-6.73	8.49
	3	65.9	2.17	0.0	57.75	74.10
	4	17.01	1.91	0.0	9.83	24.19
	5	14.52	1.75	0.0	7.91	21.12
	6	79.64	1.61	0.0	73.50	85.79
	7	83.59	1.58	0.0	77.56	89.61
	8	87.77	1.47	0.0	82.04	93.50
	9	64.26	1.98	0.0	56.83	71.69
	10	81.90	1.59	0.0	75.84	87.97
	11	78.48	1.67	0.0	72.16	84.80
	12	86.46	1.53	0.0	80.57	92.35
	13	15.40	1.84	0.0	8.49	22.31
	14	20.19	1.79	0.0	13.46	26.93
	15	12.92	1.79	0.0	6.21	19.64
	16	20.88	1.79	0.0	14.15	27.61

Table 4.6: Tahmane post hoc test - TT Entire system

4.2.2 Resources' effects on the KPIs

Physician 1 and *Nurse 3* have the greatest overall impact in terms of decreasing throughput times after the patient has entered the physical system, either through the unplanned entry or after the patient has been registered before a physician visit. Both resources are often highly utilised with utilisation figures of 70%, which indicated that this resource combination might be a bottle neck.

The analysis of the throughput time of the external phone is very straightforward. Since *Nurse 2* is the only resource working in this part of the patient flow, paired with the patients not yet having physically entered the system, all trials containing an additional *Nurse 2* significantly decrease this throughput time KPI. It is also not possible to differentiate between the different trials containing an additional *Nurse 2*. In terms of the entire system's throughput time, KPI (1), this resource has the largest impact as the time spent in the phone queue is greatly reduced. Furthermore, the utilisation of *Nurse 2* reaches 80% in the initial system, indicating that this too might be a bottle neck in the system.

The throughput time of the internal phone is affected by changing the resource *Nurse 1*, as this was confirmed by the ANOVA test. Since *Nurse 1* is the only resource staffed on the internal phone, this is not very surprising. However, the post hoc test test showed that the difference in mean values of throughput times were not statistically significant. It is also interesting to note that the decrease in throughput time when adding an additional *Nurse 1* is fractional. This might be connected to low utilisation of *Nurse 1*, which only

reaches 37% in the initial system.

The throughput time of the fax referral highly depends on the availability of *Physician 2* on call as well as *Nurse 1*. The throughput time is significantly decreased by adding both resources. However, if only a *Nurse 1* is added it does not impact the throughput time significantly, hence *Physician 2* is the limited resource. This well aligned with *Physician 2's* high utilisation. An interesting observation is that adding a *Physician 1* and a *Nurse 3* further decreases the throughput time to the fax referral. This is due to them freeing up time for *Physician 2* who can focus on fax referrals instead of treating patients.

Regarding TT from patient registration (patient type 2), *Physician 1* and *Nurse 3* have the biggest impact in terms of decreasing throughput time. An additional *Physician 2* also shortens the throughput time but not to the same degree as an extra *Physician 1* and *Nurse 3*. For example, an extra *Physician 1* and *Nurse 3* reduces the throughput time by 39 minutes, while an extra *Physician 2* reduces the throughput time by 16 minutes.

However, adding a *Physician 2* at the same time as adding a *Physician 1* and *Nurse 3*, as when comparing Trial 7 and Trial 8, there is no significant difference in throughput time, which indicates the relatively low impact of an additional *Physician 2*. Neither *Nurse 1* or *Nurse 2* seem to impact the throughput time significantly.

TT from patient registration (patient type 3) is affected similar to the throughput time patients registered (patient type 2). An additional *Physician 1* and an addition *Nurse 3* decrease the throughput time significantly and neither *Nurse 1* or *Nurse 2* seem to decrease the throughput time. However, *Physician 2* does not impact the throughput time significantly. This could be due to patients of type 3 being prioritised. Since the *Physician 2* is not exclusively working with assessing patients at the ENT clinic and therefore treats far less emergency patients compared to the *Physician 1* and *Nurse 3*, prioritised patients will often find their way to the more available resource, i.e. *Physician 1* and *Nurse 3*.

TT Unplanned is slightly different from the other throughput times. It concerns the throughput time for patients that enter through the unplanned entry and measures throughput time throughout the entire system. Neither *Nurse 1* or *Nurse 2* significantly impact the throughput time. The reason for this might be the low utilisation of *Nurse 1* and that nurse two exclusively works with patients entering through the external phone.

However, when adding an additional *Nurse 2*, there is a tendency of increasing the throughput time, although the result is not statistically significant. This could be due to an decreased throughput time in other patient flows that later flow into the same shared patient flow. This could imply that unplanned patients experience a larger queue when waiting for a physician appointment.

All systems with an additional *Physician 1* and *Nurse 3* reduces TT Unplanned with statistical significance, and *Physician 2* does not. This can be seen when comparing Trial 1 with Trial 12, 13 and 14, where adding a *Physician 2* does not impact the throughput time. The reason for this might again be the low availability of *Physician 2*, as discussed above. In conclusion, adding one each of *Physician 1* and *Nurse 3*, the throughput time

is decreased the most, out of which the majority of the throughput time is made up of waiting time for the patient, such as waiting in a queue to the external phone or fax assessment.

4.3 Desired future state

In line with the stated research question, the desired future state is a state in which less resources manage to treat the same number of patients without increasing the throughput time radically. Since *Physician 1* and *Nurse 3* have a high utilisation rate and almost always decrease the throughput time at the higher factorial level, in the experiments above, when increased by one, these are resources that cannot be reduced or removed. Continuing on this trajectory, *Nurse 1* and *Nurse 2* are the remaining choices to reduce. Since *Nurse 1* showed a low utilisation rate and *Nurse 2* being highly utilised in the external phone, allowing *Nurse 1* to assist with the external phone was considered to be a viable solution. When this system was simulated, it was shown that *Nurse 2's* schedule could be reduced. This option was also considered to not require a complex adaptation in the physical system, as all nurses in the system are trained to work at each station, and currently staffed on a rotating schedule.

As displayed in table 4.7, the throughput time of the entire system decreases if one allows *Nurse 1* to help *Nurse 2* with the external phone, and reducing *Nurse 2's* scheduled hours by half. The obtained result can be explained by the low utilisation of *Nurse 1* and the high impact of patients entering the system by the external phone on the throughput time. Removing *Nurse 2* completely from the model was tested as well and did not show large increases in the throughput time. However, this solution showed a simulated utilisation of *Nurse 1* of almost 90%, which is a considerably high value. The model does not account for some of the personal time, such as drinking coffee or social media usage, however, it is only done by restricting the utilisation of resources to 90%, which might be somewhat optimistic. Furthermore, some process time might be prolonged by implementing changes, this is unlikely in the long run but could become the case initially. Hence, this thesis proposes that the scheduled time for *Nurse 2* in the emergency flow is reduced by half. A graph of the 11 KPIs is provided in figure 4.3, comparing the current (initial) state with the desired future state.

The desired future state enables transferring half of resource *Nurse 2* to the elective care. By doing so, it is possible to assist the elective care and also in the medium term being able to decrease the number of patients that enter through the booked patients entry point. Because, as mentioned above, some of the patients entering through this entry are patients that the elective care does not manage to treat in time. Furthermore, *Nurse 2* will not have any issue of changing to the elective care since the personnel at the ENT clinic rotate between the emergency flow and the elective care unit.

TT Entire system [min]			
State of the model	**-95%**	**Average**	**95%**
Initial state model	169.5	172.4	175.2
Nurse 2 working half-day	115.1	118.9	122.6
Nurse 2 not working	186.3	191.6	196.8

Table 4.7: Table of desired output

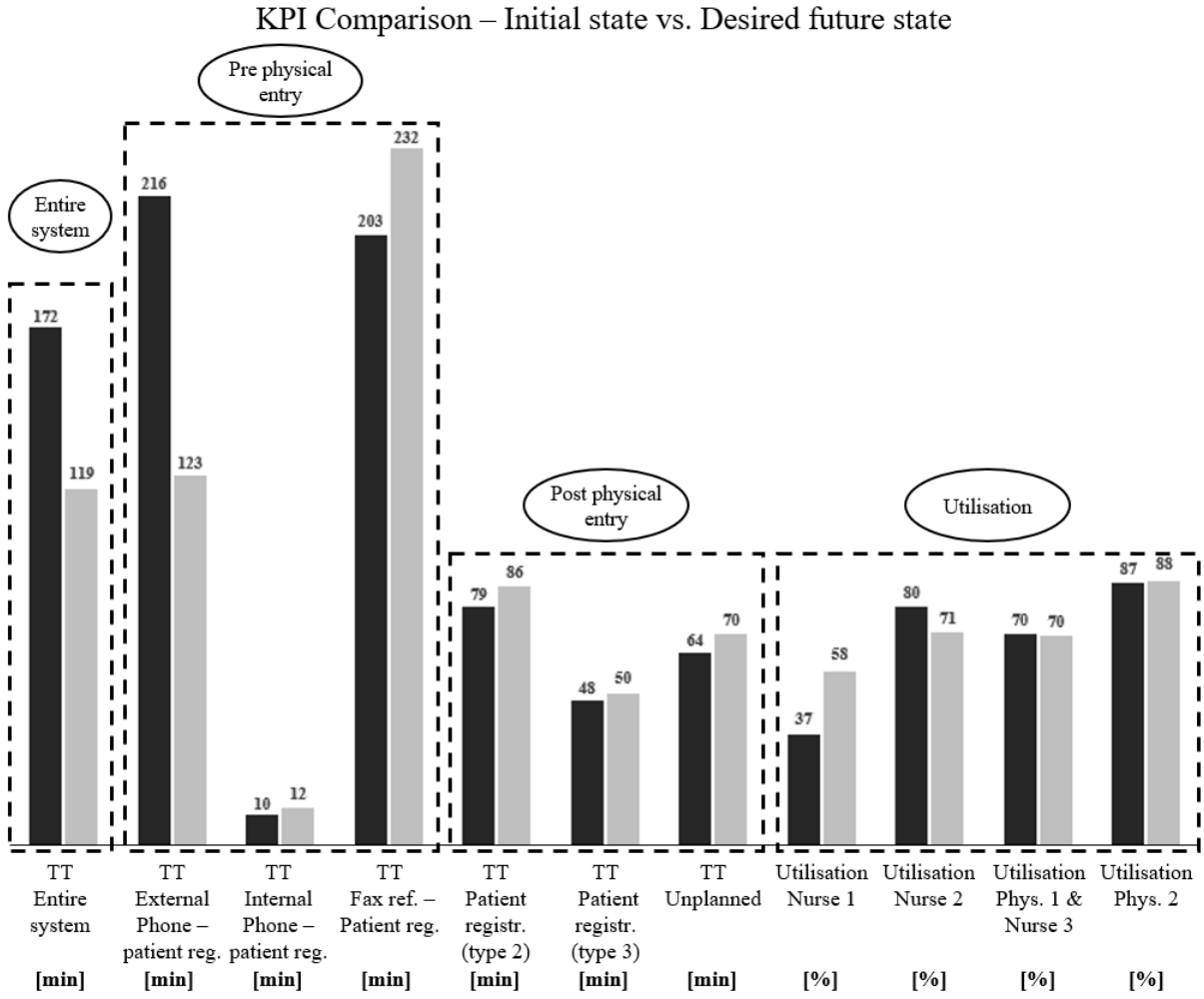


Figure 4.3: KPI comparison - Initial vs. Desired future state

4.3.1 Sensitivity analysis

In order to hedge some of the initial limitations and their impact on the KPIs of the desired future state, a sensibility analysis was conducted. Similar to the experimental design, a $2k - p$ factorial design, with the number of levels in each factor reduced to two, was conducted. This time the factorial design was limited to three combinations. In this factorial design, the initial state, Trial 1, was the desired future state with *Nurse 1* assisting with the external phone and *Nurse 2* scheduled for half days, and the variance of a certain process. The variance was changed by multiplying the initial variance by a factor of two, and is indicated by a (+) sign in table 4.8.

Since all factors were variances of process distributions, the initial state, Trial 1 was compared to the worst case, Trial 2. Of all processes variances that constitute a factor in the sensitivity analysis, *Nurse 1* is responsible for Fax referral registration and *Nurse 2* for none.

Factorial design matrix: += High, -= Low				
Resource	**Fax referral assessment**	**Fax referral registration**	**Further assessment**	**Further treatment**
Trial 1	-	-	-	-
Trial 2	+	+	+	+

Table 4.8: Table of factorial design matrix - Sensitivity analysis

None of the trials yielded any significant differences in throughput time, when measuring the throughput time of the entire system, KPI (1). This could be due to the processes constituting a factor not being highly utilised, hence the increased variance does not necessarily increase existing queues. Another reason might be that the majority of patients never enter these processes and therefore have limited impact on the overall throughput time.

Nevertheless, the impact of the uncertainty of variance in the adjusted processes is determined to be limited. Therefore, on the basis of this sensitivity analysis, the desired future state is judged to be reliable.

5

Discussion

5.1 Method discussion

As mentioned in the theoretical framework, Chung (2004) and Jingshan et al. (2017) describe simulation modelling as a good way to comprehensively analyse complex system without disturbing them. During the project, the choice of simulation modelling, in line with Chung (2004), has helped both the personnel working at the clinic and the authors of this thesis in better understanding the physical system. Among others, mapping the patient flow and quantitatively estimating throughput time and utilisation were activities that increased the understanding. Up until this thesis, there were no detailed descriptions of how individual processes are carried out, and most process times were estimates as no detailed data had been collected by the Hospital. Being involved in the process of mapping out the system as well as being presented with the final model and the desired future state, increases employee understanding and awareness of the operations of the emergency flow. This involvement of the ENT clinic personnel could even lead to a future increase in improvements to the physical system initialised from within the ENT clinic.

In the specific case of healthcare, as Boaden et al. (2008) describe it, the implementation of change is strenuous. Therefore, according to the authors of this thesis, simulation modelling could be an adequate tool. It is a comprehensive tool that allows for easier analysis of complex and dynamic systems, paired with ease of visualisation (Chung, 2004). This makes it easier to present, understand and trust the results. Since it could be perceived as a more scientific approach over other more common used practices in a similar context, a result achieved through simulation modelling could also experience a higher probability of implementation. Simulation modelling also allows practitioners to, once the model is built, run it in compressed time. This means that the model can be examined for a large number of runs leading to both a better understanding of the model and better statistical analysis.

Simulation within healthcare has also shown to come with its difficulties. As stated earlier, access to data could be difficult as the collected data, at this Hospital, mainly covers when patients are admitted along with respective diagnosis codes. No statistics are collected on a process level, which required the simulation practitioner to collect most data themselves. Many physicians and nurses tend to approach tasks in their own ways,

possibly due to a lack of detailed process descriptions. Additionally, the work requires specific professional skills, making it difficult to fit data to highly accurate distribution functions. Furthermore, high variation increases the data quantity requirement.

A clear drawback to simulation modelling as Chung (2004) states is that it cannot solve any issues on its own, provides no easy answers to complex problems and is very dependent on the quality and amount of input data. In addition to this it is resource intensive in terms of knowledge and time. These drawbacks to simulation modelling are important to take into consideration during the process of choosing a method. In this particular case, it was determined that simulation modelling would be a suitable fit, since the physical system was very complex and one where changes could be difficult to promote and implement. However, due to COVID-19, the data collection phase was interrupted and input data limited, further discussed in section 5.5. This could not have been anticipated and therefore simulation modelling still is considered to be a suitable tool in this context.

Due to the limited input data, the data analysis method had to be altered for some of the processes described in the conceptual model in figure 4.1. Due to a lower volume of underlying data points than desired, these distributions may be less valid. However, they where validated by qualitative methods and are believed to be good-enough approximations. Therefore, it is not believed that the probability distributions nor the model as a whole make up a non-desirable representation of the physical system, as the practitioners experienced it. Yet, the limited observations made it difficult to account for the high variance between different employees at the clinic and the seasonality of the system. Due to a rotational staffing system the suggested future state could easily be implemented as all employees have knowledge of all stations, but this also means that there are, for example, several nurses that occupy the post of *Nurse 1*. This imposes a major variance on both average times and the appurtenant variance. The practitioners hedged some of these effects in the sensitivity analysis, but it is suspected that not all effects could be caught.

The desired future state was derived by an experimental design. This method imposes a quantitative approach to a complex task. By understanding the impact of every resource on the entire system as well as subsystems, it becomes easier to design a new system. Both Chung (2004) and Law (2014) describe the factorial analysis as a way of determining the design of new systems. Due to the research question of this thesis and the difficulty of change within healthcare as described above, only resources composed factors. With 11 KPIs and 16 factors the factorial analysis became rather complex, however, it could have been conducted in a even more extensive matter by increasing the number of factors. There is no clear disadvantage to increasing the number of factors or their levels in terms of better understanding the model. Yet, an increased analysis makes it easier to mix up results and is time demanding. Therefore, this experimental design was limited to 120 comparisons of mean values for every KPI. This comparison of mean values is displayed in **??** for the entire throughput time, i.e. KPI (1).

In summary, this thesis reiterates that simulation can be used as a powerful tool for understanding complex systems with large degrees of variation. As could be expected, higher quality of collected data could lead to a more precise the model. However, even though a model can become very precise the more work is put into it, how detailed does

the model have to be, in order to draw relevant conclusions? The authors of this thesis argue that if one wants to draw precise solutions about outcomes from system changes, detailed data is needed. But as this thesis has shown, relevant results can still be found, even though the data collection phase is limited.

5.2 Discussion of system description and operations characteristics

As mentioned in section 4.1.1, the emergency flow is characterised by a high volume, variety, variation, and visibility. According to Slack and Lewis (2017), this would indicate that three out of four V's would be associated with higher costs, while the higher volume helps drive down cost, since the physicians and nurses can become more specialised. However, Swedish healthcare is typically costly (Björnberg & Phang, 2019). This leads to a position where the high variety, variation and visibility could have a higher weight from a cost perspective.

In relation to Slack et al. (2013), quality and speed seemed to be the most important performance objectives for the clinic, since it is important to deliver the right care in order to minimise second visits and utilise the capacity on new patients, while the patients need to be treated quickly in emergency situations. The system also needs to be flexible since although the clinic provides specialised care within ENT, the emergency flow is more generalised than the elective care. Dependability does not seem to be of utmost importance, as patients are booked in preliminary time slots and receive care as the queue before the physician is worked through.

The above factors had to be taken into consideration when designing the new model. As the emergency flow is predominately staffed with a single resource where a resource is required, it was not as simple as removing a resource from a process to free up capacity. Instead, every resource's individual schedule and utilisation had to be taken into account, resulting in shortening the day of *Nurse 2* by 50% while allowing for *Nurse 1* to assist in handling the external phone, and suggesting to redirect *Nurse 2* to the elective care unit for the rest of the day.

Relating to Slack and Lewis (2017), the above illustrates a trade-off between quality, speed and cost. Removing a whole resource could free up capacity from the emergency flow and thus reduce direct cost, but could lead to a slower throughput time and indirectly impact the quality of the provided care. Furthermore, taking into account the high variation and high potential capacity utilisation for *Nurse 1*, this could lead to longer queue times in reality, as increased variation and utilisation have an exponential impact on queue length according to the Kingman Formula (Holweg et al., 2018).

As Modig and Åhlström (2012) wrote, the efficiency paradox means that an organisation typically strives for higher resource utilisation, while customers tend to want higher variety and short queues. Therefore, a case is made for maintaining some slack resources in the emergency flow, since reallocating *Nurse 2* completely flow leads to a utilisation

rate for *Nurse 1* that was deemed too high.

5.3 Discussion of simulation model and results

The simulation base model was constructed under unusual circumstances with COVID-19 occurring during the study. This only impacted the data collection, however, that in turn impacted the model's process times. Anyhow, given that the model reached high face validity, sensitivity analysis was performed on the future state model, and both the practitioners and supervisors have experience from healthcare improvement projects, the results are thought to be reliable. This does not necessary imply that the physical system, if suggested future state changes are applied, experiences the exact same throughput times as the desired future state model, but the result of this thesis provides a good-enough approximation to indicate that the suggested changes can have positive implications for the clinic.

The final result, reducing *Nurse 2's* availability by half, provides the ENT clinic emergency flow with an increased utilisation of resources and a decreased throughput time. Since the throughput time is decreased in the model and *Nurse 1* is able to perform a portion of *Nurse 2's* tasks, it seems like a feasible solution from a operations standpoint. This way of task sharing is also in line with what Sasser (1976) stated about utilising cross-trained employees for managing capacity constraints. However, to avoid wasteful travel times for *Nurse 1*, a workstation could be set up in the expedition room. This would enable *Nurse 1* to efficiently switch between physically assessing patients and doing so over the phone. This would require a small investment, but this is not expected to be a large cost in the scope of total cost for providing care. A prerequisite to this solution is that the patient flow entering through the external phone is managed the same way in the future. This means that patients call in early and all slots are filled, hence, *Nurse 2* can be fully utilised. If there was no initial queue the utilisation would decrease and the future state would not be as viable, as a basic assumption in the model is that the queue does not start filling up again as the day goes on.

As presented in the results, *Nurse 2* could be reallocated entirely but this would imply a very high utilisation of *Nurse 1*, and would probably not be viable in reality. Since the model is limited in its underlying input data, this future state was somewhat to close to a critical value of utilisation, in addition, the model does not mitigate the effects of non-value creating activities, therefore, this future state was rejected. The only mitigation of non-value creating activities is performed by decreasing the overall ability of each resource to 90%.

The desired future state, if implemented, decreases throughput time for the entire system as well as allows for *Nurse 2* to partly be reallocated to the elective care of the ENT clinic. This is not a drastic change but offers several possibilities for the ENT clinic. Since the emergency flow's throughput time is decreased, it is, in theory, possible to increase the number of treated patients per day. It is also possible to decrease the number of patients that enter through the booked patients entry, since some of these patients are

patients redirected from the elective care. These patients are usually not able to fit into the strained schedule of the elective care and cannot wait for a later assigned time slot. However, one concern is raised regarding the need for an additional nurses in the elective care. Regarding empirical data from interviews, it seems like there is a need for additional nurses, yet, the practitioners have not conducted any additional research on this subject. This limits the thesis' ability to predict what effects this would have on the elective care system.

Overall, the proposed future state helps the entire ENT clinic with increasing its accessibility and decreasing patients' waiting times. Although this is a small step towards a bigger goal, reaching a potential sub-goal of gaining momentum and experiencing the positive effects of providing better care, could impact the personnel at the ENT clinic positively in terms of work satisfaction and psychosocial work environment Göransson et al. (2018).

5.4 Implementation considerations

Making changes to a healthcare system is not entirely straight forward. As Boaden et al. (2008) stated, healthcare workers have considerable discretion and autonomy, and usually need to be persuaded in change initiatives. Therefore, it is important to include the staff early on in both problem assessment, solution design and implementation to ensure that change efforts meet as little resistance as possible. Given that the healthcare system is also very complex and there generally exists a lack of whole-system views, paired with Druckler (2011)'s point about healthcare being financed by public tax money, implementation work becomes even more difficult.

Any implementation should, according to one of the interviewed physicians, be considered with quality and patient safety (Swedish: "patientsäkerhet") in mind, to not let speed and cost dominate too much. However, this is a matter of debate. This thesis' view is that maximising resource utilisation is not necessary preferable in Swedish healthcare, due to the long queues and waiting times. Therefore, the implementation of this thesis would imply increased resources within the elective care unit, without increasing labour costs. Although the proposed solution increases the utilisation in the emergency flow, it does decrease the throughput time. This is a rare but desirable combination.

After consulting with one of the main stakeholders at the ENT clinic, a potential hinder for implementation could be that unforeseen emergency circumstances might make *Nurse 1* unable to assist in the external phone as the resource is needed elsewhere. The external phone must also be reachable during the entire day, for example in the case of external caregivers calling for consultation. Furthermore, patients that have a scheduled time, as in the case of a time where they should be expecting a call from the clinic, might be less tolerant towards waiting or undependable time slots, which is in line with Slack and Lewis (2017)'s view on process visibility, waiting times and dependability. In the event of exceptional emergencies, these perspectives in combination might lead to difficulties when planning the schedule to balance availability and resource utilisation. Since the

thesis takes variation and today's demand into account, this is assessed to not affect the potential impact of the suggested future state significantly.

The healthcare stakeholder perspective presented by Glouberman and Mintzberg (2001) adds to the complexity of implementation. Balancing the needs of each stakeholder can make it difficult to implement system changes like the one proposed, possibly due to a lack of clear ownership between stakeholders as a large number of people tends to become involved and each stakeholder has its own interest and agenda. In brief terms, there might be risk that one stakeholder who is not satisfied with where the project is heading could hinder the progress, e.g. a physician that doesn't want to adhere to new standards, or politicians who don't share the view of the change project.

Even though a change like this might be difficult to implement, it is not impossible. Given that the proposed change is not of a drastic or radical nature, simply altering the resource schedule and allowing for more task sharing between *Nurse 1* and *Nurse 2* is a relatively small change and could be tried out on a small scale before assessing the suggestion's feasibility. Therefore, this thesis recommends that the proposed change is discussed and piloted before taking a final decision to implement.

5.5 Extraordinary circumstances impacting the study

As mentioned throughout the report, there were external factors impacting the work that was carried out in the thesis. The COVID-19 pandemic brought drastic restrictions to a majority of the world's countries, with Sweden being no exception. Access to the Hospital was restricted to anyone who didn't have a specific reason to visit, i.e. only healthcare workers and patients were allowed. This meant that the data collection phase at the ENT clinic had to be discontinued.

Luckily, the restrictions didn't come in force before the data collection started, meaning that data collection had already been conducted in the form of interviews and initial time studies of a majority of the processes. However, the extraordinary circumstances led to a general lack of data volume, impacting the ease of translating data into highly accurate distribution functions. After discussions with thesis supervisors, it was determined that data had to be completed with interviews and more face validation, as well as use software to assist in assigning distribution functions.

Although a higher volume of data could be desired, the study could be carried out without much difficulty after adjustments to the situation had been made. While statistical validity could not be reached, the model is an accurate description of reality in terms of logical flows and decision points. What is left for an accurate statistical description is to simply complement with more data points and compare the model output to the physical system's output. Furthermore, it is believed that for the aim of this study, relevant conclusions about the current and desired state could be drawn to answer the outlined research questions.

5.6 Generalisability

The desired future state can only be applied to the emergency flow of the ENT clinic or emergency flows designed in a similar way. Therefore, it is a very specific result. This however is the case with any new tailor-made layout for an operation. However, this does not mean that there are no general results that can be drawn from this study.

Being able to decrease throughput time as well as increasing utilisation of resources, by relocating parts of resources, implies a sub-optimised operation. As cited earlier, Boaden et al. (2008) describe healthcare as complex and often lacking a whole-system view. This could be an underlying reason to suspect that other clinics and inpatient wards within the Hospital, as well as within other hospitals, face similar sub-optimised operations. Therefore, the broader result of this study, the ability to create a more efficient patients flow by changing the operational layout, could be applied to other potential studies. This could by done by using a tool such as simulation modelling, but also be accomplished by leveraging other tools.

Describing and understanding the current state of any operation is a spillover effect of applying simulation modelling. As Chung (2004) mentions, thorough understanding and a conceptual model of a physical system are steps that should be included in any simulation modelling project. By taking these steps, practitioners will be able to describe the current state of any operation that is to be modelled. This spillover effect is determined to be rather general.

Jingshan et al. (2017) describes several cases where simulation has been used as a effective tool within healthcare. After conducting the study, it has become clear that simulation modelling is adequate tool to tackle optimisation tasks within a complex environment. Two advantages found in this thesis are the probability of impact and the diffusion during the process of building a model. This refers to the better understanding of the system of people working within it and the perception of simulation modelling as complex and scientific. It has been noticed that people within healthcare often are critical towards change, as described by Göransson et al. (2018), however, in this particular case they were rather receptive towards simulation modelling. This is not a final conclusion since the empirical data is too limited, but it opens up the possibility of an additional reason to continue using simulation modelling in solving complex problems within healthcare. Altogether, it appears like simulation modelling is a suitable tool for solving complex problems, both within and outside of healthcare. From this thesis' perspective, it is a particularly suitable method when relevant secondary data exists and primary data is easily accessible.

5.7 Suggestions for further research

Building upon this project by simulating the emergency flow of the ENT clinic with a broader and larger conducted data collection is a consideration for future research. The

existing model, or a new one, complemented with more data, could be used to conduct a different experimental design with other objectives, for example with decreasing the the throughput time as the main objective. Also, by using the findings from this thesis, practitioners can gain a base understanding of one part of the patient flow and connect it to a larger model, incorporating several units like the elective care flow, referral booking process, etc. Researchers could base their study in this thesis' model in both the above described future research suggestions. Additionally, the finding that the emergency flow at the ENT clinic was sub-optimised gives rise to the possibility of sub-optimised patient flows at other clinics and inpatient wards within the Hospital and the region. Assuming that this statement is valid, simulation modelling could be applied in the aforementioned cases and this thesis' model could act as a reference and guideline in these studies.

As simulation was a helpful tool for mapping, understanding and quantifying factors in a patient flow, the authors of this thesis argue that simulation can be used in more settings than just at the Hospital. An advantage of simulation is the ability to quantify factors in both the base system and effects on system changes, as well as account for variation in processes and arrival times. These advantages might not be as prominent in other methods such as standard flowcharts or Value Stream Mapping, to name a few. However, during the process of conducting this thesis, the authors have experienced several inefficiencies within the physical system that do not necessarily need to be quantified, and could serve as basis for future research within the Hospital without the use of simulation modelling.

6

Conclusion

The current state of the ENT clinic deals with a moderate volume and high variety, variation and visibility. There are several entry points through which the patients can enter, which are either walking in unplanned through the front door, through an external phone, referrals to the internal phone or fax, or being pre-booked for appointments. The external phone and the fax referral are currently the entry points where the throughput times are the longest.

The physical system consists of several processes and different resources are required for different processes. From studying the different processes, it could be concluded that *Nurse 2*, *Nurse 3*, *Physician 1*, and *Physician 2* are highly utilised, and that Nurse 3 and Physician 1 constitute a bottle neck in regards to physical throughput time, and *Nurse 2* constitutes a bottle neck as a large phone queue is built up each day at the external phone. However, neither resource is currently utilised to the degree where the throughput time is impacted to an extent that is regarded as non-feasible. With the parts of the research question aiming towards freeing up resources the conclusion was to emphasise on *Nurse 1* and *Nurse 2*.

The desired future state presented in this report suggests allowing for task sharing between *Nurse 1* and *Nurse 2*, where *Nurse 1* assists with the external phone. By doing so, the throughput time for patients waiting in the external phone queue can be shortened. Furthermore, by decreasing the hours of *Nurse 2* working within the emergency flow of the ENT clinic and reallocating the rest to the hours to the elective care of the ENT clinic, resources are freed and the throughput time of the entire system is decreased. This result is reached by letting *Nurse 1* aid with the assessment of patients entering through the external phone. Since the nurses rotate within the ENT clinic and a relatively small investment is required, in order to enable the task sharing, this is judged to be a feasible solution.

Decreasing the availability of resources within a system often leads to increasing throughput times. Yet, it is possible to reallocate part of *Nurse 2's* hours and still be able to shorten the throughput time of the entire system. This is somewhat counter-intuitive but becomes clear when understanding how the assessment of patients through the external phone is designed. As patients call in the early morning and book time slots, the external phone creates a large queue that gradually decreases as each patient is assessed. This construction in combination with the new advocated state allows more patients to enter other parts of the system earlier, decreasing throughput times and slightly increasing

utilisation. The decrease in overall throughput time is reached by a large decrease of the throughput time from the external phone outweighing the slight increases in the rest of the throughput times.

Starting with setting up an additional work station to answer and assess patients calling in through the external phone is the suggested first step in implementing the desired future state. This enables *Nurse 1* to aid *Nurse 2* without increasing travel times. Consequently, the hours of *Nurse 2* can continuously be decreased to the level of a half day, one hour at a time. This way, the system will not have to adapt immediately and it could also experiment with a further decrease of *Nurse 2's* hours. Since the computer based model suggests that it is possible to fully reschedule *Nurse 2* to a half day, the ENT clinic could continue reallocating *Nurse 2's* hours until reaching an optimal state.

The specific clinic, in which this thesis was conducted, is viewed as open to change and willing to test new solutions. Although they were not instantly convinced when presented with the thesis' results, there is a good chance that the desired future state will be trialled for implementation. If this project were to be implemented, it is possible that in the trajectory of this, several other simulation modelling projects could be conducted in healthcare in the region by future researchers.

As discussed in section 5.6, a takeaway for future research is that a lack of whole-system view might make it difficult to see where processes or schedules could be optimised, as processes or resources may appear strained as a result of sub-optimisation. However, the results from this thesis show that there still existed room for improvement, giving reason to believe that this phenomena might exist in other healthcare systems as well.

References

Alänge, S. (1994). The new paradigm for industrial practices - total quality management, CIM-working papers - Department of Industrial Management, Chalmers University of Technology: WP 1994-01.

Barnes, R. M. (1956). *Work sampling*. New York: John Wiley & Sons, Inc.

Belvedere, V. (2014). Defining the scope of service operations management: An investigation on the factors that affect the span of responsibility of the operations department in service companies. *Production Planning and Control*, *25*(6), 447–461.

Berg, B. L. (2009). *Qualitative research methods for the social sciences* (7th ed.). Boston: Allyn & Bacon.

Björnberg, A., & Phang, A. Y. (2019). *Health consumer powerhouse: European health consumer index 2018*. https://healthpowerhouse.com/media/EHCI-2018/EHCI-2018-report.pdf Accessed 2020-01-22

Boaden, R., Harvey, G., Moxham, C., & Proudlove, N. (2008). *Quality improvement: Theory and practice in healthcare*. Coventry: NHS Institute for Innovation; Improvement.

Bryman, A., Bell, E., & Nilsson, B. (2005). *Företagsekonomiska forskningsmetoder*. Stockholm: Liber Ekonomi.

Chan, C. W., & Green, L. V. (2013). Improving access to healthcare: Models of adaptive behaviour (Ed. B. T. Denton), In *Handbook of healthcare operations management: Methods and applications*. New York: Springer Science+Business Media.

Chase, R. B. (1978). Where does the customer fit in a service operation. *Harvard Business Review*, *56*(6), 137–142.

Chung, C. (2004). *Simulation modeling handbook: A practical approach*. Boca Raton: CRC Press.

Colldén, C., & Hellström, A. (2018). Value-based healthcare translated: A complementary view of implementation. *BMC Health Services Research*, *18*(1).

Conway, K. (2019). *Value-based healthcare starts with standard definitions*. https://www.hpnonline.com/sourcing-logistics/article/13002078/valuebased-healthcare-starts-with-standard-definitions Accessed 2020-01-24

Devore, J. (2012). *Probability and statistics for engineering and the sciences, 8th ed.* Boston: Brooks/Cole.

Druckler, P. F. (2011). *The new realities*. Oxford; New York: Routledge.

Gillham, B. (2008). *Forskningsintervjun: Tekniker och genomförande* (E. J. Gromark, Trans.). Lund: Studentlitteratur.

Glouberman, S., & Mintzberg, H. (2001). Managing the care of health and the cure of disease—part i: Differentiation. *Health care management review*, *26*(1), 56–69.

Göransson, P., Johansson, E., Kröss, A., Lambrecht, S., Rye, S., & Westberg, C. (2018). Leadership and psychosocial work environment in an intensive care setting, Chalmers University of Technology, Sweden.

Holweg, M., Davies, J., De Meyer, A., Lawson, B., & Schmenner, R. W. (2018). *Process theory: The principles of operations management*. Oxford University Press.

Jacobsson, T. (2012). Operations management in healthcare – principles for creating swift even patient flow and increased accessibility, Chalmers University of Technology, Sweden.

Jingshan, L., Nan, K., & Xiaolei, X. (2017). *Stochastic modeling and analytics in healthcare delivery systems*. World Scientific.

Jones, D. T. (2006). Leaning healthcare. *Management Services*, *50*(2), 16–17.

Kaplan, R. S., & Porter, M. E. (2011). The big idea: How to solve the cost crisis in healthcare. *Harvard Business Review*, *89*(9), 46–64.

Kjøller, A., & Westergaard, N. (2017). The hidden potential of managing failure demand. *Implement Consulting Group*, *October*.

KTH. (2019). *Hållbar utveckling*. https://www.kth.se/om/miljo-hallbar-utveckling/utbildning-miljo-hallbar-utveckling/verktygslada/sustainable-development Accessed 2020-01-27

Kumar, R. (2011). *Research methodology: A step-by-step guide for beginners*. London: Sage Publications Limited.

Law, A. (2014). *Simulation modeling and analysis, 5th ed.* New York: McGraw-Hill Education.

McLaughlin, D. B., & Olson, J. R. (2017). *Healthcare operations management, third edition*. Chicago: Health Administration Press.

Modig, N., & Åhlström, P. (2012). *This is lean: Resolving the efficiency paradox*. Rheologica.

Pruzan, P. (2016). *Research methodology: The aims, practices and ethics of science*. Cham: Springer International Publishing Switzerland.

Ross, T. K. (2014). *Health care quality management: Tools and applications*. New York: John Wiley & Sons, Inc.

Sasser, W. E. (1976). Match supply and demand in service industries. *Harvard Business Review*, *November 1976*, 133–140.

Skinner, W. (1969). Manufacturing – missing link in corporate strategy. *Harvard Business Review*, *47*(3), 136–145.

Slack, N., Brandon-Jones, A., & Johnston, R. (2013). *Operations management*. Harlow: Pearson Education Limited.

Slack, N., & Lewis, M. (2017). *Operations strategy: Fifth edition*. Harlow, UK: Pearson Education Limited.

Wallén, G. (1993). *Vetenskapsteori och forskningsmetodik*. Lund: Studentlitteratur.

World Commission on Environment and Development. (1987). *Our common future*. Oxford: Oxford University Press.

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISION OF SUPPLY AND OPERATIONS MANAGEMENT
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY