Eliminating Waiting time at a 24 hour Medical Emergency Facility.

by

Elize Smuts
25165552

Project leader: Dr. J W. Joubert

submitted in partial fulfilment of the requirements for the degree of
BACHELORS OF INDUSTRIAL ENGINEERING

in the

FACULTY OF ENGINEERING, BUILT ENVIRONMENT
AND INFORMATION TECHNOLOGY

UNIVERSITY OF PRETORIA

October 2010

Executive Summary

Medipark is a 24-hour Medical emergency facility situated in Rooihuiskraal, Centurion. Currently this practice is providing a daily average of 350 patients with vital healthcare services.

The practice is constantly aiming to be the benchmark for the delivery of quality healthcare within the community. Therefore it follows that a patient's negative experience regarding waiting time will radically influence the perception of the quality of the service provided. In emergency situations every minute wasted by being confined to a waiting room will certainly have a detrimental effect on the outcome of the healthcare intervention.

Medipark therefore strives towards establishing innovative procedures that would eliminate patient waiting time.

The first and foremost aim of this project is to reduce patient waiting time without causing an increase in doctor idle time. The focus of this project will however be on: *unscheduled* and *emergency* patients as this has been determined as the category with the most improvement potential.

As solution strategy, a unique triage system (called MediTriage) has been identified and accompanying implementation procedures are suggested. In order to prove that reduced waiting time can be achieved, MediTriage is modeled in the form of a priority queuing system with the use of Lingo® software and verified with the qtsPlus® addin for Microsoft Excel®.

Ultimately this project culminates into the amount of waiting time that can be reduced. This would determine whether or not the endeavour is worth pursuing. In the case of MediTriage, the savings outrank the liability on part of the practice. Eliminating waiting time altogether is unrealistic; however the proven reduction in *unscheduled* and *emergency* waiting time indicates a feasible solution to the initial problem.

Contents

List of Tables

List of Figures

List of Acronyms

ED Emergency Department

FIFO First–In–First–Out

HPCSA Health Professions Council of South Africa

IHI Institute for Healthcare Improvement

JIT Just In Time

LCFS Last–Come–First–Served

SRSC Standard Room Stocking Checklist

TSM Telephone Scripting Method

U/E Unscheduled and Emergency

Chapter 1

Introduction

1.1 How Medipark strives to serve the community

> *"From birth to death, we, as humans, are spills around which healthcare systems revolve. We rely on both public and private organisations to provide preventative care and treat our ailments. The quality of healthcare can therefore not only play a vital role regarding longevity, but also and especially regarding the Quality of Life." - Randolph W. Hall*

Medipark is a 24-hour Medical facility situated in Rooihuiskraal, Centurion. At present Medipark is providing a daily average of 350 patients with valuable healthcare services. Fourteen doctors and specialist physicians are employed on a permanent basis whilst they are supported by another ten locum doctors. These doctors work on a shift basis in order to provide care 24 hours per day and seven days per week. At any given point in time a minimum of one doctor in the small hours of the night to a maximum of 12 doctors during the peak hours are on duty. This is due to the constraining factor of only 12 available consulting rooms.

Furthermore, the remaining permanent positions are filled by three receptionists, four permanent pharmacists as well as two cashiers.

Patients make appointments with their preferred doctor in advance and a 15 minute timeslot is allocated per patient. Unscheduled appointments and emergencies are allowed as each doctor is assigned to one hour per day where he/she is the 'on-call' doctor and will only consult with these specific patients. If at any point in time one of the other doctors has an opening or a patient doesn't keep an appointment, he/she will also attend to emergencies.

The consequences of the system are frequent delays varying between 2 minutes and 2 hours:

- At times there are more patients than the available doctors can attend to;
- at times you will have doctors in excess.

As different doctors have different modus operandi, the allocated 15 minute timeslots cause bottlenecks for certain doctors whereas the same 15 minutes can leave their colleagues with idle time. Doctors are remunerated per patient which implies that idle time can result in suboptimal salaries.

On the other hand, it is significant to realize that long queues in turn will result in unhappy patients.

1.2 Problem statement

1.2.1 Examining the problem

Medipark strives to be the benchmark for quality healthcare in the community. A patient‘s negative experience regarding waiting time will radically influence the perception of the quality of the service provided. Medipark therefore aims to establish innovative procedures that would eliminate patient waiting time.

In emergency circumstances every minute wasted by waiting can have a detrimental effect on the outcomes of the healthcare to be provided. As queues lengthen, workloads increase and the capacity to render quality healthcare to patients, deteriorates (Hall, 2006).

In order to get to the root of this problem, the patient flow at the practice need to be investigated. Patient flow represents the ability of the healthcare system to serve patients quickly and effectively as they move through the stages of care. Hall (2006) explains that when a system works well, patients flow like a river, meaning that each stage is completed with minimal delay. When the system is broken, patients accumulate like a reservoir.

Simplified, good patient flow means that waiting time is minimised; poor patient flow means that patients are exposed to significant delays.

1.2.2 Evident flow problems

The individuality of each doctor results in that some take 10 minutes to consult with a patient whilst others might take up to an hour. This seems to be the most obvious bottleneck in patient flow. Further examinations reveal that emergency patients are seen on an ad-hoc last come first served basis, bumping scheduled patients back in line. Furthermore, when patients arrive late for their scheduled appointment they are still allowed to see the doctor and this incurs an extra waiting period for all the successive patients.

Figure 1.1 shows how the individuality of doctors causes delays in patient waiting time and influences patient satisfaction.

1.3 Project Aim

The key issue is to reduce patient waiting time without causing a significant increase in doctor idle time, which will incur a significant cost for the doctor in terms of loss of income. It is important to note that the aim is not to improve cycle time. The Institute for Healthcare Improvement (IHI) defines cycle time as the total time a patient spends at the practice starting at arrival and stopping after he has finished with the consultation, received medicines from the pharmacy and proceeded to pay and leave. It is necessary to distinguish between the time the patient spends with

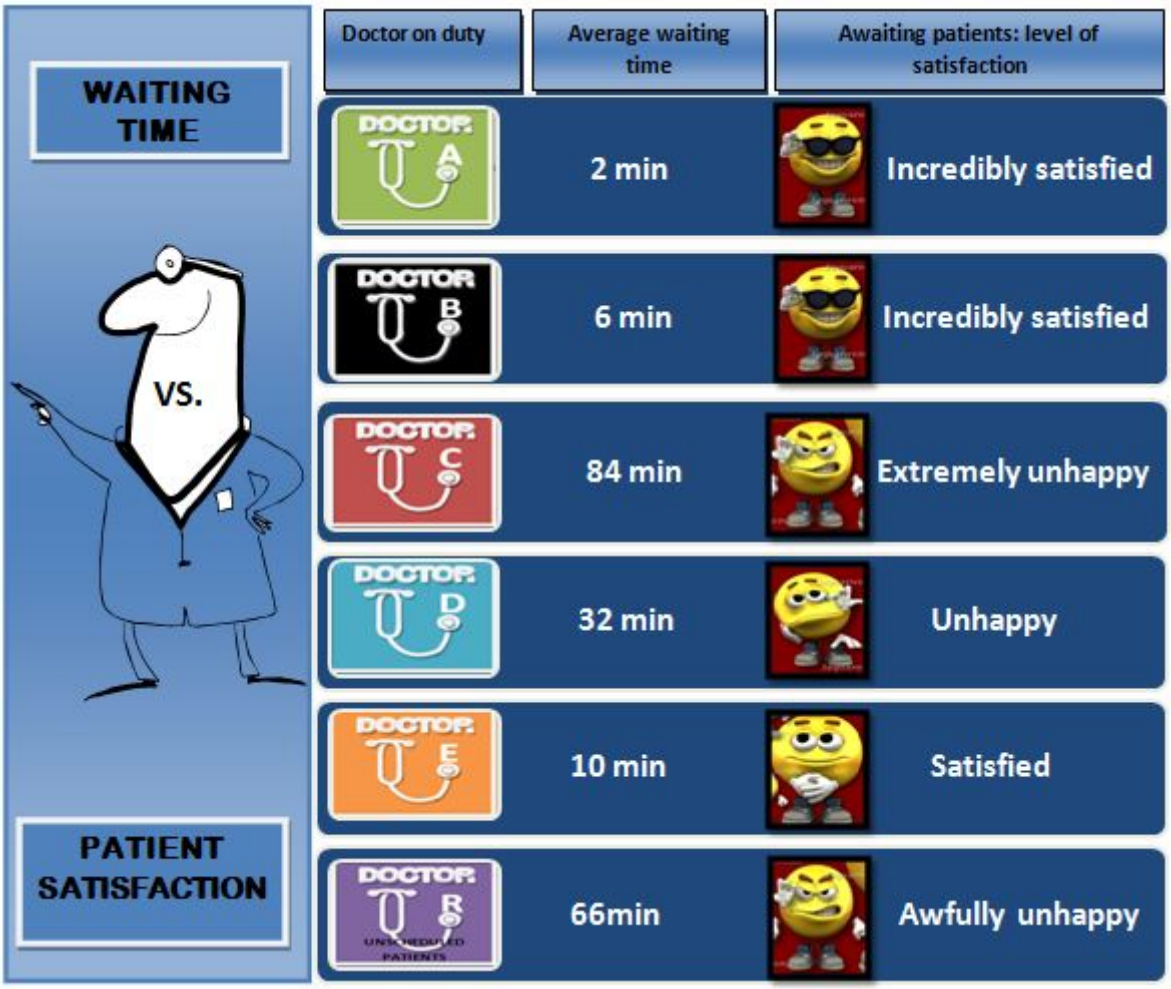


Figure 1.1: Waiting time vs. Patient satisaction

the doctor or other members of the care team (*value-added time*) and the time spent waiting (*non-value-added time*). The goal is not to reduce total cycle time but to maximize the time the patient spends with the doctor or other members of the care team and minimizing idle waiting time. The focus is still on a quality healthcare service, and not on seeing as many patients as possible in the shortest period of time for the highest amount of profit.

At this stage in time, it must be noted that this project will be divided into two segments. One will deal with the permanent doctors and their daily schedules and the other will solely focus on Unscheduled and Emergency (U/E) patients. The reason being that these two categories of patients each have their own unique characteristics and will need contrasting methods to solve each individual need.

1.4 Proposed solution

"No matter how great a system might be, if people don't use the system, the system remains quite useless" (Kelton et al., 2003).

The foremost goal of this project is to establish innovative procedures that would eliminate patient waiting time. To achieve this, the unique model should include the relevant engineering techniques discussed in Chapter 2 and would be created in conjunction with management, doctors, nurses and admin staff to assure the relevance to Medipark.

1.4.1 Reducing emergency patient waiting time

As stated, U/E patients will be dealt with as a separate entity. It is proposed that a triage system is implemented as it is a tried and tested method used around the world in hospital emergency departments. Triage is a lean management technique, discussed in detail in Chapter 2. Triage would significantly reduce the number of unscheduled patients waiting in the queue, as a considerable number of these patients could be assisted by a pharmacist or nurse and need not see a doctor. The feasibility of a triage system will be tested by means of priority queuing models in *Lingo*®.

1.5 Conclusion

In Chapter 1 the waiting time problem was discussed and the goal of the project as well as the research defined. This showed that the Medipark problem is indeed novel. Since no exact heuristic for solving the problem exists, Chapter 2 will demonstrate how a literature study of solutions to similar healthcare problems contributed towards arriving at a unique solution strategy.

Chapter 2

Literature review

2.1 Problem Classification

The elimination of waiting time through an Industrial Engineering approach has been around for a number of years. However the application of these methods in a healthcare environment is a fairly new concept. The research challenges in the healthcare environment are linked to the vast number of variables that have to be taken into account when developing solution models. With regard to patient flow, each system is unique, due to the fact that it solely relies on human elements. According to the IHI, healthcare systems can be changed for the better through a strategy that combines creativity with a number of engineering techniques. However, each unique system needs its own mix of techniques. As found during the preliminary literature review, the methods used depend on the uniqueness of each hospital, Emergency Department (ED) or outpatient clinic.

2.2 Solution Variants

A number of methods exist to solve our optimisation problem. Examination of literature indicated a number of widely used methods.

2.2.1 Lean improvement methods

Since the concept of Lean thinking originated from manufacturing industry, it may be argued that the service sector and especially the healthcare sector may not gain from it (Khurma et al., 2008). However, Womack and Jones (2003) advocate the application of lean thinking in medical systems. They argue that the first step in implementing lean thinking in healthcare is to put the patient in the foreground and include time and comfort as key performance measures of the system. Lean concepts are especially applicable in the healthcare environment due to the fact that just as in manufacturing it also strives for the *quest for zero defects*, continuous improvement and Just In Time (JIT) principles. Having multi-skilled teams taking care of the patient and an active involvement of the patient in the process is emphasized by Khurma et al. (2008). Several case studies on lean thinking initiatives in the health care sector can be found in Miller (2005) and Spear (2005). In a recent publication

by the IHI, two healthcare organizations in the United States of America showed a positive impact on productivity, cost, quality, and timely delivery of services after having applied lean principles (Miller, 2005). Lean methods proved to be very useful in large emergency departments where patients move through various processes (registration, triage, lab tests etc). A variety of lean tools can be utilised to eliminate *waste* in healthcare processes. The most widely used tools include **Six Sigma** and **Triage**.

Six Sigma Quality improvement approach

Breyfogle and Salveker (2004) advocate quality improvement in healthcare and give an example of how lean management principles can be applied to healthcare processes through the use of the Six Sigma methodology, which in many ways resemble the lean production techniques. Six Sigma seems to be a useful tool in identifying underlying problems leading to long waiting times, and could prove to be useful when classifying patient flow problems. A study of outpatient waiting time done by Abdullah (2004) at the University Hospital Kebangsaan in Malaysia showed how a Six Sigma approach can be used to improve the quality of care and patient satisfaction (see Figure 2.1). In conclusion of the study the main factors leading to long patient waiting time had been identified but no improvements were made.

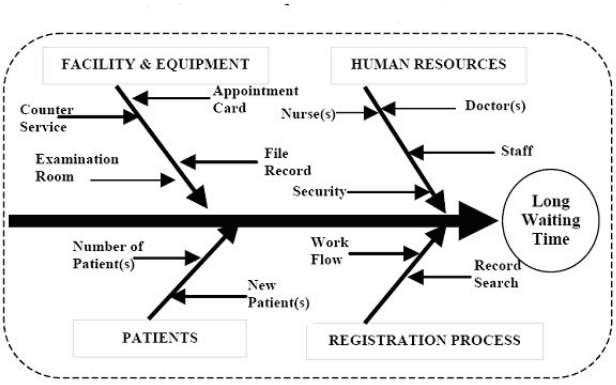


Figure 2.1: Causes of long waiting time (Abdullah, 2004)

Six Sigma was first introduced by *Motorola*®in 1986. It is a method used to measure the quality of a process when it is forced to fulfil the customers needs, which approaches perfection. Six Sigma sets the benchmark for perfection at 3.4 defects per million opportunities. Data and statistical analysis is used to identify defects in a process and reduce variation. A variety of methods can be used to measure, identify and eliminate these defects in the quest for *zero defect*. These methods mostly comprise of existing project management, statistical and analytical tools. According to Abdullah (2004) the following Six Sigma toolkits can be applied in healthcare problems:

- Descriptive statistics
- Control Chart

- Flow Chart
- Pareto Chart
- Cause and Effect Diagram
- Quality Function Deployment (QFD) Chart

All of these toolkits can be applied within the DMAIC method, which serves as the roadmap to Sigma quality improvement. DMAIC consist of five steps, as the acronym adequately states, *Define-Measure-Analysis-Improve-Control* An explanation of each step and the appropriate toolkits utilised is given in Figure 2.2.

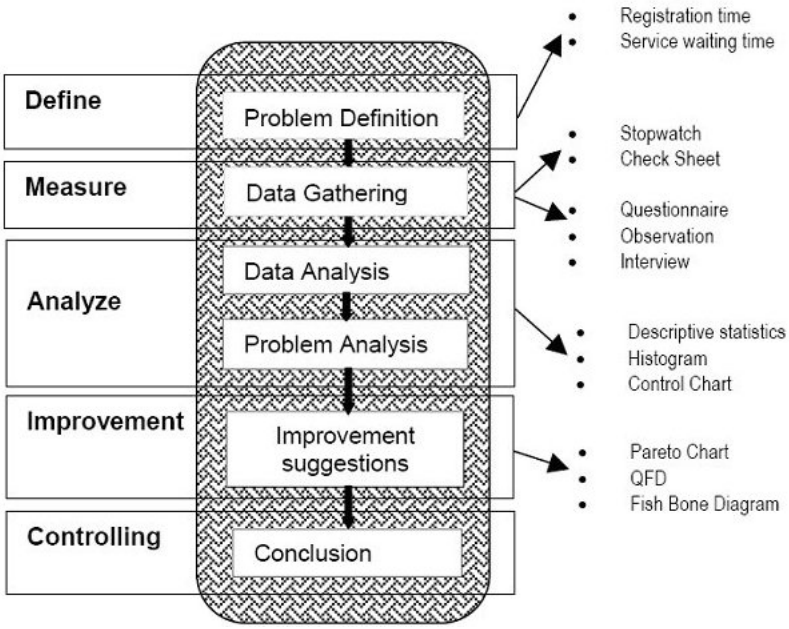


Figure 2.2: DMAIC Steps (Pyzdek, 2001)

Triage system

Triage is a internationally recognised lean management technique. It is a system that is used with great success in hospital EDs across the world. Triage is defined as the system where all U/E patients are classified according to the severity of their medical condition when they arrive at an ED. All of the patients entering a queue without an appointment will be seen by a triage nurse and given a triage number from 1–5 denoting his state of health (sometimes colours are used instead of numbers, as not to alert patients to the severity of their condition). Level 1 and 2 patients require immediate care from a doctor while level 3 patients are expected to be seen within 30 minutes. Level 4 and 5 patients are typically seen before level 3 patients simply because they can be serviced quickly and discharged shortly afterwards (Khurma et al., 2008).

2.2.2 Queuing Theory

Queuing theory is the one method that is most widely used by healthcare organisations to improve their current systems. The IHI has launched a project to educate healthcare providers in the principles of queuing theory with the aim of converting more organisations to the wonders that it can achieve for flow management and throughput optimisation. The organisations that care for persons who are ill and injured vary widely in scope and scale (Fomundam and Herrman, 2007). Despite these differences, one can view the healthcare processes that these organisations provide as queuing systems in which patients arrive, wait for service, obtain service and then depart. The resources in these queuing systems are the trained personnel and specialised equipment that these activities and procedures require.

In reality, patients do not arrive exactly every 15 minutes nor are they consulted with the same duration of time. Queuing theory is composed of a group of equations and relationships which can provide analytical solutions to show these variations in processes. Stout and Tawney (2005) feel that accounting for variability in healthcare systems has become the rule, and not the exception. They also state that queuing theory assumes steady-state conditions where the distributions of arrival rates and service times are stationary over time and the system is operating in equilibrium. The problem with this statement in the healthcare environment remains that ED queues might never reach a steady state. ED systems behave chaotically with variable service and arrival rates being the norm. To apply queuing principles to our specific problem, we would have to find an innovative way to account for the variability or force the system into a steady state. If each individual doctor is modeled as a single server, it could be possible to assume either a steady arrival rate or steady service rate. In their work on a forecasting a demand model for an ED, Stout and Tawney (2005) demonstrated what the effects would be when one keeps the arrival rate constant and the service rate variable (and visa versa). Table 2.1 and Figure 2.3 are an adaptation of the explanation provided by Stout and Tawney (2005).

List of symbols used:

$\lambda \triangleq$ Arrival rate per hour.
$\mu \triangleq$ Service rate per hour.
$\rho \triangleq$ Traffic intensity ratio (utilisation factor).
$L_q \triangleq$ Mean number of patients in system.
$L \triangleq$ Mean number of patients in queue.
$W_q \triangleq$ Mean time that a patient spends waiting in the queue (hours).
$W \triangleq$ Mean time that a patient spends in the system (hours).

Table 2.1 and Figure 2.3 illustrate conditions when the arrival rate is variable and the service time is held constant for a single server model. It can be noticed that the doctor utilisation rate is practically benevolent until the arrival rate reaches about 80%. It is evident that as the 80% level is reached the curve of the waiting time line rises sharply, increasing to infinity as utilisation approaches 100%. This places the doctor under immense strain. Due to the increased opportunity for errors, the system cannot indefinitely continue to operate under such conditions. Likewise, we can repeat the computations in table 2.1 while holding the arrival rate constant and

Table 2.1: Variable Arrival Rates versus Constant service rates

	$\mu \triangleq 8$						
λ	**6.50**	**5.00**	**5.25**	**5.50**	**5.75**	**6.00**	**6.25**
L_q	1.04	1.25	1.51	1.84	2.25	2.79	3.52
L	1.67	1.91	2.20	2.56	3.00	3.57	4.33
W_q	0.21	0.24	0.28	0.32	0.38	0.45	0.54
W	0.33	0.36	0.40	0.44	0.50	0.57	0.67
$\rho(\%)$	**62.5**	**65.6**	**68.7**	**71.8**	**75.0**	**78.1**	**81.2**

λ	**6.75**	**7.00**	**7.25**	**7.50**	**7.75**	**8**
L_q	4.56	6.13	8.76	14.06	30.03	$\propto$
L	5.40	7.00	9.67	15.00	31.00	$\propto$
W_q	0.68	0.88	1.21	1.88	3.88	$\propto$
W	0.80	1.00	1.33	2.00	4.00	$\propto$
$\rho(\%)$	**84.3**	**87.5**	**90.6**	**93.7**	**96.8**	**100**

varying the service time. Moreover the result stays the same; there is a sudden exponential increase in the queue size when working above the 80% utilisation mark. Queuing models can only handle a limited number of arrivals when there is only one doctor as server.

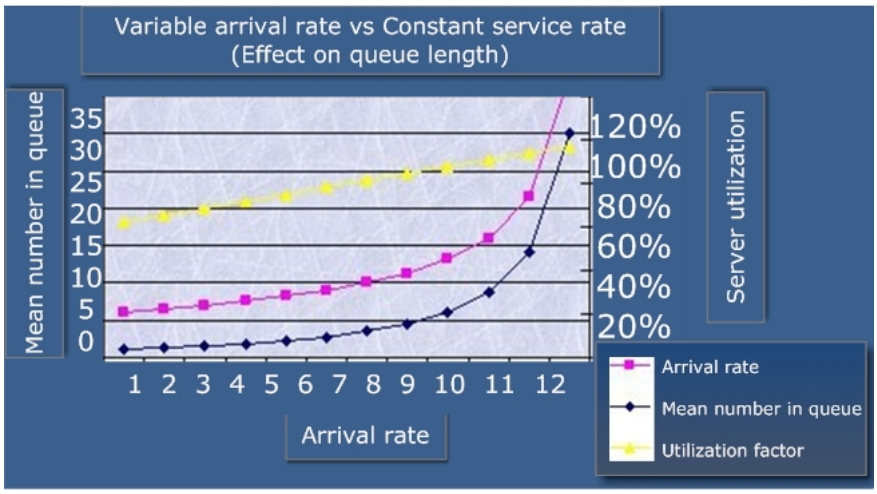


Figure 2.3: Variable arrival rate vs. constant service time

In healtcare systems where patient classes have different priorities (Triage system), Siddhartan et al. (1996) proposed the use of a priority queuing discipline. This would give priority according to the category (or triage number) and then use a First–In–First–Out (FIFO) discipline within each category. According to Fomundam and Herrman (2007) they then found that the priority discipline reduces

the average waiting time for all the patients: conversely, while the waiting time for higher priority patients reduces, lower priority patients have to undergo a longer average waiting time.

In their survey of queuing theory applications in the field healthcare, Fomundam and Herrman (2007) covered every possible angle for applying analytical queuing theory models to real world healthcare systems. They are of the opinion that with this knowledge base, it is reasonable for any analyst to grasp, adjust and apply these principles to his own situation.

2.2.3 Optimisation through scheduling

In most scheduling situations, the information available to the decision maker is both partial and vague. Earlier research on staff scheduling was primarily aimed at developing efficient heuristics. Miller et al. (1976) were the first to formally address the preference scheduling problem. Starting with an initial solution, they developed a greedy neighborhood search procedure to find local optima. Howell (1998) solved the cyclic scheduling problem by combining intuitive information of what constitutes a good schedule with greedy exchange measures. More recently, metaheuristics, such as tabu search, simulated annealing, and genetic algorithms, have been used to solve various midterm scheduling problems (Aickelin and Dowsland, 2000; Nonobe and Ibaraki, 1998). Nevertheless, it is often difficult for heuristics to cope with conflicting hard (must be satisfied) and soft (can be violated at a cost) constraints in a computationally efficient manner (Bard and Purnomo, 2009). Motivated by the need to balance solution quality with computational effort, Causmaecker and den Berghe (2003) showed how to combine metaheuristics and coverage relaxation algorithms to address practical concerns in a real scheduling environment.

Most of the reviewed literature focussed on cyclic and preference scheduling models as the basis from which to formulate a unique scheduling IP for each of their specific applications. In their work on cyclic preference scheduling, Bard and Purnomo (2009) explain the difference between the two models. *Preference scheduling* applies a common set of constraints as well as a cost measure that is designed to achieve a balance between staff approval and the use of outside resources (Burke et al., 2004). The constraints are hospital dependant, but always classified as either hard or soft. Preference scheduling can deal with issues associated with the quality of the schedule as judged by the presence of disagreeable work patterns (Causmaecker and den Berghe, 2003). *Cyclic scheduling* is defined as a method where fixed patterns are established (ex. days on and days off) and the staff is rotated through them on a continuous basis.

A vast amount of nurse scheduling literature is available (for a survey of nurse rostering see (Burke et al., 2004)), on the other hand doctor scheduling has not been done before or is simply not available for review. The reason for this might be that medical practices normally employ only a handful of doctors and have no need for intricate scheduling; and in large EDs the patients do not make appointments and all doctors are *on-call*.

The biggest difference between nurse- and doctor scheduling is the varying constraints. Nurse schedules have a fixed demand (number working per shift) with fixed

shift time–slots (days and hours of day) and would normally be scheduled so that they have two consecutive days off. With doctors the demand depends solely on the number of patients available to be seen. Nursing works with a fixed number of staff to be scheduled and rotated while doctor scheduling should stay uncapped to allow for the addition of locum doctors when patient demand shows the need for extra hands. Due to the fact that nurses are employed on a contract basis, they are forced to work a certain number of hours per week (for which they get paid per hour), doctors on the other hand are paid for every patient they see and when working in the South African private sector; they will never be fixed to a certain number of hours of service.

As a result, it would not be possible to use a specific heuristic or adapt an outline from work that has been reviewed. Bard and Purnomo (2009) suggest that it would be necessary to develop a hybrid algorithm comprising of both heuristic and exact procedures.

2.2.4 Simulation Methods

Due to the complexity of healthcare systems, discrete event simulation (DES) has proved to be an effective tool used for process improvement (Khurma et al., 2008). An ED is the main area through which thousands of patients flow every year. For this reason, several studies have been conducted to increase the efficiency of the ED using simulation tools. Most studies found in literature aim at reducing waiting times and increasing service level (throughput) by improving the actual care process (Barnes and Laughery, 1998) or by increasing the size and the operation of the ED (Benneyan, 1997).

2.3 Solution Strategies

It could be useful to combine some of the reviewed solution methods. For instance, queuing models and simulation models each have their advantages. It is clear that queuing models are simpler, require less data, and provide more generic results than simulation (Green, 2006). However, discrete event simulation permits modelling the details of complex patient flows. Kao and Tung (1981) and Tucker et al. (1999) used simulation to validate, refine or otherwise complement the results obtained by queuing theory.

In the specific problem that Medipark is facing, the Six Sigma approach will be a useful tool in establishing patient flow problems before the start of a complex IP formulation.

2.3.1 Reducing scheduled patient waiting time

Simulation modelling

Simulation is a valuable tool both to evaluate new processes and to understand and demonstrate the current causes of delay. An initial simulation would therefore serve as the benchmark for improving waiting time at Medipark. A later simulation

model can be used to verify the results and show the improvements when a scheduling system has been implemented. These models would require a vast amount of system variable and data gathering such as: consulting times, demand, arrival time, no-show rates of booked patients, late arrivals and emergency patient arrivals.

Queue analysis

Queuing analysis is invaluable when executed on a real-time basis to highlight the delays currently experienced throughout the whole patient cycle. If the findings are correctly analysed it will help to better understand delays and act on them through reallocation of doctors and appropriately prioritising patients. Queuing analysis would require in depth information gathering on the following system variables: variable arrival rates, system size, bottlenecks, reneging, blocking, cycle time, patient type priorities as well as waiting time and resource utilisation analysis.

The queuing theory techniques discussed in Section 2.2.2 can be implemented by seeing each individual doctor as a *single server*. The variables needed for each doctor such as λ , μ , ρ , W_q, W, L_q and L can now easily be extracted from the initial si - mulation model. It is now possible to use these variables and the principles of queuing theory to find the optimal size of scheduling time-slots for individual doctors. For instance, by using the optimal arrival rate value found for each doctor, we could determine that bookings should be made in differing time-slot intervals such as 10 minutes for doctor A, 15 minutes for doctor B and 40 minutes for doctor C. This method would accurately balance the patient flow (as defined in Chapter 1).

Optimisation model

At this stage, a hybrid algorithm should be modelled in the place of a traditional IP, and should be based on *Preference Scheduling* approaches. This model should incorporate the results concerning time-slot intervals as obtained through queuing analysis. A number of clinical constraints should be included in the model; such as number of rooms, available doctors, available shift time-slots etc. A crucial and difficult part of the model would be to include the seasonal demand patterns. Seasonal demand is defined as the number of patients that arrive to be consulted by a doctor and depends on the time of day, day of the week, season of the year, public and school holidays and pandemic outbreaks (such as the swine flu and measles).

The main objective of the model is to generate schedules that would optimise the number of doctors working on each shift so that there is a perfect balance between available doctors and patients to be seen. In other words, no doctor should be idle on a particular shift and patient waiting time should be kept to an absolute minimum.

2.3.2 Reducing emergency and unscheduled patient waiting time

These two types of patients constitute the largest amount of waiting time at Medipark. Unscheduled patients are the category which spends the longest time waiting, with an average of 66 minutes. Emergency patients might not always wait that long, because they are seen on an ad-hoc last come first served basis which in turn

bumps scheduled patients back in line causing considerable delays for them. As found in review of the literature, the best way of solving this problem is through implementation of the tried and tested Triage system.

Implementing a Triage system

Instead of implementing the standard hospital ED Triage system, it will be modelled as a priority queuing discipline and adapted accordingly for unique use in Medipark.

2.4 Conclusion

With a fully developed knowledge base it should now be possible to solve the two waiting time problems with the approaches discussed in this chapter.

Chapter 3

Eliminating Waiting Time - minute by minute

This chapter will show how engineering techniques are utilised to arrive at a hybrid model to solve the U/E leg of our waiting time problem as defined in Chapter 1. Firstly, we show how a few practical changes can influence patient flow in the practice before moving on to solving *U/E* patient waiting time.

3.1 The practical way to reduce waiting time minutes

According to Barnes and Laughery (1998) the first step to reduce waiting time and increase the service level would be to improve the actual care process. As part of Lean improvement through a Six Sigma approach, a fishbone diagram (see Figure 3.1) and Unplanned Activity Cards (see Appendix A) had been utilised to show the main factors leading to long patient waiting times.

From the analysis it is clear that most of these factors originate from faults at the reception desk. Causes of delay include:

1. Patients arriving late. These patients will still be put into the queue which incurs extra waiting time for all consecutive patients. This means that the specific doctor sat idle for the 15 minutes that had been allocated to this patient and thus will not be on time for the rest of the day.
2. Entire families showing up for one 15 minute appointment slot. Meaning that the doctor will have to take more than the allocated 15 minutes and will run late for all consecutive appointments.
3. A receptionist forgets to draw a file and place the patient in queue. Resulting in an idle doctor and furious patient that can sometimes wait for an hour before asking whether or not he is still in line to see the doctor.

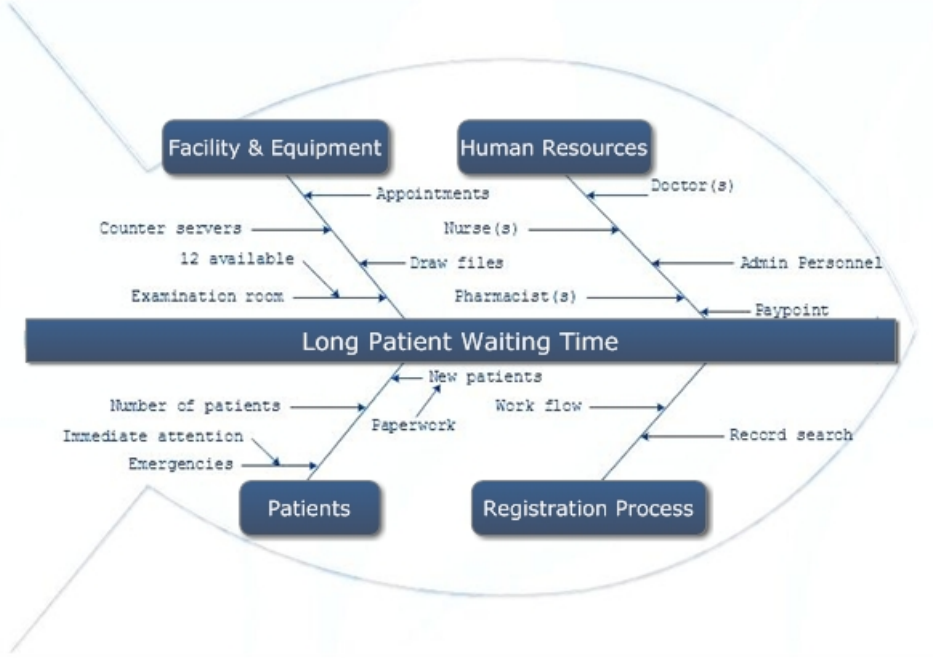


Figure 3.1: Causes of long waiting time

4. Wrong files are being drawn, which also leaves a patient out of the queue (and furious).

5. Patients that arrive early are immediately filed in the queue (taking up another patients time slot and forcing them to wait).

The unplanned activity cards showed that the only cause of delays not linked to reception is due to missing equipment and supplies. This can be rectified by using a *Standard Room Stocking Checklist* (SRSC, see Appendix A) to ensure efficient stock levels at all times.

3.1.1 A new protocol for reception

A few practical changes must be made to reception before we can start with intricate engineering models to save time. To start off with, we are implementing a new Telephone Scripting Method (TSM, see Appendix A), that has been adapted from an outline developed by the IHI. TSM is an outline set for use by all reception personnel that will ensure that all patients are treated fairly and all future bookings made equally. The new protocol set to solve problems 1–5 above is as follows:

1. When a patient arrives more than 5 minutes late for their appointment, he/she should be given the option to either wait in line as an U/E patient, or to reschedule their appointment at a later time or date.

2. Reception should ask who will be seeing the doctor. If they want more than one member of the family to consult when they have made only one appointment, should be forced to reschedule with enough time allocated to them or wait in line with the U/E patients.

3. Receptionists should keep track of what they are doing by ticking the patient names in the appointment book upon arrival and highlighting the surname once the file has been placed in queue. Thus ensuring that no patient is erroneously left out of the queue.

4. The receptionist should ensure that the surname on the file being drawn corresponds to the surname on the computer screen before highlighting the surname and adding it to the queue.

5. When a patient arrives early, the receptionist should take great care in ensuring that the right order is maintained when adding the file to the queue.

These are all practical ways to save valuable bits and pieces of waiting time. Unfortunately they can not be measured with engineering techniques. The savings that will have a significant impact and that can be measured, will follow in the sections below.

3.2 Patient Demand Patterns

In order to arrive at a feasible solution, it is crucial to include demand patterns in the final model. This would enable the model to resemble a real world situation as closely as possible. *Patient demand* is defined as *the number of patients that arrive to be consulted by a doctor each day.*

Strenuous time studies were conducted to collect consultation data from the past 6 months in the form of day–to–day totals. The patterns were expected to fluctuate depending on the day, season and public as well as school holidays. The results however, painted a surprisingly different picture. Findings were tabulated in the form of Excel® sheets as indicated in Appendix B.

3.2.1 Demand analysis

The findings of the time studies have to be analysed to find patterns and averages. It is important to note that we cannot make use of quality control methods such as variable and attribute control charts. These methods can be used in a manufacturing environment where we can force a system into a steady state and strive for zero perfection. In healthcare, however, the system will always resemble reality and cannot be forced to behave otherwise. For example; some days will have much lower daily totals than others. With control charts; these days would be taken out of the system due to the fact that they fall outside *specification limits.* In actual fact these days fall below specification, because they fell on public holidays. Thus we will not discard these days as accidental defects.

Limitations: For simplistic reasons, we will leave pandemic outbreaks out of the analysis as this could be covered in an entire forecasting project of its own.

Assumptions: U/E patient demand is categorised according to peak and off-peak times as these patients arrive at different rates throughout the day. Scheduled

patient demand on the other hand is not set by the time of day as it is always busy and the appointments are scheduled evenly throughout the day.

Unscheduled and Emergency Patient demand analysis

An analysis of the daily totals found during time studies at the practice, indicated that U/E patients arrive in a steady stream throughout the day. However, this stream will fluctuate according to the time of day. For this reason arrivals will be dealt into 3 categories (Table 3.1) and handled accordingly. Please note that School Holidays were found to have a negligible effect on daily averages.

Table 3.1: Arrival Times

	Off–Peak	**Peak**	**Night**
Monday–Friday	8h00–16h00	16h00–23h00	23h00–8h00
Saturday & Sunday	8h00–14h00	14h00–23h00	23h00–8h00
Public Holiday	8h00–14h00	14h00–23h00	23h00–8h00

By counting all of the U/E patients that arrived for service for the months of February to June, the daily averages are indicated in Table 3.2.

Table 3.2: U/E Arrival Patterns

Time:	**Off–Peak**	**Peak**	**Night**
Monday	27.55	30.61	8.11
Tuesday	18.76	31.99	5.78
Wednesday	24.06	27.94	6.69
Thursday	17.8	25.26	6.65
Friday	20.00	18.36	5.94
Saturday	21.74	24.76	7.80
Sunday	25.50	31.39	7.39
Public Holiday	19.67	34.33	8

Scheduled Patient demand analysis

Scheduled patient arrivals were found to be evenly distributed throughout the day. However it does fluctuate according to seasons. The months following May to August will hence be know as *peak* season with September to April as *off–peak*.

The values in Table 3.3 refer to the number of patients that made and held appointments during the peak and off–peak months respectively.

Table 3.3: Scheduled Arrival Patterns

Season:	**Off–Peak**	**Peak**
Monday	310	320
Tuesday	247	270
Wednesday	240	244
Thursday	232	241
Friday	197	201
Saturday	168	181
Sunday	164	176
Public Holiday	187	207

3.3 Solving Unscheduled and Emergency Waiting time

As discussed in Chapter 2, the use of a special Triage system is proposed in order to solve the current U/E waiting time problem. This unique version of the Triage system will henceforth be known as *MediTriage* (as manner of distinction). Firstly, we make use of an Arena® simulation model to demonstrate the current scenario and set the benchmark for improvement accordingly. A detailed description of the concepts behind *MediTriage* will follow.

3.3.1 Setting the benchmark for improvement

In order to show the impact that the project will have in terms of the savings that can be achieved, we have to set a benchmark from which to improve. The chosen method was to model the current U/E waiting time scenario with the use of an Arena® simulation model. The logic behind this model is shown in Figure 3.2.

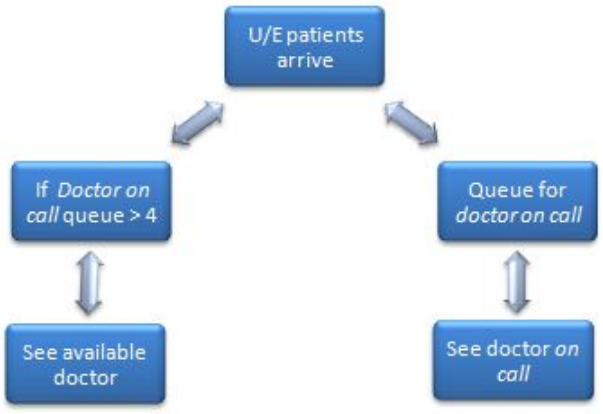


Figure 3.2: Modelling logic for the current scenario in Arena®

The diagram shows the patients arriving at a Poisson distribution rate and then

sends them off to the *doctor–on–call* queue. If, however, the queue here grows larger that 4 patients waiting, any other idle doctor is allowed to take a patient from the queue to be served. See Appendix B1 for the complete Arena® model, which includes all the doctors and other functions such as the pharmacy. See Table 3.4 for the results obtained from the simulation.

Table 3.4: Current waiting time scenario

	λ	W_q
Monday	3.88	25.38
Tuesday	3.38	30.92
Wednesday	3.47	33.72
Thursday	2.87	19.13
Friday	2.56	14.62
Saturday	3.10	23.50
Sunday	3.79	49.29
Public Holiday	3.60	39.04
Mean Service Time ($\frac{1}{\mu}$)	0.21	

Legend: $\lambda \triangleq$ Arrival Rate.
$W_q \triangleq$ Mean waiting minutes in the queue (per patient).

3.3.2 MediTriage

Triage is a process of determining the priority of patients treatments based on the severity of their condition. This rations patient treatment efficiently when resources are insufficient for all to be treated immediately. The term comes from the French verb *trier*, meaning to separate, sort, sift or select. For the purpose of designing a unique triage system for Medipark we will sort patients according to the urgency of their need for care, with the intention of reducing waiting time for all patients involved. MediTriage will sort patients into 5 categories or levels. Level 1 patients are seen as life threatening and should be served on an ad-hoc Last–Come–First–Served (LCFS) basis. The patients with triage level 2 and 3 will be required to wait for the doctor *on-call*. Furthermore, those with triage level 4 can be seen by a nurse, and level 5 should be referred to the pharmacist.

MediTriage levels explained:

- **Level 1:** *Patients that require immediate life–saving intervention.* These patients have first priority over available resources. Afflictions of this level include: cardiac arrest, strokes, seizures, severe head and chest wounds, stab and gunshot wounds, 1st and 2nd degree burns, unconsciousness, unresponsiveness and severe trauma.

- **Level 2:** *Patients in stable condition that require care from a doctor within 30 minutes.* Conditions include: minor trauma, dog bites, major cuts, 3rd degree burn wounds, broken bones, psychiatric mania and migraine.

- **Level 3:** *Patients that require care from a doctor, but not immediately.* These can afford to wait for more than 30 minutes and include general ailments such as the flu, chronic conditions (high blood pressure, diabetes, cholesterol, depression, contraception etc.), prescriptions, sick notes and patients that have two or more symptoms when consulting the pharmacist (one being a high fever).

- **Level 4:** *Patients that can be seen by a nurse.* These might have minor injuries (scrapes, cuts, burns, dog bites etc.) for which first–aid care would be sufficient. Other needs include: taking of blood pressure and temperature, testing of urine samples, family planning, administering injections (gout, vitamins etc.), dressing of wounds, removing stitches, removing plaster of paris, vaccinations (travel, flu, rabies etc.) and baby immunisations.

- **Level 5:** *Patients that can be referred to the pharmacist.* These are patients presenting with minor ailments such as general aches and pains, common colds, nausea, gastro-intestinal problems, headaches, coughs and sore throats etc. (Note: only if less than 3 of the mentioned symptoms are present and there is no presence of dehydration or fever.)

A feasibility study of 100 consecutive U/E *walk-in* patients revealed the arrival probabilities as shown in Figure 3.5.

Table 3.5: Arrival probability of patients for triage

Triage Level	**1**	**2**	**3**	**4**	**5**
Probability	0.08	0.2	0.32	0.24	0.16

It was found that each triage level takes a certain amount of time to be served by a member of the care team. These times will be refered to as the mean service rates as indicated in Table 3.6.

Table 3.6: Mean service times and rates for triage patients

Triage Level	**1**	**2**	**3**	**4**	**5**
Service Time (minutes)	25	20	12	10	5
Service Rate (patients/hr)	2.4	3	5	6	12

3.4 Conclusion

With a firm understanding of the concepts behind *MediTriage* and all the necessary input data collected, it is possible to proceed with the development of the unique hybrid model.

Chapter 4

MediTriage Conceptual Framework

The aim of this chapter is to provide the reader with the academic background for *priority queuing* disciplines and the knowledge as to why these are used to model the *MediTriage* system. Furthermore, to prove what the proposed *MediTriage* system can achieve in terms of savings, various solution scenarios will be modeled in the form of basic and adapted *priority queuing* disciplines.

4.1 Priority Queuing Models

As found during the literature study, MediTriage can be seen as a *priority queue* where level 1 corresponds to service priority 1. In priority schemes customers (*patients*) with the highest priorities (*triage level 1*) are selected for service ahead of those with lower priorities, independent of their time of arrival into the system. There are two further refinements possible in priority situations, namely *preemption* and *non–preemption*. In the latter, low priority patients receive service only when no high priority patients are waiting, but the low priority patient who is currently receiving service is not interrupted if a high priority patient arrives and all servers are busy. In the preemptive queue discipline, however, the service to a low priority patient is interrupted in this event. This section will show the logic behind basic non–preemptive and preemptive system modeling (with reference to Gross et al. (2008)) as well as refinements necessary when using multiple servers. Thereafter, an explanation of our unique MediTriage model will follow.

4.1.1 Non–preemptive System with Many Classes

Suppose that patients of the kth priority (the smaller the number, the higher the priority) arrive at a single channel queue according to a Poisson process and that these patients wait on a FIFO basis within their respective priorities.

We define:

$\mathbf{K} \triangleq \{1, \ldots, 5\}$.

$\lambda_k \triangleq$ Rate at which patients of priority k arrive to join the service queue, where $k \epsilon \mathbf{K}$.

$\mu_k \triangleq$ The service rate (consultation time in minutes) with which priority class k is served, where $k \epsilon \mathbf{K}$.

$\rho_k \triangleq$ The traffic intensity (service ratio) for patients of priority class k, where $k \epsilon \mathbf{K}$.

$\sigma_k \triangleq$ Sum of traffic intensities (service ratio) for priority class k, where $k \epsilon \mathbf{K}$.

$n_k \triangleq$ The number of patients of priority class k that are in the queue ahead of any new arrival, where $k \epsilon \mathbf{K}$.

$S_k \triangleq$ The time required to serve (consult) n_k patients of priority class k already in the queue ahead of an arriving patient, where $k \epsilon \mathbf{K}$.

$S_0 \triangleq$ The time required to finish the patient already in service (where this value could be zero in the event that the system is empty upon arrival), where $k \epsilon \mathbf{K}$.

$T_q \triangleq$ Any new arrival´s waiting time.

$n'_k \triangleq$ The number of patients of priority class k who arrive later and go to service ahead of a new arriving patient , where $k \epsilon \mathbf{K}$.

$S'_k \triangleq$ The time required to serve (consult) n'_k customers of priority class k already in the queue ahead of an arriving customer, where $k \epsilon \mathbf{K}$.

$W_q^{(i)} \triangleq$ The expected waiting time (in minutes) in the queue for a patient of priority class i, where $i \epsilon \mathbf{K}$.

$W_q \triangleq$ The expected waiting time (in minutes) in the queue for all patients.

$L_q^{(i)} \triangleq$ The expected queue size (in number of patients) for the queue of patients with priority class i, where $i \epsilon \mathbf{K}$.

$L_q \triangleq$ The expected queue size (in number of patients) for queues of all priorities.

Let the service distribution for the kth priority be exponential with mean $\frac{1}{\mu_k}$. A unit that begins service completes its service before another item is admitted, regardless of priorities. Then;

$$\rho_k = \frac{\lambda_k}{\mu_k} \quad \forall \quad k \text{ with } k \epsilon \mathbf{K} \tag{4.1}$$

$$\sigma_k = \sum_{i=1}^{k} \rho_i \quad (\sigma_0 \equiv 0, \sigma_5 \equiv \rho) \tag{4.2}$$

The system is stationary for $\sigma_k = \rho \leq 1$

$$T_q = \sum_{k=1}^{i-1} S'_k + \sum_{k=1}^{i} S_k + S_0 \tag{4.3}$$

Taking expected values of both sides gives

$$W_q^{(i)} \equiv E[T_q] = \sum_{k=1}^{i-1} E[S'_k] \quad + \quad \sum_{k=1}^{i} E[S_k] + E[S_0] \tag{4.4}$$

From the uniform property of the Poisson process, $E[n_k]$ equals the time–average number of priority-k patients in the queue, or $L_q^{(k)}$. Little´s formula then gives

$$E[n_k] = L_q^{(k)} = \lambda_k W_q^{(k)} \tag{4.5}$$

Because the service times are independent of n_k,

$$E[S_k] = \frac{E[n_k]}{\mu_k} = \frac{\lambda_k W_q^{(k)}}{\mu_k} = \rho_k W_{q^{(k)}} \tag{4.6}$$

For a Poisson arrival process, the average number of priority-k arrivals during the current arrival´s wait in queue is

$$E[n'_k] = \lambda_k W_q^{(i)} \tag{4.7}$$

Therefore

$$E[S'_k] = \frac{E[n'_k]}{\mu_k} = \frac{\lambda_k W_q^{(i)}}{\mu_k} = \rho_k W_{q^{(i)}} \tag{4.8}$$

Combining equations 4.4 and 4.6, we have

$$W_q^{(i)} = \frac{\sum_{k=1}^{i} \rho_k W_q^{(k)} + E[S_0]}{1 - \sigma_{i-1}} \quad \forall \quad i \tag{4.9}$$

These equations are linear in $W_q^{(i)}$. The solution to 4.9 was found by Cobham (1954), by induction on i. That solution is

$$W_q^{(i)} = \frac{E[S_0]}{(1 - \sigma_{i-1})(1 - \sigma_i)} \tag{4.10}$$

Now, S_0 (the remaining service time of the customer in service at the time of the arrival) has the value 0 if the system is idle; hence

$$E[S_0] = Pr\{\text{system is busy}\} \cdot E[S_0 | \text{ system is busy}]$$

The probability that the system is busy is

$$\lambda \cdot (\text{expected service time}) = \lambda \sum_{k=1}^{5} \frac{\lambda_k}{\lambda} \frac{1}{\mu_k} = \rho \tag{4.11}$$

Also,

$$E[S_0|\text{ system is busy}] = \sum_{k=1}^{5}(E[S_0|\text{system is busy with priority–}k\text{ customer}]$$

$$\times\ Pr\{\text{system is busy with priority–}k\text{ customer} \mid \text{system is busy}\}) = \sum_{k=1}^{5}\frac{1}{\mu_k}\frac{\rho_k}{\rho}$$

Therefore

$$E[S_0] = \rho\sum_{k=1}^{5}\frac{1}{\mu_k}\frac{\rho_k}{\rho} = \sum_{k=1}^{5}\frac{\rho_k}{\mu_k} \tag{4.12}$$

Plugging 4.12 into 4.10 finally gives

$$\boxed{W_q^{(i)} = \frac{\sum_{k=1}^{5}\frac{\rho_k}{\mu_k}}{(1-\sigma_{i-1})(1-\sigma_i)}} \tag{4.13}$$

Note that 4.13 holds as long as $\sigma_i = \sum_{k=1}^{i}\rho_k \leq 1$.

The total expected queue size can also be obtained from this result using Little´s formula as

$$L_q = \sum_{i=1}^{5}L_q^{(i)} = \sum_{i=1}^{5}\frac{\lambda_i\sum_{k=1}^{5}\frac{\rho_k}{\mu_k}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{4.14}$$

The wait in queue averaged over all customers is

$$W_q = \sum_{i=1}^{5}\frac{\lambda_i W_q^{(i)}}{\lambda} \tag{4.15}$$

4.1.2 Preemptive System with Many Classes

This section modifies the previous non–preemptive model so that patients of a higher priority *preempts* patients of a lower priority in service. Lower priority patients that are ejected from service cannot reenter service until the system is free of all higher priority patients. Generally, for such queues, we must specify how the system handles ejected patients that receive only partial service. Two common assumptions are that ejected patients must start over thereby losing the partial work already completed, or that ejected patients resume service from the point of interruption. Since we assume here that service times are exponential, this issue is irrelevant in view of the Markovian memoryless property. For a Markovian preemptive queue, the system state is completely determined by the number of patients of each class in the system. For non–preemptive queues, we also had to specify the class of the patient in service. Here, the class of the patient in service is always the highest priority class in the system, so the extra parameter is not needed in the state space. Changes to the non–preemptive model include:

$S_j \triangleq$ The random service time (consultation time in minutes) with which priority class j is served, where $j = \{1, \ldots, 5\}$.

$L^{(i)} \triangleq$ The average number of priority class i patients in the system at steady state, where $i = \{1, \ldots, 5\}$.

Then equations 4.1 and 4.2 change to 4.16 and 4.17 respectively.

$$\rho_i = \lambda_i E[S_i] \quad \forall \quad i \tag{4.16}$$

$$\sigma_i = \sum_{j=1}^{i} \rho_j \quad (\sigma_0 \equiv 0, \sigma_5 \equiv \rho) \tag{4.17}$$

In the case of 5 customer classes and general service distributions, where preempted customers resume service from the point of interruption, the results are (replacing 4.14).

$$L^{(i)} = \frac{\rho_i}{1 - \sigma_{i-1}} \quad + \quad \frac{\lambda_i \sum_{j=1}^{i} E[S_j^2]}{2(1 - \sigma_{i-1})(1 - \sigma_i)} \quad \forall \quad i \tag{4.18}$$

The rest of the equations in the previous section still hold and stay unchanged. In conclusion, the preemptive model has no effect on the patients´waiting time in the queue or system. Seeing as we are aiming to reduce *waiting time* and not queue length. This proves not to be a viable solution when modeling the MediTriage system. To strengthen the argument, it should be noted that in reality we cannot simply *pull a patient from service*. Once a doctor is in consultation, for ethical and professional reasons, they should not be interrupted, unless a life–or–death emergency emerges for which no other doctor is idle. This situation, however, has a very small probability of occurring and will be left out of the model for simplistic reasons.

4.1.3 Multiple Servers

The analysis for the multiple-server case is very similar to that of non–preemption except that it must now be assumed that service is governed by identical exponential distributions for each priority at each of the service channels. Unfortunately, for multichannels we must assume no service–time distinction between priorities, or else the mathematics becomes quite intractable. Let us define

$c \triangleq$ The number of doctors available to consult with E/U patients.

$p_n \triangleq$ The probability that channel n is busy, where $n = \{1, \ldots, c\}$.

$$\rho_k = \frac{\lambda_k}{c\mu} \quad \forall \quad k \tag{4.19}$$

$$\sigma_k = \sum_{i=1}^{k} \rho_i \quad (\sigma_5 \equiv \rho = \frac{\lambda}{c\mu}) \tag{4.20}$$

Again the system is stationary for $\sigma_k = \rho \leq 1$

$$W_q^{(i)} = \sum_{k=1}^{i-1} E[S'_k] \quad + \quad \sum_{k=1}^{i} E[S_k] + E[S_0] \tag{4.21}$$

where, as before, S_k is the time required to serve n_k patients of the kth priority in the line ahead of the patient arriving. S'_k is the service time of the n'_k items of priority k which arrive during $W_q^{(i)}$, and S_0 is the amount of time remaining until the next server becomes available. To derive $E[S_0]$, consider

$$E[S_0] = \Pr\{\text{ all channels busy}\} \cdot E[S_0|\text{all channels busy}]$$

The probability that all channels are busy is

$$\sum_{n=c}^{\infty} p_n = p_0 \sum_{n=c}^{\infty} \frac{(cp)^n}{c^{n-c}c!} = p_0 \frac{(cp)^c}{c!(1-\rho)} \tag{4.22}$$

and
$E[S_0|\text{all channels busy}] = \dfrac{1}{c\mu}$

from the memorylessness of the exponential,

$$E[S_0] = \frac{(cp)^c}{c!(1-\rho)(c\mu)} \left(\sum_{n=0}^{c-1} \frac{(cp)^n}{n!} + \frac{(cp)^c}{c!(1-\rho)} \right)^{-1} \tag{4.23}$$

Therefore from 4.13,

$$W_q^{(i)} = \frac{E[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)} = \frac{[c!(1-\rho)(c\mu)\sum_{n=0}^{c-1} \frac{(c\rho)^{(n-c)}}{n!} + c\mu]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{4.24}$$

and the expected line wait taken over all priorities is thus

$$W_q = \sum_{i=1}^{5} \frac{\lambda_i}{\lambda} W_q^{(i)} \tag{4.25}$$

4.2 MediTriage Hybrid Model

We can now use the building blocks from Section 4.1 to form a unique hybrid model that would show to which extent the MediTriage system can save valuable waiting time for patients at Medipark.

The model will be dealt into 4 sections (which together would form the hybrid model). Each of these sections will be iterated for each day of the week and once more for public holidays. The sections are as follows:

1. **Off–Peak queuing model for Triage levels 1–3**
2. **Peak queuing model for Triage levels 1–3**
3. **Nurse queuing model for Triage level 4**
4. **Pharmacy queuing model for Triage level 5**

Limitations:

• It is important to note that MediTriage cannot be implemented during the night–shift (from 23h00–8h00) and we should also rule out using more than one doctor (server). This is due to the fact that we only have one doctor on duty without assistance from a nurse. If we should appoint one more doctor as well as a nurse, this would incur an extra cost of R2500/night to the practice in terms of wages. Now if we take into account that the average arrivals is only a mere 7 patients per night, and the cost of a consultation is set at R320/patient. This would already cause a loss of R260/night even when we leave the utility costs out of the equation thus is clearly an unrealistic option. (Please note that it would not be possible to increase consultation fees in order to make up for this loss, as these are annually set by guidelines given by the Health Professions Council of South Africa (HPCSA)). Also, with 7 patients arriving over a period of 9 hours, the probability of waiting is already very low and needs no improvement.

• Also note that we can only utilise 1 server during *off–peak* times. With the current arrival rate during these times, we run the risk of creating too much idle time for doctors when we appoint more than one to be *on–call* in the same hour. Another problem arises when we appoint another doctor to attend to U/E patients exclusively. By doing this, we are taking away at least 32 scheduled appointment time slots per shift. When this happens, a patient may phone to make a booking but find that there are no more slots available. In this case the patient would have to come in and join the U/E queue, which in turn could cause the U/E arrival rate to increase by at least 32 patients. Again causing queues to explode and waiting times to become unbearable. Please note that this trend is not applicable to *peak* times, as patients finding no open slots at night will tend to reschedule to the following morning.

Verification:

All of the results obtained from the Lingo® model will be verified with the use of the QtsPlus® add in software package for Microsoft Excel®.

QtsPlus® software is a collection of Excel® workbooks designed to solve many common queuing problems. The collection is based on the textbook Fundamentals of Queueing Theory, 3rd Ed. by Gross et al. (2008). The workbooks implement closed–form analytic, numerical and simulation methods. Due to the fact that these workbooks are based on the same logic as defined in Section 3.3.3, the results proved to be an exact match to those found by the Lingo® model. For a merged version of both Lingo® and QtsPlus® results, see Appendix C.

4.2.1 Section 1: Off–Peak queuing model for Triage levels 1–3

This section was modeled based on the logic behind the equations as defined in Section 4.1.1: *Non–preemptive System with Many Classes* . This part of the model will only iterate for Triage levels 1 to 3, due to the fact that both Triage levels 4 and 5 will leave the main queue and join the nurse and pharmacy queues respectively. To recap, this section has many service rates and triage levels, with lower number having higher priority. For a full report on the results of this section, refer to Appendix C2.

4.2.2 Section 2: Peak queuing model for Triage levels 1–3

This section was modeled by adapting the equations from Section 4.1.1: *Non–preemptive System with Many Classes* to make room for more than one server as shown in the *Multiple server* part of section 3.3.3. Once again, we will only iterate for Triage levels 1 to 3, due to the fact that both Triage levels 4 and 5 will leave the main queue and join the nurse and pharmacy queues respectively. To recap, this section has only one service rate but 2 servers and 3 triage levels, with lower numbers having higher priority. For a full report on the results of this section, refer to Appendix C3.

4.2.3 Section 3: Nurse queuing model for Triage level 4

This section takes on the form of a basic Markov queuing model, where patients arrive in a Poisson fashion at one server, where they are served with an exponential service distribution. Lingo® has a built in function called PEB for these types of queues. The following excerpt in Figure4.1 is explained by the parenthesis in green.

```
! Average no. of busy servers;
LOADSr(t) = (Arv_Rate_Peak(t,4)/60) * Mean(4);
! Probability a patient must wait for SR;
PWAITSr(t) = @PEB( LOADSr(t),1);
! Conditional expected wait, i.e., given must wait;
WAITCNDSr(t)  =  Mean(4)/( 1 - LOADSr(t));
! Unconditional expected waiting time in minutes;
WAITSr(t) = PWAITSr(t) * WAITCNDSr(t);
```

Figure 4.1: The nurse queue calculations for Triage level 4 in Lingo®

The QtsPlus® software provides a model with much more detail than the Lingo® capabilities. See Appendix C4 for the full modeling input, computations and results.

4.2.4 Section 4: Pharmacy queuing model for Triage level 5

This section also takes on the form of a basic Markov queuing model, where patients arrive in a Poisson fashion at 2 servers, where they are served with an exponential service distribution. Lingo® has a built in function called @PEB for these types of queues. The following excerpt in Figure4.2 is explained by the parenthesis in green.

```
! Average no. of busy servers- 2servers;
LOADPh(t) = (Arv_Rate_Peak(t,5)/60) * Mean(5);
! Probability that a patient must wait for SR;
PWAITPh(t) = @PEB( LOADPh(t),2);
! Conditional expected wait, i.e., given must wait;
WAITCNDPh(t)  =  Mean(5)/( 2 - LOADPh(t));
! Unconditional expected waiting time in minutes;
WAITPh(t) = PWAITPh(t) * WAITCNDPh(t);
```

Figure 4.2: The pharmacy queue calculations for Triage level 5 in Lingo®

Once more, the QtsPlus® software provides a model with much more detail than the Lingo® capabilities. See Appendix C4 for the full modeling input, computations and results.

The final model incorporates all 4 of these sections into one hybrid model that computes all of the waiting times for each triage level, shift (peak and off–peak) and day. See Appendix C1 for the full Lingo® model.

Chapter 5

Savings Measured

5.1 MediTriage model – waiting time results

The following is a summary of the results obtained by each individual section of the model. In conclusion, Table 5.1 tabulates the total *Off–Peak* waiting times over all triage numbers with the *Peak* totals shown in Table 5.2.

Table 5.1: Total Off–Peak waiting times (minutes/patient)

QUEUE	**Triage 1–3**	**Nurse**	**Pharmacy**	**Total Waiting**
Monday	20.87	1.59	≤ 0.1	**22.47**
Tuesday	12.04	1.04	≤ 0.1	**13.07**
Wednesday	19.27	1.36	≤ 0.1	**20.64**
Thursday	11.02	0.97	≤ 0.1	**11.99**
Friday	13.47	1.11	≤ 0.1	**14.58**
Saturday	29.84	1.69	≤ 0.1	**31.54**
Sunday	48.54	2.05	≤ 0.1	**50.59**
Public Holiday	23.31	1.51	≤ 0.1	**24.82**

Table 5.2: Total Peak waiting times (minutes/patient)

QUEUE	**Triage 1–3**	**Nurse**	**Pharmacy**	**Total Waiting**
Monday	2.42	2.12	≤ 0.1	**4.54**
Tuesday	2.68	2.24	≤ 0.1	**4.92**
Wednesday	1.97	1.89	≤ 0.1	**3.87**
Thursday	1.58	1.69	≤ 0.1	**3.26**
Friday	0.79	1.17	≤ 0.1	**1.97**
Saturday	0.88	1.24	≤ 0.1	**2.12**
Sunday	1.46	1.62	≤ 0.1	**3.08**
Public Holiday	0.02	0.18	≤ 0.1	**0.19**

5.2 MediTriage Savings measured

By comparison of tables 5.1 and 5.2 versus the benchmark values set in table 3.4 the MediTriage system has revealed the following in terms of savings.

Table 5.3: Total savings (minutes/patient)

	Current Scenario	**OFF-PEAK Total Wait**	**Savings**	**PEAK Total Wait**	Savings
Monday	25.38	22.48	2.90	4.54	20.84
Tuesday	30.92	13.08	17.85	4.92	26.00
Wednesday	33.72	20.64	13.08	3.87	29.85
Thursday	19.13	11.99	7.13	3.26	15.86
Friday	14.62	14.58	0.04	1.97	12.65
Saturday	23.50	31.54	-8.04	2.12	21.39
Sunday	49.29	50.59	-1.30	3.08	46.21
Public Holiday	39.04	24.82	14.22	0.20	38.84

From the tables, the difference in the amounts of waiting time saved during the *peak* and *off–peak* times are quite noticeable. The reason for this occurrence is the fact that we can only utilise 1 server (doctor) during *off–peak* times. The appointment of one more doctor would greatly increase the amount of savings possible, but unfortunately this would drive the *server utilisation* factor down to a mere 36%. As a result, these 2 doctors would be idle for 64% of the day.

To recap, from Chapter 1, we are aiming to decrease waiting time without causing an increase in doctor idle time. By trail and error, the best solution to balance waiting time and doctor idle time was found by appointing 1 doctor during *offpeak* times and 2 doctors during *peak* times.

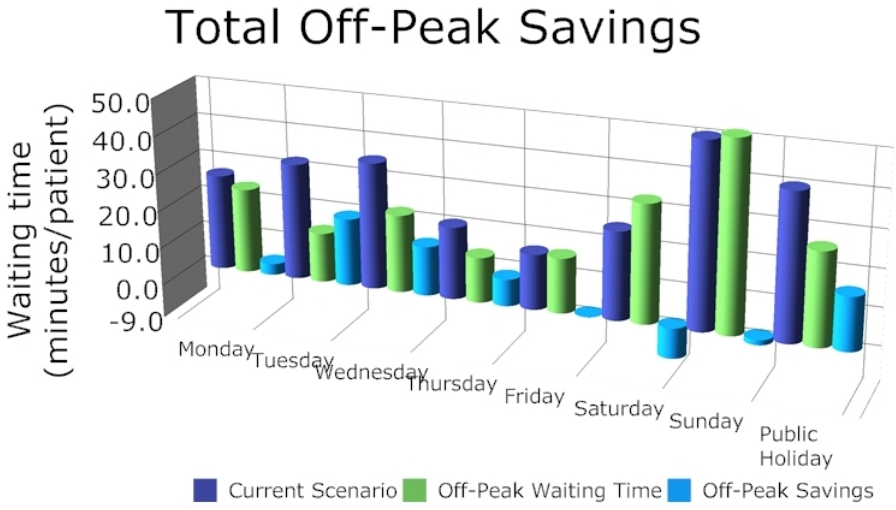


Figure 5.1: Total Off-Peak waiting minutes saved

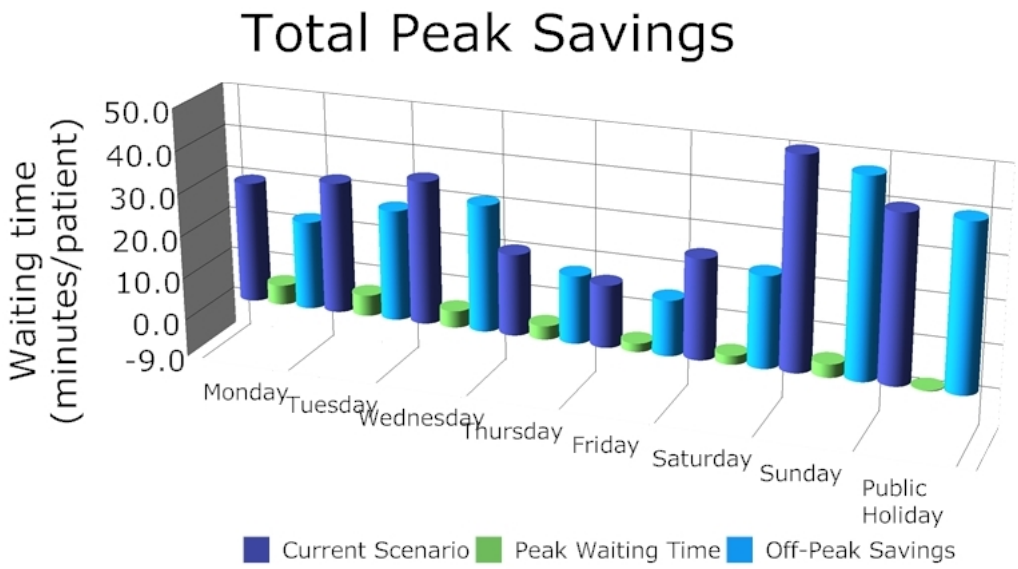


Figure 5.2: Total Peak waiting minutes saved

5.3 Conclusion

In conclusion, the MediTriage system has proved to be a valuable time saving tool. Against an average of 47.5 U/E patients that arrive per day, the saving amounts to 1520 waiting minutes per day. This translates into savings that amount to a total of 32 minutes per patient per day (see Figure 5.3).

Figure 5.3: Total daily savings

Furthermore, Medipark already has the systems in place for implementing a Triage system. The resources needed would be a triage room and a triage nurse. In this regard, there is room available and one of the reception ladies is a retired nurse. She can double as the triage nurse and this would incur no extra costs in terms

of salaries for the practice. The practice manager at *Morningside Mediclinic* has offered to supply the necessary training for triage personnel under the framework set by the HPCSA.

In terms of the monetary value attached to these savings, it could be stated that with each 32 minutes saved, one extra patient can be consulted. Thus the practice is provided the opportunity of consulting an additional 30 patients per day. The current consultation fee is set at R230 per patient, indicating that the savings method proposed could generate an additional income of R6900 per day or R213900 per month. For Medipark, this indicates a 13.4% increase in total consultation revenue. This is most definitely a feasible solution to our original U/E waiting time problem.

Even though the project aim was only to eliminate waiting time, the outcome exceeded expectations in the sense that a significant financial gain can also be achieved.

This study underwrites the integration of triage resources and processes into cohesive strategies, structures and systems for the effective and efficient delivery of services by Medipark.

Bibliography

Abdullah, M. H. (2004). Study on outpatients' waiting time in hukm through the six sigma approach. *Department of Statistics Malaysia*, pages 39–53.

Aickelin, U. and Dowsland, K. (2000). Exploiting problem structure in a genetic algorith approach to a nurse rostering problem. *Journal of Scheduling*, 3(3):139–153.

Bard, J. F. and Purnomo, H. W. (2009). *Cyclic preference scheduling of nurses using a Langrangian-based relaxation heuristic*. Spinger Science and Business Media.

Barnes, C. D. and Laughery, R. K. (1998). Advanced uses for micro saint simulation software. In *Proceedings of the 1998 Winter Simulation Conference, Washington, D.C*, pages 271–274. IEEE.

Benneyan, J. C. (1997). An introduction to using computer simulation in healthcare - a patient wait case study. *Journal of the Society for Health Systems*, 5(3):1–15.

Breyfogle, F. and Salveker, A. (2004). *Lean Six Sigma in Sickness and in Healthcare*. Smarter Solutions, Austin, Texas.

Burke, E. K., de Causmaecker, P., vanden Berghe, G., and van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499.

Causmaecker, P. D. and den Berghe, G. V. (2003). Relaxation of coverage constraints in hospital personnel rostering. In *Practice and Theory of Automated Timetabling*, volume 4, pages 129–147. Springer, Berlin.

Cobham, A. (1954). Priority assignment in waiting line problems. *Operations Research*, 2(3):70–76.

Fomundam, S. and Herrman, J. (2007). A survey of queuing theory applications in healtcare. *The Institute for Systems Reasearch Technical Report*, 24(2):1–22.

Green, L. (2006). Queuing analysis in healtcare: Reducing delay in healthcare delivery. In Hall, R. W., editor, *Patient Flow*, pages 281–308. Springer, New York.

Gross, D., Harris, C. M., Shortle, J. F., and Thompson, J. M. (2008). *Fundamentals of Queuing Theory, 4th ed.* Wiley, Hoboken, New Jersey.

Hall, R. W. (2006). *Patient Flow - The new queuing theory for healthcare.* Lionheart Publishing, USA.

Howell, J. P. (1998). Cyclic scheduling of nursing personnel. *Hospital J.A.H.A*, 40:77–85.

Kao, E. P. and Tung, G. G. (1981). Bed allocation in a public healthcare delivery system. *Management Science*, 27(1):507–520.

Kelton, W. D., Sadowski, R. P., and Sturrock, D. T. (2003). *Simulation with Arena.* McGraw-Hill,New York.

Khurma, N., Bacioiu, G. M., and Pasek, Z. J. (2008). Simulation based verification of lean improvement for emergency room process. In *Proceedings of the 2008 Winter Simulation Conference.*

Miller, D. (2005). Going lean in healtcare. *Institute for Healthcare Improvement.*

Miller, H. E., Pierskalla, W. P., and Rath, G. J. (1976). Nurse scheduling using mathematical programing. *Operations Research*, 24(5):857–870.

Nonobe, K. and Ibaraki, T. (1998). A tabu search approach to the constraint satisfaction problem as a general problem solver. *European Journal of Operations Research*, 106(1):599–623.

Pyzdek, T. (2001). *The six sigma handbook: A complete guide for greenbelts, blackbelts, and managers at all levels.* McGraw-Hill.

Siddhartan, K., Jones, W. J., and Johnson, J. A. (1996). A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance.*, 9(1):10–16.

Spear, S. J. (2005). Fixing healthcare from the inside, today. *Harvard Business Review*, 83(8):78–91.

Stout, W. A. and Tawney, B. (2005). An excel forecasting model to aid in decisionmaking that affects hospital resource/bed utilization. In Bass, E. J., editor, *Proceedings of the 2005 Systems and Information Engineering Design Symposium.*

Tucker, J. B., Barone, J. E., Cecere, J., Blabey, R. G., and Rha, C. (1999). Using queuing theory to determine operating room staffing needs. *Journal of Trauma*, 46(3):71–79.

Womack, J. P. and Jones, D. T. (2003). *Lean Thinking.* Simon and Schuster, London.

Appendices

Appendix A:

A1: Unplanned Activity Chart

A2: Standard Room Stocking Checklist

A3: TSM Scripting for Appointment Scheduling

A1: Unplanned Activity Chart

The Unplanned Activity Chart assists in identifying waits and delays in the process of providing smooth uninterrupted patient care. Each doctor carried the card during a consultation session(shift) and documented when and why patient care is delayed or interrupted.

Unplanned Activities/Indirect Patient "Pulls"

Provider Name: Dr Gerrit _____ Date: 20/06/10 _____

Time: 8:00-14:00 _____

Place a "tic" mark for each incident of an unplanned activity.		Total
▪ Phone Interruptions	𝍸 ///	8
▪ Support Staff Interruptions	///	3
▪ Patients out of order	𝍸 /	6
▪ Late patient arrival	𝍸 𝍸 /	11
▪ Hospital Admissions	/	1
▪ Patient Phone Calls	//	2
▪ More than 1 patient	𝍸 /	6
▪ Missing Equipment	𝍸	5
▪ Missing Supplies	𝍸 //	7
▪ Missing Chart	////	4
▪ Missing test results	//	2
▪ Psychiatric consultation	////	4

A2: Standard Room Stocking Checlist

The Standard Room Stocking Checklist(SRSC) makes sure that all consultation and emergency rooms are adaquetly stocked. This will save valuable consultation time(as shown in the Unplanned Activity Card) and lead to savings in waiting time. The original SRSC is a 12 page document. For relevance the following is only an explanatory excerpt.

A3: TSM Scripting for Appointment Scheduling

Receptionist: 'Which doctor do you regularly see?'
Patient: 'Dr Moore, but it really doesn´t matter to me.'
Receptionist: 'It really is better for you to see the same one as frequently as possible, so that he gets to know you better and can take better care of you. Dr Moore is not in today, but I can schedule you tomorrow with him when he returns.'
Patient: 'I would rather come in today.'
Receptionist: 'That´s fine, you can see one of his colleagues today, and next time we will try to get you in with Dr. Moore.'

or

Patient: 'I would like to make an appointment with Dr Moore.'
Receptionist: 'When would you like to come in?'
Patient: 'Tomorrow sometime'
Receptionist: 'Dr Moore is not in tomorrow. He could see you at 3:00 today, or he will be back in on Thursday and I could schedule you then.' (Patient gets to choose)
or

Patient: 'I would like to make an appointment for next month with Dr Moore for my physical'
Receptionist: 'We really try not to schedule out so far, since plans change and it can be hard to keep an appointment that is scheduled so far in advance. Would you like to come in sooner, or would you like to call back within a few days of when you would like to be seen? We will have appointments available then'
(If patient is insistent and the schedule is open, go ahead and schedule, but make a note for someone to confirm appointment the day before)

or

Receptionist: 'Dr Moore´s schedule is full today and we have already worked in a few emergencies. Since you are requesting a routine physical, I will need to schedule you for another day with Dr Moore. What day is best? '
Patient: '#$%&# !!! You people first tell me something about a Same Day appointment and have asked me to call on the same day, and now that I do, you tell me that I can´t come in today! When are you going to get your #$%&# act together??'
Receptionist: (Pleasant and smiling) 'We are doing the best that we can. We have gotten so busy that we have had to schedule out a few days, but we are working hard to get back to the same day appointments. Remember when you used to call and it took a month to get in? If you really can´t wait, one of Dr Moore´s colleagues can get you in today, but I know that Dr Moore would really like to see you himself, since he knows all about you. He can see you at 8:00 tomorrow and you will be his first patient of the day'

or

Receptionist: 'Dr Moore´s schedule is full today, but you can see him tomorrow morning or one of his colleagues today'

Patient: 'I want to see Dr Moore, but I don´t know what I am doing tomorrow. I want to call back tomorrow.'
Receptionist: 'If that works better for you, that is fine. Try to call as early in the day as you can, since the schedules fill up fast and I can´t guarantee that you will get the time that you want.'

Remember

- It´s the patient´s choice – accommodate them whenever possible

- Always confirm patients´choice provider and schedule with that doctor whenever possible.

- Try not to schedule out any further than 2 weeks, if possible, since the no show rate rises after that length of time

- Anything that you are scheduling for another day, try to encourage the early morning appointments. If the patient insists on a later time, go ahead and schedule (it´s the patient´s choice!)

- If the conversation is getting tense, get the point across to the patient that we want his appointment time to work for him so that he will be sure to make it.

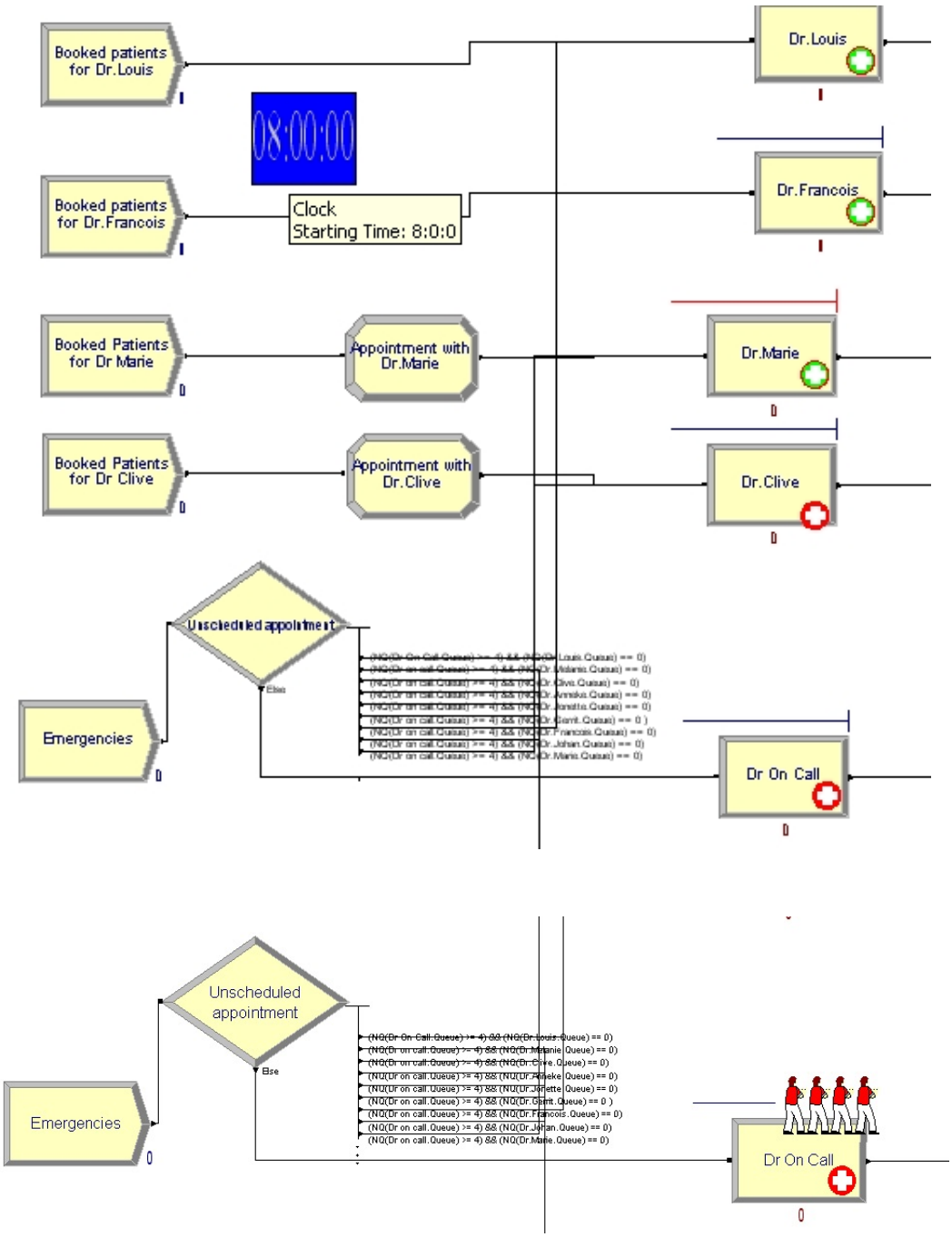
Appendix B:

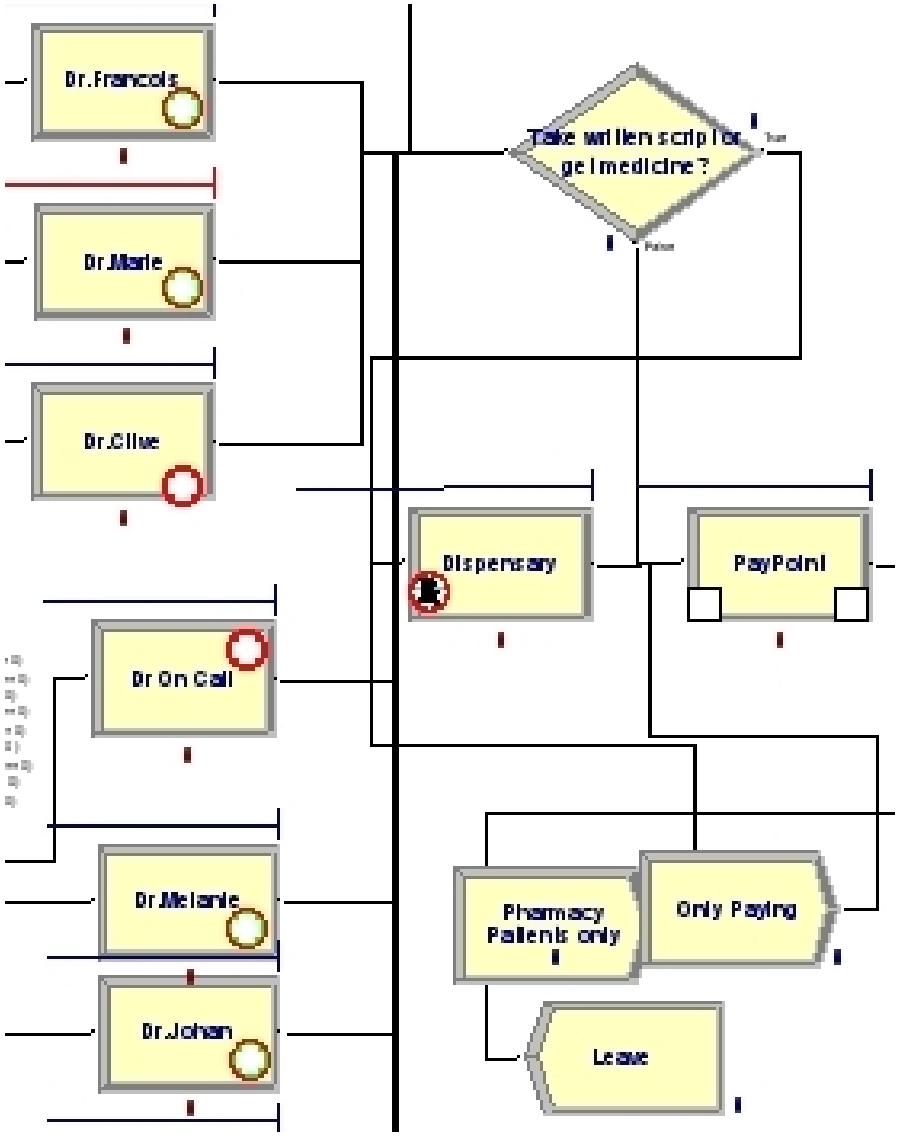
B1: Arena®model

B2: Time Studies - U/E Arrivals

B3: Time Studies - Scheduled Arrivals

B1: Arena® model




Dr.Francois
Dr.Marie
Dr.Clive
Take written script for ge I medicine?
Dispensary
PayPoint
Dr On Call
Dr.Melanie
Dr.Johan
Pharmacy Patients only
Only Paying
Leave

B2: Time Studies - U/E Arrivals

B3: Time Studies - Scheduled Arrivals

Appendix C:

C1: Lingo®model

C2: Off-Peak times – Single–Server, Non-preemptive Markov Model for Triage Levels 1–3

C3: Peak times – 2–Servers, Non–preemptive Markov Model for Triage Levels 1–3

C4: Peak and Off–peak results for the Nurse Queue (Triage level 4)

C5: Peak and Off–peak results for the Pharmacy Queue (Triage level 5)

C1: Lingo®model

C2: Off-Peak – Single–Server, Non–preemptive Markov Model for Triage Levels 1–3

The following pages are a summary of the results obtained by the Lingo®model during *Off–peak* times for Monday through Sunday, including public holidays. The results were verified using the QtsPlus® add in software for Microsoft Excel®.

Off–Peak waiting time (minutes/patient)

Triage level	**1**	**2**	**3**	**Total**
	W_{q1}	W_{q2}	W_{q3}	W_q
Monday	10.23	14.87	27.30	**1.60**
Tuesday	7.47	9.75	14.61	**12.04**
Wednesday	9.81	14.03	24.91	**19.27**
Thursday	7.05	9.07	13.22	**11.02**
Friday	8.00	10.67	16.58	**13.47**
Saturday	12.09	18.96	41.08	**29.84**
Sunday	14.53	25.27	71.59	**48.54**
Public Holiday	10.80	16.07	30.96	**23.31**

Legend: $W_q \triangleq$ **Mean waiting minutes in the queue.**

C3: Peak – 2–Servers, Non–preemptive Markov Model for Triage Levels 1–3

The following pages are a summary of the results obtained by the Lingo® model during Peak times for Monday through Sunday, including public holidays. The results were verified using the QtsPlus® add in software for Microsoft Excel®.

C4: Peak and Off–peak times results for the Nurse Queue – Triage level 4

The following pages are a summary of the results obtained by the Lingo® model for the Nurse queue(Triage level 4) during Monday through Sunday, including public holidays. The results were verified using the QtsPlus® add in software for Microsoft Excel®.

C5: Peak and Off–peak times results for the Pharmacy Queue – Triage level 5

The following pages are a summary of the results obtained by the Lingo® model for the Pharmacy queue(Triage level 5) during Monday through Sunday, including public holidays. The results were verified using the QtsPlus® add in software for Microsoft Excel®.