Lean Thinking and Queue Modelling in Healthcare

By Saeideh Dehghan Nasiri
Supervisor Dr. Dave Worthington

September 2009


**Executive Summary**

In this project we start by studying lean principles from where they originated and the limitations that manufacturing industries have faced in the process of going lean. Then we look at a number of different practical examples of lean approach in the healthcare system. This leads us to our initial observation that for simple cases where patients are not directly involved and changes such as restructuring and data handling are recommended to improve the system, relying only on lean can actually bring value to the system.

For more complicated issues, lean does have simple rules. To examine how efficient these rules will operate on some of the queue issues within the healthcare system, we proceed by designing a simple appointment based clinic to do our own analysis of the lean approach. With the aim to study the effectiveness of lean principles towards reduction of the waiting times within the clinic and see more clearly through other example cases of employing such policies that are mentioned in this report. Two types of patients with different service time characteristics are considered to arrive at this clinic. A simulation model in Microsoft Excel was built for three different appointment schedules. The lean approach in this case will recommend separation of different streams and giving each type the time they need according to their mean service time. A summary of our findings are listed below.

Non-overbooked clinics:

1. Variation rather than the average length of the service times causes the longer waiting times.
2. In the case where service times of both patient categories have equal big variances, streaming or different time slot appointments does not improve the waiting times much.
3. For cases where only one service time has big variation, streaming has a dramatic effect in reduction of the waiting times under the condition that the category with the smaller service time variation is given the priority and is given the first part of the clinic.
4. For service times with equal small variations, just giving each category the appointment time slot based on their own service time average, can have the same effect as streaming, with the benefit of convenience for patients.

Overbooked clinics:

1. For the cases of different variance for the two service time category, streaming under the lean policy does reduce the waiting times considerably.
2. For the cases of equal variance in the service time, just allocating different time slots for the appointment of the two categories of patients, without streaming them, does reduce the waiting times nearly the same as complete streaming.

Although the lean appointment system was recommended for most cases after all, but without modeling one could have easily been mislead in its implementation. Our conclusions in this analysis also makes us realize that in other practical examples that the improvements were reported as a result of using lean principles, a lot of other issues are also present that lean could not resolve. We do see modeling as one of the powerful analytical tools which often need to be used in a lean framework. This seems as a necessity especially in more complex systems to ignore inadequate decisions or confusions which can exist due to lack of analytical support.
Acknowledgement

This work would not have been possible without the great help and support I got from different people.

First and foremost, I am grateful to my supervisor Dr Dave Worthington who dedicated his time and expertise to give me the guidance that made the completion of this work possible.

I would also like to thank my husband and son, Alireza and Shayan for their patience and support throughout this project.
1. Introduction

The cost of medical care is increasing at an unsustainable rate worldwide. While a significant percentage of these cost increases can be attributed to the inevitable factors of the aging population and technological advances the other significant source of healthcare costs can be broadly characterized as unnecessary operational inefficiency. Healthcare professionals have more control over this factor.

1.1. Lean methodology

A variety of process improvement methodologies have been proposed to address the reported inefficiencies in health care delivery. Lean production, also referred to as the Toyota Production System (TPS) is one such method. The management philosophy and tools of lean production come from the manufacturing industry, where they were pioneered by Toyota Motor Corporation, which is viewed as the leader in utilizing these performance improvement methods.

The system of lean production has spread rapidly across the world since its emergence in Japanese manufacturing in the 1960s and 1970s. It represents an approach to production that strives to eliminate waste from the production process by making workers responsible for productivity and product quality. This management method, referred to as just-in-time (JIT) production and quality management, in the 1980s and ‘Lean production’ in the 1990s (Krafcik, 1988), denote the approach of Japanese automobile manufacturers toward several areas of management. The methodologies are applicable in manufacturing, design, supplier relations and sales, which were spelled out in the book _The Machine that Changed the World_ (Womack & Jones, 1990). In the turmoil of the economic recession at the beginning of the 1990s, the message of Womack et al.’s book proved to be powerful and many non-Japanese producers were trying to implement lean management techniques and systems (Oliver, 1992).

The development of lean was little known or understood outside Japan until the 1970's. Britain gained early experience of lean manufacturing from the establishment of Toyota, Nissan and Honda plants in the UK. Nevertheless, until the 1990's it was really only the automotive industry that had adopted lean manufacturing. Since then it has spread into aerospace and general manufacturing, consumer electronics, construction and, more recently, to food manufacturing, meat processing and healthcare (Lean Manufacturing Overview).

The lean methodology has been gaining popularity in the health system over the last years because of the belief behind it, to be able to produce the same output with a fraction of the organizational resources. Medical care is delivered in extraordinarily complex organizations, with thousands of interacting processes, much like the manufacturing industry. Therefore many aspects of the Toyota Production System and other lean tools are seen to be applicable to the processes of delivering care. The Lean philosophy has grown over the past 60 years developing, evolving and adapting to the needs of business. At the core of the lean strategy lay the following five principles (Womack, 2003):
1. Specify what creates value from the customers perspective
2. Identify the value stream – line up activities which contribute value, eliminate those which add no value
3. Create the conditions for value to flow smoothly through the stream
4. Have the customer pull value from the stream
5. Pursue perfection – work on improving the responsiveness of the production system to the customer demand for value

Lean principles hold the promise of reducing or eliminating wasted time, money, and energy in health care, creating a system that is efficient, effective, and truly responsive to the needs of patients. However, in practice they seem to encounter additional problems if not taken cautiously. Some of the disadvantages of a lean system in manufacturing have already been reported in the literature, which is briefly discussed in the next section.

1.2. Limits of Lean

Toyota developed the lean principles in manufacturing, in response to the needs of the post-world war II Japanese auto market. The market at the time had some characteristics which were served by lean perfectly well, helping Toyota and Honda the leading lean manufacturers to reduce their development times to 42 months compared to 65 months for the US and European producers and were able to adopt full model changes every four years. The Japanese auto market was very small but was a rapidly growing demand for different types of car and truck models, which could well be served by the small-lot, just-in-time (JIT) lean principle. This Japanese system contrasted with the use of ‘functional’ departments, where each would hand off work slowly to the next in a sequential manner, rather than in over-lapping phases guided by a project manager. Although this combination of JIT manufacturing and product development skills did help the Japanese to dominate the world industry through the early 1990s, it has also resulted in a new set of problems and some practical limits. Some of the problems that other companies or the environment were faced are listed in the table below.

| Table 1.1. Limitations of “Lean”: Japan in the 1990s (Cuzumano, 1994) |
|------------------------|-----------------------------|
| **Problems**           | **Production**              |
|                        | Urban congestion            |
|                        | Long geographical distances |
|                        | Overseas locations          |
|                        | Stress on suppliers         |
|                        | Too much product variety    |
|                        | Shortage of blue-collar workers |
| **Product Development**| High cost of frequent model replacement |
|                        | High cost of frequent model line expansion |
|                        | Environmental and recycling costs |
|                        | Too much product variety    |
The strategy of ‘Just in time’ would need frequent deliveries of parts to the manufacturing plants which causes traffic to build up. Apart from the environmental hazards, this would add waste\(^1\) somewhere else to the system by people spending time in traffic and in the production line, waiting for parts to arrive. Another limitation of lean is the need for a cooperative and reliable supplier to agree to manufacture components in small lots and deliver frequently to assembly plants. Japanese companies had a problem with finding the reliable suppliers outside Japan, as they dispersed their plants throughout the world. Also they experienced severe shortage of factory labor that had to employ foreign workers. This had additional costs and problems of training these new workers some of whom had no literacy in Japanese. The consequence of using less-skilled foreigners was reduction in the supplier’s productivity by forcing managers to spend more time on inspection and rework.

The application of lean in healthcare does not have as long history as it does in manufacturing. Therefore, reported faults in a lean health system are rare in the literature. However, due to sensitivity of such system where its product is a patient’s life or well being, a close look at these limits in manufacturing is worthwhile and can provide us with an insight on the possible downsides of a lean health system. A further discussion on this area follows in chapter 2 and 5 of this report.

1.3. **Tackling queue issues in the healthcare system**

Any system in which arrivals place demands upon a finite capacity resource may be termed as a queuing system. In case of health care facilities, it can be found wherever patients arrive or demand for the services randomly, such as emergency room, walk in patients in outpatient setting, etc.

Health Care systems need to ration their medical services in different ways, because no nation has the resources to match the insatiable demand for services. Britain’s National Health Service (NHS), which offers patients medical Care that is free at the point of service, practices queue based rationing. People face all sorts of time delays before medical problems are addressed. The NHS has a long and undistinguished history of attempts to manage the gap between demands and supply (Harrison, 2000). Among different strategies for tackling queue and service system problems, queuing models and simulation are well established approaches.

Queues or waiting lines or queuing theory, was first analyzed by A.K. Erlang in 1913 in the context of telephone facilities. The body of knowledge that developed from it after further research and analysis came to be known as “Queuing Theory”. It is extensively practiced or utilized in industrial settings and retail sector operations management, and falls under the purview of decision sciences. The decision problem involved here is balancing the cost of providing services with the costs of customer waiting. The ultimate goal of queuing analysis and its application in healthcare organizations is to minimize different sorts of cost such as capacity

---

\(^1\) Anything or process that does not add value to a product.
and waiting costs to the organization which might be achieved through a primarily study and balance of queues in the system.

Discrete-event simulation is one of the other tools available to health care decision-makers that can assist in this endeavor. Discrete-event simulation is an operational research technique that allows the end user (namely, hospital administrators or clinic managers) to assess the efficiency of existing health care delivery systems, to ask ‘what if?’ questions, and to design new systems. Discrete-event simulation can also be used to forecast the impact of changes in patient flow, to examine resource needs (either in staffing or in physical capacity), or to investigate the complex relationships among the different model variables (for example, rate of arrivals or rate of service). This information allows managers to select management alternatives that can be used to reconfigure existing systems, to improve system performance or design, and to plan new systems.

Even though TPS, or lean, is widely accepted in the management literature as the most efficient production system developed to-date, its applicability in healthcare and other services outside manufacturing is little known (Hines, 2004) and healthcare managers do not know when it can or should not be used. The central purpose of our research is to study use of lean thinking in healthcare, focusing on its impact on the queue issues in this system. Our approach is to critically review literature on use of lean thinking in healthcare, describe some successful and failure example cases of its appliance in healthcare and provide a better understanding as to what extent lean thinking can be applied on queuing cases in the health system.

1.4. Outline of the report

The report will continue with a brief history on lean and an introduction on its principles in chapter 2, which continues with a review of literature on limits of lean in manufacturing. This part is aimed to shed light on some of the already encountered problems of a lean manufacturing system, to guide us towards the potential problems of a lean health system. In chapter 3 we look at some practical examples of lean implementation within the healthcare to have a better understanding of how these principles approach the health queue issues. After looking at a couple of simpler cases which have the potential for rapid improvements, this chapter continues on the study of more complex queuing systems. This part includes a comparison of the simulation and queue modeling approaches with lean approach on similar cases to give a better understanding of significant differences and capabilities of each method. Chapter 4 then builds a hypothetical simplistic case of queues in an appointment based clinic and investigates the scope of lean principles on improvement of such a system. This follows with a discussion on whether lean can be taken as a sufficient management policy or whether complementary methods are needed. In the final chapter we discuss some of the potential limits of lean in the health system, in the light of our discussions in chapter two, three and own analysis in chapter 4. Then a conclusion and summary of our findings on the degrees of applicability of lean principles in the healthcare system, ends the report, in chapter 5.
2. An Overview of lean

The starting point for the emergence of production process plans goes back to 1913 (Lean Enterprise Institute, 2009), when Henry Ford created what he called a ‘Flow Production’ plan. He focused on elimination of waste while developing his mass assembly manufacturing system. Although Ford’s plan made it possible to manage variation in work activities, his approach did not respond well to uncertain, dynamic business conditions.

The problem with Ford’s system was his inability to provide variety. Just after World War II, Kiichiro Toyoda, Taiichi Ohno, and others at Toyota looked at this situation and came up with a series of simple innovations to make it more possible to provide both continuity in process flow and a wide variety in product offerings. They therefore revisited Ford’s original thinking, and invented the Toyota Production System.

Levels of demand in the Post War economy of Japan were low and the focus of mass production plans therefore had little application. The problem with Ford’s system was his inability to provide variety. Taiichi Ohno recognized the scheduling of work should not be driven by production targets but by actual sales and over production should be avoided. Taiichi Ohno, Kiichiro Toyoda and others at Toyota came up with a series of innovations to make it more possible to provide both continuity in process flow and a wide variety in product offerings. They therefore revisited Ford’s original thinking, and invented the Toyota Production System.

Toyota’s development of ideas later became Lean and its thought process was thoroughly described in the book The Machine That Changed the World (1990) by James P. Womack, Daniel Roos, and Daniel T. Jones. In a subsequent volume, Lean Thinking (1996), James P. Womack and Daniel T. Jones distilled these lean principles further to five. As lean thinking spreads to other service areas, leaders are also adapting the tools and principles beyond manufacturing, to logistics and distribution, retail, maintenance, construction and healthcare. The next section provides the list of five lean principles and a discussion around their implementation in the healthcare system.

2.1. Principles of Lean

Lean thinking brings together several strands of process improvement. Lean starts by defining the purpose of the process (value for the customer), then redesigns the process to deliver this value (with minimum wasted time, effort and cost). It then organizes people and organizations to manage this value delivery process. At its heart lay the following five principles (Jones, 2006):

- Specify **value** from the patient’s point of view
- Identify the **value stream** to diagnose and treat the patient and remove wasted steps
- Enable the patient to **flow** smoothly and quickly through every step
- Match capacity with demand so work is done in line with the **pull** of the patient
• Pursue **perfection** through continuous improvement of the value stream

A number of steps are taken for a system to establish these principles (Byrne, 2005). The summary and order of the steps that have also been found common in all the practical examples over study in this project, which we look at in the next chapter, are illustrated in the figure below.

![Figure 2.1. The process of implementing the Lean principles.](image)

The following subsections provide an explanation for the concepts introduced, in the principles within the lean framework.

### 2.2.1 Value and Value Stream

Under Lean, value is defined from the customer’s (patient’s) perspective. Anything that helps treat the patient is value adding. Everything else is waste. Lean eliminates waste and reinvests released resources in value creation.

The first step for implementation of lean is to identify the value streams. A value stream is all the actions (both value-adding and non value-adding) and associated information required to bring a product (patient) through the value-adding process from beginning to end. In hospitals it is a natural tendency to group patients by clinical similarity. The difference with Lean is that it focuses not on similar clinical conditions but similar processes. Similar value streams flow at a similar pace and require similar infrastructure, processes, etc. Once a value stream has been identified, it can be worked on, end-to-end, to remove obstacles and improve the flow.

It is often hard to see value streams. A preliminary high-level map can provide a big-picture overview that allows the value streams to become visible. Ideally, all levels of staff should be involved in drawing this big-picture map, and then they will all see the process end to end.
The next step is to map every action that is currently taken along a particular value stream, whether necessary or unnecessary to get the patient moving through the system from one stage to another. Considering issues like who does what, when, and how long does it take them? What materials or equipment do they need? What information do they use, input or pass on? Mapping the ‘current state’ of the process is likely to reveal absurdities, possibilities for error and confusion, blockages and bottlenecks. This highlights activities and procedures that are not necessary, do not add value or could be redesigned.

The stage is to create an ideal ‘future state’ map which shows what the process should look like if it were working perfectly without non-value adding (waste) activities.

2.2.2. Flow

To ease the flow of the process each patient is worked with, one unit at a time, and passed on for the next step of the process without any delay. A preoccupation of Lean is to identify blockages and obstacles that cause delay, and to remove them. One way for easing the flow is by smoothing demand where it can be smoothed, and by developing the flexibility to cope with variability when it is unavoidable. Also by processing smaller volumes more frequently the system will move from batch and queue towards flow. Below are three different approaches in some practical examples in a health system for moving towards ease of flow.

- An example of this in hospitals is scheduling surgical operations so that instead of conducting many of the same operations on one day and many of another type of operation on the next day (batch & queue), the system will move to a schedule where it does a few of each type of operation each day (More frequent & smaller volume). If such a system was to be practical for the resources, especially the doctors, then has the positive effect of reducing waiting times for patients while also reducing pressure on wards.

- A second kind of example for improving the flow of work in a hospital is the streaming of patients in an Accident and Emergency (A&E) department. This department is a walk in centre where patients arrive without appointment and with wide spectrum of clinical conditions. Many A&E departments prioritize the patients on a triage basis, with the need for more urgent cases to be seen quicker. At the point of arrival, much effort must be made to triage the patients in 4 or more categories. Normally patients in the less urgent categories will be continually pushed down the queue with the arrival of a more urgent case. However, according to the lean principle of identifying the value streams, only two can be identified in the system:
  - patients who can be treated and discharged more or less immediately
  - patients who need to be admitted into a ward for further treatment

The lean system would then separate out, at triage, the two groups of patients, literally placing them in different physical locations and treating them in different ways, with different priority schemes for each group. A comparison for such a scenario is cars on motorway. If slow moving lorries were to travel in both lanes, they would slow down
faster moving cars and cause traffic. Separation of different value streams would enable them to flow according to their own logic and pace, without interference.

The third type of example for easing the flow in a process is to connect different parts in a sufficient way. Some examples of poor connection between different sections in the healthcare system may include the delayed arrival of a test result before a consultant does his round, so treatment will be postponed or the breakdown of liaison with social services and transport services, so the patient isn’t discharged as planned. Any contact needed between members of staff to achieve a task is known as a hand-off and can be a source of potential delay or error. To reduce the number of hand-offs one way is to create standardized processes for the hand-offs between staff so that issues can be identified and addressed as and when they need to be, regardless of who is on duty, on leave or tied up elsewhere.

2.2.3 Pull

To create value without waste, the services need to operate in line with demand. Delivering care in line with demand means not producing it to meet some imposed metric such as a productivity, asset utilization or unit cost target. Delivering services in line with demand also means all work materials and information should be pulled towards the task as and when needed. Any time spent waiting or queuing is another form of waste. To gain a better understanding of this concept, we look at an example of two extreme pull and push systems.

- A ward push for admitted patients for allocating beds to patients, who are to be admitted from the Accident and Emergency department, will operate according to an assessment of the patient’s clinical priority, with urgent cases put in any bed that becomes vacant. This usually managed by developing a complex central bed management role. Bed managers are responsible for pushing patients into wards. This often may lead patients ending up in wards that are not specialized in their illness or injury. Therefore the clinical teams who are responsible for the patient have to spend increasing amounts of time and travelling to different wards to visit their patient.

- A ward pull system, under lean would operate in such a way that the specialist wards pull appropriate patients towards them as and when beds become free. To cope with situations when the best ward is full, the next best wards for each category of patients will be identified. The purpose of this new system is for doctors, nurses and equipment appropriate to the condition to be closer to hand more of the time, meaning less travel and fewer occasions when people or equipment are not available. In reality though still cases occur where patients need to be transferred to wards, not related to their condition but the fraction would certainly be less than the ward push system.

2.2.4 Perfection

Finally, continuous improvements are to be made in the system through, constant observations. While creating clear, easily seen, standardized processes, makes it easier to observe the system and recognize faults and opportunities for improvements.
2.2. Philosophy
To get started on a Lean journey, one needs to consult an expert individual or team in this field who have experience of what to do and how. Yet, for successful changes in any system and Lean to work, the active, enthusiastic cooperation of staff is needed. This will not simply happen by order of the management. One important factor is to involve all levels of staff from porter to consultant surgeon, from ward assistant to top-ranking administrator. While every individual staff member knows more about his or her particular job than anyone else, this helps everybody see how the complete ‘value stream’ works from end to end, and where the waste is.

In the start of a lean journey, after the initial mapping of process and the identification of improvement opportunities which is normally done by involving the existing staff, a rapid improvement event will take place. The real aim for rapid improvement events is to create a culture of continuous improvement.

This involves simple changes in the system such as redesign of work, machine relocation, data accessibility, data input processes and so on. This stage also, helps in encouraging the staff to follow on to next steps. The actual observation of these initial improvements by the staff will also help to create an environment for cooperation and trust which is essential for the implementation of any change in the system.

The system of lean production has spread rapidly across the world since its emergence in Japanese manufacturing. However, there have also been many criticisms by unions and social scientists, pointing to the negative aspects of the ‘Just in Time’ (JIT) production systems and personnel practices. Next section summarizes these critical aspects which are already present in the literature. After a detailed study of lean approach on queue issues in the healthcare system in the next two chapters this overview will then help us to come to our final discussion and conclusion regarding the health system in the last chapter.

2.3. Limits of a lean system
The Lean management policy is usually referred to as ‘management by stress’ by critics of this methodology (Parker, 1988). The Lean production methodology is claimed to possess the ability to cope with the customer demand volatility and heterogeneity due to its flexibility. The flexibility is believed to stem from the commitment and problem solving ability of the workers involved in the production. Frequent model changes and long working hours under lean can put workers under a great deal of stress. Fucini and Fucini (1990) provide a detailed report on the stressful working situations in Mazda’s American transplant. A summary of the limits of lean follows in the next subsections.

2.3.1. Labor Shortage
Production work in manufacturing industries in general (Nuki, 1993), and the car industry in particular, was increasingly perceived as unattractive, leading to severe recruitment problems due to the restricted labor market. Due to low unemployment rates and the high turnover rates
(Shimuzu, 1995), the recruitment problems led to a severe shortage of production workers, and representatives of car producers started complaining about the difficulties in attracting and retaining personnel. The following two approaches have been taken to overcome the labor shortage.

- **Mechanization and Automation**

A labor saving strategy can be pursued by means of mechanization and automation. Automaton, unless it is flexible can lead to high development costs and many Japanese firms had decided to avoid automation in order to maintain competitiveness (Fujimoto, 1992). Toyota introduced automated transfer machinery cautiously and used robots in modest numbers only in 1980s, after they had become programmable, reliable, and inexpensive compared to human workers (Cuzumano, 1994).

- **Increase the Supply of Labor**

Several strategies have been developed to increase the supply of labor in the Japanese manufacturing industry, namely:

1. Establishing new plants in remote areas, where more labor is available.
2. Recruiting more from other segments of the labor market than formerly had been the case.
3. Increasing the attractiveness of jobs by improving the quality of working life (QWL).

These strategies are discussed in turn.

- Most of the Japanese assembly plants and suppliers were located in the center of Honshu Island and Aichi, in central Japan. However in 1990s, Nissan, Isuzu, and even Toyota started building plants on the southern island of Kyushu (Benders, 1996). The policy towards geographical dispersion reduces the possibility of sustaining the JIT system. However this system was strongly criticized because of its contribution to traffic congestion in traditional industrial areas (Kyosei, 1993).

- The subcontracting companies supplying car firms were reported to have been recruiting illegal foreign workers (Smith, 1993) who performed unskilled work. This also, result in reduction of the supplier productivity by forcing managers to reduce work schedules and use more inspection and rework to ensure high quality of their products (Cuzumano, 1994). The larger car manufacturers started recruiting from another labor market segment within Japan, namely women. Also they adopted policies to retain ‘older’ employees longer and additionally were forces to contract many temporarily workers, which led to friction (Shimuzu, 1995).

- QWL can be split up into work content, employment conditions, working conditions and labor relations. In order to improve the working conditions for workers Toyota made a change in its traditional policy of eliminating all forms of waste by splitting up its final
assembly line into four zones, divided by small buffers (Shimuzu, 1995). As, some 'waste' was now seen as necessary for improving working conditions. Other actions taken by Toyota to enhance the working conditions include (Benders, 1996):
- A modification of the line design in order to make cycle times and workloads equal for all work stations.
- The reduction of heavy and demanding work tasks as well as improvements in the physical working environment.
- The maintenance of a smooth and stable work process despite fluctuations in the demand for cars.

The improvements in the field of quality of working life in the 1990s are not unique for Toyota. For instance, Takezawa (1976) mentions Honda’s changed policy with respect to career planning and Nissan’s efforts at helping former workers find jobs as mechanics.

2.4.2. Conditions of Work and Worker Resistance
The multifunctional worker is a precondition for the flexibility of lean production system, as argued by Best (1990). According to Doshe (1985), lean production means diminishing opportunities to slacken the pace, to take a break, etc. Hence, they maintain that there is some intensification of work.

2.4.3. Reduce autonomy of the workforce
Skorstad (1994) refers to reduction of buffers in a lean system and claims that this increases the dependency between the stations on the line and consequently the dependency of individual workers is reduced. Buffers allow for variation in speed, production planning of one’s own, and possibilities for shortened or prolonged breaks according to how fit one feels on the particular day. The elimination of buffers therefore, brings an end to this and places management in a position to control over what workers actually do, reducing the autonomy of individual worker.

2.4.4. Managing the Defects
The product oriented layout and scheduling according to the pull principle are the two main pillars of the lean production when considering its relevance for industry in general (Skorstad, 1994). However, production according to the lean principle will not work if something goes wrong in the production process. For example, buffers cover up deviations from formal plans and their reduction, both in number and individual size, makes it imperative to balance the line. This means that workers must have adequate skill to allow for deployment according to capacity considerations and other needs. Defective parts or components and machinery breakdowns ought not to happen because, if they do, this will reduce the capability of delivering according to demands. Consequently, both quality control and preventive maintenance become more important than before.
2.4.5. High Traffic volume
By the increase in the number of Japanese factories that adopted the Toyota practice in different industries, the traffic in the urban areas worsened to the point where, in the 1990s the Japanese government mounted a media campaign encouraging companies to reduce their frequent deliveries (Cuzumano, 1994). Apart from the environmental pollution as a consequence of the traffic, a new kind of waste would also be created somewhere else in the system, while people are stranded in traffic and in manufacturing plants, waiting for components to arrive. This contradicts with the lean’s ‘waste elimination’ principle.

2.4.6. Supplier Management
Another limitation of the lean manufacturing is the need for a cooperative and reliable supplier. For this system to work, the supplier must agree to manufacture the components in small lots and deliver them frequently to assembly plants, to prevent huge inventories for themselves. This however has become a problem for the Japanese companies as they disperse their plants throughout the world (Cuzumano, 1994).

In the next chapter we look at some practical examples in the healthcare system for which lean has been applied successfully or with failure. Then focus on two particular cases of queues in the Accident and Emergency department and Outpatient clinic, from different angles with the aim to particularly compare and distinguish between the simulation, queuing model and lean approach.
3. **Practical examples of Lean in healthcare**

In the practical context of moving towards a lean healthcare system, after the initial stage of process mapping, some changes toward improvement of the system become apparent for the observer. These usually, do not require further quantitative analysis or modeling and after some rapid improvements great achievements can be obtained towards a more efficient system. On the other hand, some of the issues are not that straight forward and require a more detailed analytical approach to come to an optimal. The principles of lean have brought benefits for some manufacturing industries, but since this strategy on its own lacks a quantitative support, the actual decision boundaries remain doubtful. As argued in the previous chapter, these boundaries have sometimes been found by try and error or after experiencing the unexpected consequences.

In this chapter first we briefly discuss some of the first kind of practical examples of lean implementations on queue issues in the health system, in which rather simple changes do make significant improvements. Then continue by going through two particular kinds of case studies regarding queues in the A&E departments and outpatient clinics. The aim in this part is to compare different approaches of queue modeling, simulation and lean on similar issues and take into account the advantages and disadvantages of each. In some cases we will show that the modeling technique can be considered as a complementary for the lean strategy.

### 3.1. Different applications of Lean

#### 3.1.1. Case 1

This case study was carried out at Bradford pathology department in 2008 (The Spotlight Project, 2008). The purpose of the study was regarding the problem that arose after division of the budget on the blood tests that were ordered from the Oncology and Hematology clinics within that department. The number of tests performed could not meet the target.

The approach was to bring into action a team of consultants trained in an overview of lean. The team started by studying the spreadsheet data on the blood tests from the two departments over the last two years, 2006 and 2007. By analyzing where the tests were ordered from, who ordered them and the types of tests, and using pivot tables\(^2\) and Pareto charts\(^3\) it was understood that due to lack of communication between members of staff and fear of not having the results, a number of unnecessary tests were being ordered. It was decided to provide a guidance form for the junior doctors on the ward. Through discussions with colleagues and searches it was found that some general tests had to be performed on all patients and a certain number of tests, on exception regimes.

---

\(^2\) Pivot tables are a data summarization tool found in data visualization programs such as spreadsheets.

\(^3\) Pareto charts are one of the seven basic tools of quality control and they contain both bars and a line graph. The bars display the values in descending order, and the line graph shows the cumulative totals of each category, left to right.
The outcome was that after introducing these guidelines, the number of blood tests carried out in excess of the proposed guidelines were 140 in the Oncology and 136 in the Hematology department, proving that there was potential to set a higher target. The implementation of the guidance prevents unnecessary tests and also ensures that people know what type of tests is needed and when to request them.

3.1.2. Case 2
This second case study which is related to the flow of patient records in a NHS trust was carried out in three hospital departments in 2007 (Souza, 2008). This example is particularly chosen for presentation here as a case that does not involve the patients directly. The problem of inefficient patient record flows across the trust had been causing delays and cancellation in the clinics prior to the study.

To improve the situation, Lean techniques were chosen to be implemented across the core departments involved in handling medical records which consisted of Medical Records, Medical Secretaries and Main Outpatient Clinics. In the first step an information flow map was developed from the moment a patient presents a health issue to the point of discharge. It was observed that usually 150 procedures were taking place in an average patient treatment with more than 50 handoffs. Having so many steps was seen as one of the obstacles for improvement. The other defect was seen to be poor communication between departments. The consultant team focused on the elimination of wasteful procedures that add cost and increase time to respond to patients. A group of managers, secretaries, clerks and nurses was trained in lean principles and a 7-week implementation followed.

The outcome in each of the departments was positive. In the medical records department, 57 hours per week were saved by streamlining the processes in the clerk area, for example by reduced over processing due to using standardized letters and forms, reduced waiting and over processing for finding files in the library and reduced motion from the clerk’s desk to the printer. Within the medical secretaries, the backlog was reduced by hiring extra staff for a one month period and this together with lean improvements such as standardizing procedures, letters, forms and dictating processes saved a total 8 hour per week for each secretary. For the outpatient clinic, a 4% reduction in the canceled clinics (defects in the patient treatment) and a 30% reduction walking distance for the nurses were achieved. These were made possible by reducing the inventory and making place to store patient records one day earlier and being able to identify missing files and redesign of the layout.

3.1.3. Discussion
In the two case studies, discussed above, the positive outcomes are apparent. However, in the process of lean implementation in these cases and other similar ones that we have come across in this project (Jones, 2006; Mathieson, 2006; Byrne, 2005; Lendon, 2004; Silvester, 2004), one

---

4 The term handoff refers to the process of transferring an ongoing procedure from one person in charge to another.
common challenge is present. Skepticism among practitioners to go along with the changes recommended by lean is the main problem. Part of this resistance would be understandable due to lean being seen as an industrial pilot philosophy, among the health practitioners. In the lean cases discussed earlier and the ones that will follow, the involvement of individuals from all levels, as discussed in the previous chapter would be an important factor in the acceptance of its guidelines within the system and successful implementation. We will look at some of the limitation of lean in chapter 5.

These practical cases and other similar ones that do not involve the patient directly are typical examples of the initial fast improvement made by lean. These are usually done through elimination of wasteful activities by standardization and restructuring of facilities or rooms. Such changes reduce the queues either directly or indirectly while not involving any kind of modeling. In the following sections we aim to focus more complicated examples, where a simple solution with significant improvement, does not become apparent after initial process mapping. These cases which usually affect the patients directly require more analytical investigations. A comparison of the lean approach with the two approaches of simulation and queue modeling on similar cases then will provide us better scope for discussion around the advantages and disadvantages of each procedure.

### 3.2. Accident and Emergency Department

The Accident and Emergency (A&E) departments are ones that people can come to without any barriers and they have yearly demand of around 13 million (Fletcher, 2007) in the UK. Around 80% of attendees are referred to as minor cases and can be dealt with and discharged through the department and 20% are major cases that need to be admitted to an inpatient hospital bed. In the UK there has been much public concern towards waiting times in the National Health Service (NHS). NHS set the plan target that by December 2004, 98% of patients arriving at A&E should be discharged, admitted or transferred within 4h. According to King’s Fund report (2004), two of the public’s top priorities for public healthcare were improved waiting times for patients in A&E departments and for cancer and cardiac patients.

Accident and Emergency departments have to operate many forms of elective treatments and this sets a different and more difficult target for speed and flexibility within them. Many attempts have been made worldwide towards modeling and redesigning and improving different parts of healthcare, including Accident and Emergency departments. Experts and researches have used different techniques in trying to investigate different issues involved within this system, some of which take it as a general matter and some make it specific to a local department. These techniques include simulation models, queuing models, bottleneck analysis and lean. Some of the proposals in each research category have been practically implemented and some have been published. In order to better understand these approaches and make a comparison, in the next section, we look at three different case studies which are typical examples of other research work.
within that approach, on the Accident and Emergency department and point out their similarities and differences.

3.2.1. National Generic Model for the A&E Department

This study was carried out by the operational research analysts within the Department of Health in 2006 with the aim to develop a generic national simulation model for the A&E department to inform the national policy team of the significant barriers to the national targets. Specifically for England, the target was set in 2002 to complete the journey for 98% of attendees within 4h by the year 2004. The model objective was to provide a communication tool representing a typical A&E department, with focus on the following (Fletcher, 2007):

- Show, visually and numerically, how an A&E department works (key outputs being 4-h performance by patient group and overall);
- Build a consensus within DH about the key issues facing A&E;
- Identify ‘quick wins’ to improve A&E;
- Communicate the effects of variability in demand and service provision;
- Show what success and failure look like;
- Run high-level ‘what if’ scenarios to quantify the potential impact of key policy options;
- Direct available DH resources into the correct areas;
- Establish a baseline of performance and issues in a ‘typical’ department to measure changes against.

The model in this study is a discrete event simulation with focusing on the individuals behavior and investigating the ‘what ifs’ quicker. The first step in building the generic model is to understand the flow of patients within the department. This is done through consulting with the nursing and clinical advisors. The patients are categorized to three groups of minor, major and admitted. For simplification purposes, the model takes the three different patient flows, separately although the latter two groups are usually indistinguishable at the early stages of their journey. Another simplification attempt in modeling the patient flows was to consider the process outside the control of A&E department (e.g. diagnostic rests and admission) as capacity unconstrained time distributions.

The model is then collaborated for the patient demand, staffing levels, waits for beds, percentage of patient in each group and etc, based on 7-day survey from 12 trusts across the country, expert opinion and the data from a previously built model for the Oldham hospital. In terms of model validation, as it was developed to match a typical A&E development, there was no real case, for the model to be checked against (Black box validation). However, to make it more realistic, the staffing levels demand level and department size were taken as average, some by comparison against the 7-day survey and some according to the Department of Health statistics. In terms of process times, the data from the Oldham model was used and then this data was adjusted until the model outcome in terms of 4h performance by patient groups matched the 7-day surveys.
Finally, the bottleneck processes times and patient flows in the model were shown to the clinical advisors and A&E consultants to be approved (Open box validation). The outcomes of the model, run for each patient category separately, provide an insight to the 4h performance overall and by patient group, number of patients completed in hourly time intervals and the average time in each process on the patient journey for each group. One of the several issues, investigated by the model, was improving the department’s performance by reducing the minor group’s demand. This was showed to make no significant difference, since most of the A&E delays were due to the factors that were out of the department’s control such as waits for beds and diagnostics.

The model created visual and analytical representation of a typical A&E department with a tool to investigate the ‘what if’ scenarios and was used in the seminars with the national A&E team to discuss patient flows and the impact of potential policies. Modification of this baseline model by an OR analyst can characterize it with any local A&E department. The attempt to demonstrate the model in 10 trusts resulted in 8 of them participating in working with local analysts to gather accurate data on demand, process times and performance levels. Only 3 out of 10 trusts proceeded to the stage of implementing the improvement strategies outlined by the model through running the ‘what if’ scenarios.

The main limitations in this study as pointed out by the authors (Fletcher, 2007) are as follows:

- Data availability: The lack of available data and its poor quality, prevented effective model use.
- Organizational dysfunction: The main reason for delays within the A&E department, were discovered to be related to whole hospital issues such as long admission processes due to lack of capacity. To be able to investigate these issues, a system dynamics model needs to be developed in order to take into account the interaction of different departments with the A&E.
- Lack of motivation: Trusts not being enthusiastic towards cooperating in the study, made it more difficult to find and/or improve data.
- Transferring between patient groups are ignored

### 3.2.2. Application of Lean in Redesigning the A&E Department

This study was carried out in the Emergency Department, Flinders Medical Centre, Australia (King, 2006). This department operates on quite a large scale of around 50000 yearly demands of which 43% require admission. The objective of this study is to describe an application of concepts from Lean Thinking in establishing streams for patient flows in a teaching general hospital Emergency Department (ED). The patients are divided into 5 triage categories, in the Australian Triage Scale (ATS), with triage 1 and 5 allowing for zero and 120 minutes waiting time respectively. Prior to this study, patients were prioritized to be seen based on the ATS.

The first step was to create a detailed schematic representation of how the department actually functioned, emphasizing real life operations rather than intended modes of functioning. This was
accomplished by creating a process map for ED processes that detailed both the movement of patients through the department, and the communication processes that facilitated them. Several sessions were held, involving 15-20 people from a full range of working staff in the ED to obtain an overview of the department processes. Then, the flows of patients were restructured. The triage nurses, who were experienced in allocating an Australasian Triage Scale (ATS) score, were asked to make their initial assessments, record a brief description of the presenting complaint and predict whether in their judgment the patient was likely to be able to return home from the ED after treatment, or was more likely to require admission to hospital. Based on previous studies within the department, the predictions were accurate in 80% of the cases. Patients were then streamed in relation to their predicted outcome, based on the hypothesis that the dischargeable patients require similar care and form one value stream and those in need of admission are a separate value stream. Patients predicted as being likely to be discharged directly from the ED were allocated to the ‘B-side’ team, which was a functional team of nurses and doctors located shift by shift in an allocated group of cubicles in the department. There was a chance for these patients to be transferred to the ‘A-side’ if required. Patients predicted to require admission were allocated to the ‘A-side’ team, a separate team of nurses and doctors located in the A-side cubicles. Triage category 1 patients continued to be taken straight into the designated resuscitation area of the ED. In the absence of a threat to life or severe pain, the patients were to be seen on FIFO basis on the B-side, with the emphasis to complete one patient’s journey as far as possible before starting the next case. On the A-side, the patients were give priority to be seen on a triage basis.

Hospital administrative and clinical data systems were accessed and used for comparisons made between the 12 months before and 12 months after the changes were made with \( \chi^2 \) tests, judging the significance of difference. After streaming, the average number of patients in an hour, in the department was reduced by three to four. The average time that all patients spent in the department was reduced from 5.8h to 5h (King, 2006). Also comparisons were made on the average and earliest waiting time to see a doctor, see a nurse or a doctor and percentages of patients meeting the ATS targets within each triage category. The streaming process in this way seems to be more beneficiary for the ATS category 4 and 5 patients. The improvements were highly due to the fact that despite some increase in the waiting times for the ATS categories of 2 and 3, a more significant decrease was observed in the waiting times for the ATS categories of 4 and 5. The data on the actual number of deaths within the department showed no evidence of a decline in the safety of the care provided.

The changes were supported at the most senior levels in the department. They were introduced on one specific day and have continued 24 h per day, 7 days per week, since then. The nursing and medical staffs were given very brief training on the concepts underlying Lean Thinking and streaming prior to its introduction. The FIFO discipline and seeing potentially dischargeable patients in order of arrival also reduced the sense of frustration among patients.
### 3.2.3. Queuing Network Model of Patient Flow in an A&E Department

In this study, a multiclass Markovian queuing network model of patient flow in the Accident and Emergency department of a major London hospital is developed (Au-Yeung, 2006). The model is parameterized using real data on arrivals, waiting times and mean service times and patient routing probabilities. The objective is to provide some insights into the effects of prioritizing different classes of patients in the department, using a discrete event simulation model.

The patient journey in this department is quite similar to the Flinders Medical Centre. The patients are streamed into Minor, Major and Blue Call groups. This is done either at the reception or the patients are sent for an assessment. Each stream forms a separate queue and goes through different flows with probabilities of transferring to the other groups along the way, until the point of discharge. The simulation model runs for different priorities of patient groups in their shared resources such as lab tests, radiology and other specialists, solving for moments and probability density functions of patient response time for three scenarios of No priority, Majors Priority and Minors Priority. The results reveal that prioritization of treatment for minors over majors can lead to reduction of the mean response times (and corresponding variances) for walk-in arrivals while not affecting the corresponding result for the other two arrivals of ambulance and Blue Call.

This is a particularly interesting result in light of UK government waiting time targets, which encourage the prioritization of minors. In fact comparison of the first two moments of patient response time for the simulated results with the real data over three consecutive years from 2002 to 2005 for three different arrivals of Walk-In, Ambulance and Blue Calls does suggest moving from the major’s priority to a minor’s priority. The limitations with this model are listed as follows:

- The model does not include constraints put on the A&E unit from other parts of the hospital.
- Limiting the resources by allocating specific tasks to members of staff, where in reality, they can be multifunctional.
- Replacing variable staffing level and patient arrivals by averages.
- Taking less specialist types into the model than actually are involved.

### 3.2.4. Discussion

The three cases discussed in previous sections, refer to the study of Accident and Emergency Department with the objective of providing an insight to the processes involved and aim to improve the situation, in terms of reducing the patient’s waiting times.

The Lean approach focuses on the redesign of the department and the main part of it is mapping and gathering data on the current situation, involving members of staff from all levels. This is the first step in every one of the cases. However, Lean approach does emphasis on it and its strategy is to involve all staff levels. This gives the feeling of having a power and responsibility to the
staff members so they see themselves as part of the research team rather than feeling to be modeled. Therefore, under the Lean strategy, the collaboration, motivation and excitement of the staff is usually reported, something which was specifically complained about in one of the modeling cases (Fletcher, 2007).

Lean provides a less complicated tool for the redesign of the system, without the need for modeling. In the Flinders Medical center, based on the ease of flow principle, the arrivals were streamed into minor and major cases. This was also implemented in the north London hospital prior to the Network Queue study but is not discussed whether Lean had a role in it or not. Walley (2003) points out that allocation of priorities beyond a simple minor and major category, disrupts the flow of patients through the system and uses resources in non-value adding task. The evidence for this conclusion is not clear, since if it can be reached based on facts and through the process mapping, then has the advantage of saving a time and effort trying to model different streaming options. Within this system, though next stage of improvements can only be investigated through modeling the system. Issues like differing staff levels and cubical capacities for each stream or sharing some resources and priorities need to be modeled for an optimal setting.

One main limitation which was apparent in all cases, is neglecting the interaction of the A&E department with other hospital units (e.g. waiting for beds, admissions and diagnostics). Although, it seems to be one of the most problematic in causing delays in the department (Fletcher, 2007), but makes the models too complex.

3.3. Outpatient Clinic
Healthcare resources are in great demand. Patients often have to wait long periods of time before being seen by physicians or other medical providers. However, there has been much concern over the length of time that patients wait for service within different departments. An effective appointment system is a critical component in controlling patient waiting times within clinic sessions and reducing the response time from the point of patient’s first contact to the service. The next section provides different approaches that have been taken for appointment scheduling improvements and follows by a brief discussion on the advantages and disadvantages of each study.

3.3.1. Simulation modeling for appointment scheduling
This study describes the development and use of a simulation model of an Ear, Nose and Throat (ENT) outpatient department (Harper, 2003). This case is particularly chosen here as an example of simulation modeling, because it does investigate 9 different scheduling policies, including block bookings to save the doctor’s time, reserving buffer periods with no booking and 7 more. Some of these policies are implemented in the health system and have endorsement in the literature. Although the data collected in this study is from one specific outpatient department,
still can be typical of similar departments. The simulation allows various appointment schedules to be examined and their effects on the clinic evaluated and shows alternative appointment schedules can effectively reduce patient waiting times, without the need for extra resources.

The ENT department under study, serves a population of 0.5 million, sees over 20,000 patients each year and has 4 consultants, 6.5 full time equivalent nursing staff and 6.3 administration and clerical staff. There are 22 different clinic sessions that take place within the department each month, with 10 clinics per week (Monday to Friday, morning and afternoon). Current schedules for each of the 22 sessions had evolved over many years, and were felt by staff to be inefficient and ineffective.

The first step was observing patient flows, data collection and discussions with the clinic staff. One week was spent on data collection, and for nine out of ten clinics in the week, necessary data regarding the patient flows, waiting times and length of consultations with the clinic staff, was collected. Currently, a computerized and manual appointment system is being used. Patient Management System is in charge of controlling the number of available appointment slots on a computerized system. However, a manual diary is also kept which can include up to another fifteen patients and bookings are entered into the diary, by hand, when it is necessary for that patient to be seen. It also became apparent that there are a number of different types of patients attending the clinic with varying health care needs and care pathways through the clinic. These include follow up, new patients, casualties, emergency referrals and ward discharges. Casualty patients are seen by a senior house officer and all other types of patients are seen by consultants, registrars or specialists.

A study of the consultation times according to patient type revealed that the service times do vary for different types of patients. A two-way analysis of variance proved that patient type was significant at the 5% level. Therefore, to reflect real-life service time, probability distribution functions were fitted to each patient type by service activity and individual service times, depending on patient type, were sampled from the corresponding distribution. The model also takes into account the number of people who do not attend for their appointment. Monitoring the arrival times of consultants indicated their lateness for the start of the session (due to services due in other parts of hospital) whereas patients on average arrive 8 minutes early.

A Simul8 model was then built for each clinic session in the week with each model being validated using a black box and open box approach. In particular the predicted percentage of patients who wait for more than 30 minutes (target time set by the Department of Health) for their first service were compared with the observed data and statistical t-tests showed that the weekly sessions were statistically valid. The model then runs 40 runs for a selected clinic as a control model, which operates 2pm-5pm on Tuesday for nine alternative scheduling scenarios. The scenarios include: specifying different blocks for different patient types, reserving a buffer period with no booking, Block booking, on time start of the clinics with the hypothesis of the
consultant being on time and use of the whole clinic time without any buffers or overbooked time slots.

The improvements are based on three performance measure of Average time to 1st Service, % of patients who wait for > 30 minutes for their 1st service and Average time spent in the clinic. T-tests at the 5% level of significance are performed to test for significance of differences. For all three performance measures, the consultant’s punctuation does bring the most reduction in the waiting compared with the control model. The second effective factor proves to be the scheduling of appointments throughout the whole clinic time without any block booking. In fact the scenario with overbooked clinics is the only one that has a significant increase in all the measures compared with the control schedule.

### 3.3.2. Lean technique for improvement of patient scheduling process


Prior to this study, scheduling patient appointments could take as little as 5 minutes on the phone and up to 36 days for some patients. Appointment requests arrive to the reception through mail, phone, fax and email. In order to book the patient for the appropriate medical provider, medical records and images were sent to the physicians to be reviewed and following their advice, the patient were booked and a notification of their appointment was sent to them by mail.

The first step for improvement of the system under lean approach is to identify value for the customers. It was determined that patients and referring physicians value prompt access to appropriate medical providers, while medical providers value an efficient system that accurately triages patients to their clinics. Then a current state value stream map (CS VSM) was created, capturing patient and insurance information, previous medical records, test results, and imaging films that were collected by the clinic coordinator before an appointment was scheduled. Also the CS VSM identified who was responsible for each step, the technology used in the process, how the process flowed from one step to the next, the process time, and the wait times within and in between each step. The mapping process allowed everyone involved to visualize the big picture of the process for the first time. The CS VSM was then analyzed by a process improvement team, which included representatives of all stakeholders in the process, with the goal of scheduling at least 90% of the patients during the first phone contact. A future state VSM (FS VSM) was developed to visualize the new process. A major change in the new system was giving the autonomy to the schedulers and the clinic coordinator to schedule patients without seeking other approvals and only telephone referrals were accepted. In order to help the call
center assign patients to the correct providers, the clinic coordinator and physicians designed algorithms based on clinical problems that the schedulers were likely to encounter. These algorithms were used by the call center to identify provider preferences and the types of injuries and conditions they treat. As expected the algorithms could not match unique injuries and new cases, but during the first month after implementation of the lean process, (75%) patient requests for appointments were managed with one phone call. This process was also more valuable for the referring physicians as it reduced their paperwork and the need to make repetitive calls due to scheduling delays for appointment requests on behalf of their patients.

3.3.3. Discussion

The first study on simulation modeling for the appointment scheduling in the outpatient department provides analytical evidence in support of scheduling different categories of patients regarding their own required service time, with no buffers or block appointments in the system. The performance measure here is set with regard to patient’s waiting times. The same conclusion could have been reached under lean approach. As streaming of different patient categories and allowing each one to process with their own speed lies within lean principles. Also the just in time principle would have recommended elimination of buffers, block bookings and also would have pointed out on the delayed start of the clinics to be avoided.

The same recommendation therefore would have been given, under lean without the need for modeling. This would be due to the setting of performance measures in line with lean, in that reducing waiting times is the final goal rather than the doctor’s utilization or increasing the number of appointments. As with the latter measures, the recommendation would have also included block booking of patients in the beginning of the clinic (Worthington, 2005).

A potential improvement factor, as referred to under the study, is to deliver the clinics on time. The lateness of consultants is the reason for delayed start and it is due to the consultant being held on for an emergency call or for a ward visit prior to the clinic. Although a simulation model can point out the problem but in order to overcome the issue a soft management approach could be more effective. Collection of data and process maps on where, how often and how long the doctor is required for service before the clinic might lead to better designs for doctor’s schedule.

In the case of appointment scheduling for the outpatient departments or other departments with typical characteristics, it can be concluded that were the health authorities willing to set the performance measures based on the patient’s waiting times, the lean approach could be implemented to improve the process. It also has the advantage of simplicity and with the soft approach has also the capability of finding the cause and effects for delays and focus on redesigns. However, the change in performance levels back to doctor’s utilization would lead to loss of lean’s power and other operational research methods would need to be employed for the better designs of the system.
In the next chapter, we investigate to what level the lean approach can be trusted in terms of improving the patient’s waiting times in an appointment based clinic. The case study on the outpatient clinic with the lean approach, as discussed in the previous section, tackles the same problem by focusing on the waiting times prior to the patient’s arrival at the clinic. The point of interest in that study is reduction of waiting from the point that the patient contacts for the appointment.

However, patients who are given appointments, still face queues after attending their appointment. Our interest is to study the lean approach in dealing with queues within the clinic. Although the A&E department was also of our interest and is an issue worth further research, but for the scope of this project, taking into account the time limitations, modeling of the queues within an appointment based clinic is more feasible. This case is also less influenced by interaction between different departments which is something that requires more direct observation on the real system to be modeled. Our approach is to build a simulation model for a simple hypothesized clinic and study different scheduling policies, specifically ones that could be recommended by a lean consultant. This case will also try to answer the question of whether lean principles on their own could bring much value to the system or will there be additional modeling required.
4. Lean system experiment design

Operational research applies scientific techniques to complex problems in the management of large systems of machines, men or materials. The approach relies on the technique of modeling which allows predictions and comparisons of outcomes of alternative decisions. In Britain in the 1970s, operation research gained wide application in the planning of health services (Duncan, 1978). Besides a major role in planning services, operational research techniques still have many potential applications, particularly in service organization.

Greater attention is now being focused on the effectiveness and efficiency of outpatient clinics with particular emphasis on the delivery of care, for example, shorter waiting times for appointments and shorter waiting times in clinic. More recently, inefficiencies in outpatient clinics in the British health service have been blamed on consultant practices of patient "recycling" in clinic which reduced the ability to see new patients (Armstrong, 1995). This suggestion excited vigorous correspondence which included proposals aimed at improving clinic efficiency (Roberts, 1995). Edwards (1994) describes how changes in patient processing in clinic can reduce waiting time and improve clinic efficiency. Wojtys (2009) also provides an example of reduction in waiting times prior to patient’s arrival at the clinic.

Attempts to improve patient waiting time by adjusting appointment schedules have been reported previously, but these studies described highly specific clinic settings (Jennings, 1991) and (Marshall, 1986)). This part of the report aims to investigate to what extent can lean be effective in appointment scheduling to reduce the patient’s waiting times within the clinic. In light of lean principles, value streams need to be separated and let to flow with their own speed. This, in an appointment system would mean different patient types must be separated and given different time slots based on their own requirements.

In order to investigate if lean can improve the waiting times within the clinic and if so to what extent, we consider a rather simplified case of one server with two customer types. In practice there might be more than one patient type and more than one server. However, we believe the general results of our experiment model will still hold for such clinics. A simulation model is built to run for different appointment schedules based on lean and two alternatives, to compare the average waiting times for several service time distributions. The following sections describe the objective and structure of the model and the conclusions.

4.1. The designed experiment

The main objective of the model is to investigate the effectiveness of allocating separate time slots for the two patient categories in reducing their overall average waiting times. We choose to build the simulation model in Excel, because it is much easier to manipulate the arrivals for different appointment schedules, compared to the Micro saint simulation software. Also Excel can serve for our purposes and while its open surface enables high interaction for the modeler,
makes it clearer to check through each step of the program and less need for complex validation methods. The simulation model, built in Excel, runs for different appointment schedules for two groups of patients. The main objective is to investigate when and to what extent can streaming of patients for clinic appointments be effective in reducing the overall average waiting times of the patients. The model operates with one server (consultant) for two different patient characteristics. The arrival rates of the two categories is based on an appointment schedule and the proportion of patients in two categories can either be the same or different. The performance targets include the average waiting times of the patients, the number of patients who have to wait more than 30 minutes and the doctor’s utilization. An idealistic case with zero non attendees and non overbooked clinics is studied first to take into account other factors in detail such as length of the clinic, different arrival rates and priorities of the two patient types. Then we look at overbooked clinics with non attendance rates.

4.1.1. Non overbooked clinics
In this section appointments are spread out through the whole clinic time, with no overbooking of patients.

4.1.1. a. Model parameters

- **Arrivals**
  
  In the base schedule for comparison is the random appointment schedule, in which each category calls in for an appointment randomly and everybody is given the same time slot based on a weighted average service time of the two categories. The weights are related to the percentages of each category. This base model refers to the way, most appointment schedules would operate prior to lean. Two alternative schedules which are based on the lean principle of ease of flow and separation of value streams are then modeled to compare the performance targets with the base schedule. The first alternative, gives appointments on a FIFO basis, specifying a time slot based on that patient’s category average service time. The second alternative schedule separates the two categories completely, for example giving the first part of the clinic time to category one and the second part to category two, giving each patient the time they require according to their category’s average service time. In order to be clear which schedule we are referring to, through the rest of the report the following abbreviations are defined respectively:
  
  - Base
  - Semi lean
  - Lean

  This is not to say that lean strategy would have recommended the third one, as it is not clear whether these two categories of patients would count as two different value streams. But the semi lean schedule is completely in line with the lean principles. Since the third type of streaming has been implemented in other parts of health system (i.e. A&E) as recommended by lean consultants, we do consider it as an alternative to study here. In all three appointment
systems, the whole clinic time is used and no free period is reserved. In all cases the clinic runs for 3 hours, except one scenario where we look at the difference in longer clinics and extend the operation time to 6 hours. Figure 1 shows an example of the appointment system for the three schedules, in a case where average service time patients of category 1 and 2 are 10 and 5 minutes respectively. The time slots for the base model are based on the weighted average of the two service times, which in this case is 7.5 minutes. For the semi lean schedule, time slots of 5 and 10 minutes are given, randomly with the same proportion. For the lean schedule, the two categories are divided to two parts of the clinic, with the first part having 10 minutes time slot for the first category and the second part, 5 minutes for the second category.

Figure 1. Part of the three appointment schedules for $\mu_1=10$ and $\mu_2=5$ minutes, with equal proportions.

<table>
<thead>
<tr>
<th>Base</th>
<th>0</th>
<th>7.5</th>
<th>15</th>
<th>22.5</th>
<th>30</th>
<th>37.5</th>
<th>45</th>
<th>52.5</th>
<th>60</th>
<th>67.5</th>
<th>75</th>
<th>82.5</th>
<th>90</th>
<th>97.5</th>
<th>105</th>
<th>112.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi lean</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>105</td>
</tr>
<tr>
<td>Lean</td>
<td>0</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td>120</td>
<td>125</td>
<td>130</td>
<td>135</td>
</tr>
</tbody>
</table>

The model runs for 500 iterations and T-tests are performed to test for significance of difference in the average waiting times for different schedules. The following two idealistic assumptions are made which make the model simpler and have no effect on our comparison scenarios.

- **Assumptions**

1. Every patient arrives on time
2. Zero non attendance rates for both patient categories

In practice patients usually arrive in time and they tend to be early rather than late. Therefore our first assumption is not far from reality. The non attendance of patients for their appointments however is an issue of concern. Although it could even decrease the waiting times for patients, slightly but reduces the utilization of resources. In order to improve the utilizations, an overbooking of patients usually takes place in appointment systems which does effect patient’s waiting times. Therefore we also investigate the three schedules in an overbooked clinic in the next section.

- **Performance measures**

The performance targets which we look at in this section are the average waiting times of the patients and the number of patients who have to wait more than 30 minutes (this is the time often referred to in practice measures). Since waiting times are viewed as a source of waste for the patients, are given the highest priority to be minimized under the lean strategy as a
performance measure. In the next section we also look at another measure of the doctor’s utilization.

➢ Scenarios

The following three scenarios will be investigated for several service time distributions.

1. Equal patient arrival rates for both categories
2. Different patient arrival rates for both categories
3. Extending the clinic time

4.1.1. b. Verification and Validation of the model

The figure below shows an activity cycle diagram (ACD) which represents the scope of the simulation model used in this study. It illustrates the interaction between the two classes of entity in the model, namely patients and doctors.

A patient arrives in the clinic according to an appointment schedule. Three different schedules are under investigation, in this model which we discussed in the previous section. Then the patient enters a queue where they await to be seen by a doctor (service). This activity is dependent on a doctor being free (i.e. not engaged in another activity or elsewhere). If a doctor is free, the patient is seen and then leaves the system. Such a system, while serving for our purpose of study is a simplified version of a real appointment based clinic. For further details on the model, see appendix 1.

As like any other simulation model, this model needs to be validated to confirm that it describes correctly its intended real world process under study. As our case study, is not built specifically to match a real clinic, the black box approach or comparison against real data cannot be done here. However, as part of our open box verification, we aim to go through every stage of the model and check for the right presentation of inputs, distributions and model process.
One of the advantages of building the model in Microsoft Excel is that it enables the modeller to see through every step of the model, clearly. Therefore, apart from the distributions for the service times, every other line of the model can be checked manually. A schematic figure of the model is illustrated in appendix 1, together with the validation process for the service time distributions. The first three line of the model which is arrivals of patient for different schedules is defined with fixed time slots for the base model, two alternative time slot for the semi lean schedule which is distinguished by the program through use of random numbers. For example for the case where 50% of patients require 5 minutes of service time on average and 50% 10 minutes, the total number of patients in a 180 minute clinic must add up to 24. To validate this random generation of appointment slots, we count the total number of patients in every clinic as a @risk output, the average is as expected equal to 24 (distribution is given in appendix 1). For the base and lean schedule the total number would be fixed on 24.

The next part of the model is the service times which are sampled from Gamma distributions. Gamma distributions are a two parameter family of continuous probability distributions representing the sum of $k$ independent exponentially distributed random variables, with $k$ being one of the parameters of the distribution. Service times are best seen to be represented by such a distribution, in our model as they are positive random variables with variations around an average value.

To assure the service times are correctly sampled, we collect data on a randomly chosen cell for the 500 clinic runs and run a T-test on a Gamma distribution with a same mean and variance to test for the significance of difference between their mean. The p-values for the test confirm the hypothesis of no difference in means to the 5% confidence level. An example of the distributions and the statistical test are given in appendix 1. The system times, visit times and waiting times are calculated by summing and differencing the previous cells and for three random cells, these are checked manually. The step by step verification of the simulation model has convinced us of its proper representation of the symbolic reality.

4.1.1. Results

1. Equal proportions of the two categories in a 3h clinic

Assuming equal arrival rates for both patient categories, the average waiting time ($\bar{W}$) for all patients, is given in table 1, for different service time distributions. The notation $(\mu_1, \sigma_1)(\mu_2, \sigma_2)$, in the table, refers to service times with Gamma distribution for both patient categories with average service time of $\mu_1$ with standard deviation of $\sigma_1$ for patients in category 1 and average service time of $\mu_2$ with standard deviation of $\sigma_2$ for patients in category 2.
### Table 4.1. Average waiting times (minutes) for different service time characteristics under different appointment schedules for equal arrival rates in a 3h clinic.

<table>
<thead>
<tr>
<th>Service time</th>
<th>$\bar{W}$ for Base schedule</th>
<th>$\bar{W}$ for Semi lean schedule</th>
<th>$\bar{W}$ for Lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,2)(10,8)</td>
<td>-</td>
<td>6.30</td>
<td>3.87</td>
</tr>
<tr>
<td>(10,8)(5,8)</td>
<td>9.42</td>
<td>9.72</td>
<td>9.38</td>
</tr>
<tr>
<td>(10,2)(5,2)</td>
<td>4.01</td>
<td>2.43</td>
<td>2.55</td>
</tr>
<tr>
<td>(10,2)(5,8)</td>
<td>7.62</td>
<td>6.77</td>
<td>4.38</td>
</tr>
<tr>
<td>(10,8)(5,2)</td>
<td>8.56</td>
<td>6.91</td>
<td>8.68</td>
</tr>
</tbody>
</table>

For the service times in the first row of table 1 the base appointment system performs exactly the same as the semi lean because the average service times are equal. Therefore in a base model the time slots would be constant and 10 minutes which would be the same as the slots for the other two. In the lean system, the service times for the two categories are separated but again there would be no difference in the service time arrangements for the base and semi lean system. For this reason the first cell in the table is left blank to remind that there is no comparison in this case, between base and semi lean, only these two with the lean can be compared in this case. This argument holds for all upcoming scenarios.

The results in table 1 for this run of experiments indicate the following:

- The maximum waiting times are related to the most variant service times which are in the second row of the table. In this case, although the waiting times for the three appointment systems are significantly different statistically (T-tests, at 5% confidence level) but practically neither of these make much of a difference.

- The first and fourth row of the table show considerable advantage of the lean system over the other two systems. These two rows represent the case where there is much difference in the variances of the service times, either with the same average or different averages.

- The third row of the table, where the average service times are different but the variations for the two categories are the same, proves the advantage of the semi lean system in terms of minimum waiting times. This suggests that the complete streaming of the categories and distinction between the value streams of patients could be determined regarding the differences in their service time variation rather than their averages.

- The fifth row of the table brings an interesting outcome, in that although both the average and the variances in the service times are different, but streaming of the two categories would result in average higher waiting times compared to both the base and semi lean systems. As this seemed a strange result, the separate waiting times for each category were recorded and surprisingly, the average waiting times for category two with shorter service time were significantly higher than category one.
We are not able to explain the reason behind it, at this point. But it does bring the thought that which group is defined as category 1 and is given the first part of the clinic might actually affect the lean policy results. In order to investigate this issue we run the model by exchanging the two categories of patients. This result is shown in table 2.

<table>
<thead>
<tr>
<th>Service time</th>
<th>Base schedule</th>
<th>Semi lean schedule</th>
<th>Lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5,8)(10,8)</td>
<td>10.35</td>
<td>9.10</td>
<td>9.84</td>
</tr>
<tr>
<td>(5,2)(10,2)</td>
<td>4.10</td>
<td>2.51</td>
<td>2.52</td>
</tr>
<tr>
<td>(5,8)(10,2)</td>
<td>7.54</td>
<td>6.50</td>
<td>9.16</td>
</tr>
<tr>
<td>(5,2)(10,8)</td>
<td>7.46</td>
<td>6.83</td>
<td>4.53</td>
</tr>
</tbody>
</table>

Comparison of the results in corresponding cells of table 1 and 2 for the (10,8)(5,2), (5,2)(10,8) and (5,8)(10,2), (10,2)(5,8) proves that the service time with the highest variance creates the highest waiting times and in case of streaming, this group must be selected as category two for the second part of the clinic. As for the (5,2)(10,8) case, lean policy improves the waiting times to a great extent in comparison with the base schedule.

The other performance measure usually referred to in health systems in terms of the waiting times, is the number of patients who have to wait more than a certain time which usually in practice is set at 30 minutes. We aim to compare the three scheduling systems in terms of this measure as well. The results for the different scenarios are given in the table below.

<table>
<thead>
<tr>
<th>Service time</th>
<th>Base schedule</th>
<th>Semi lean schedule</th>
<th>Lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,2)(10,8)</td>
<td>-</td>
<td>1.05</td>
<td>.45</td>
</tr>
<tr>
<td>(10,8)(5,8)</td>
<td>3.10</td>
<td>2.97</td>
<td>2.97</td>
</tr>
<tr>
<td>(10,2)(5,2)</td>
<td>.20</td>
<td>.004</td>
<td>.002</td>
</tr>
<tr>
<td>(10,2)(5,8)</td>
<td>1.93</td>
<td>2.15</td>
<td>.91</td>
</tr>
<tr>
<td>(5,2)(10,8)</td>
<td>2.33</td>
<td>1.63</td>
<td>.86</td>
</tr>
</tbody>
</table>

As might be expected the results for this performance measure lead to the same conclusions as the average waiting times measure. The percentages of the clinics where no patients had to wait more than 30 minutes over the 500 clinics are illustrated in the figure below.
In all cases except the (10,2)(5,2) scenario, the proportion of clinics with nobody waiting more than 30 minutes is higher for the semi lean and lean compared with the base schedule. Although for the (10,2)(5,2) case, for this measure the three schedules are the same but since semi lean and lean have performed better than the base for the measure of number of patients, there is no contradiction in the results.

Further considerations

In order to have a better insight to the level of changes in the waiting times, due to longer service time characteristic, we aim to find the average waiting times for the three appointment schedules for two new cases of service times as (10,4)(20,16). These cases have doubled average and variance compared to the (5,2)(10,8) and (5,2)(10,2) cases in table 2. The results are illustrated in the table below.

<table>
<thead>
<tr>
<th>Service time</th>
<th>$\bar{W}$ for Base schedule</th>
<th>$\bar{W}$ for Semi lean schedule</th>
<th>$\bar{W}$ for Lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,4)(20,16)</td>
<td>9.80</td>
<td>9.24</td>
<td>6.11</td>
</tr>
<tr>
<td>(10,4)(20,4)</td>
<td>5.97</td>
<td>3.41</td>
<td>3.41</td>
</tr>
</tbody>
</table>

Comparison of the results with the corresponding cells of table 2 indicates an increase of the waiting times for all three schedules. The ratio of increase in the waiting times is in range of 1.35 to 1.45 which is less than double. The main reason being that increase in the average service time in a fixed clinic operation time will result less number of patients. Less number
of patients would also result in less congestion because with larger numbers, the cumulative sum of delays in the visit time keep building up the queues and waiting times. The improvement gaps between the base and recommended schedule (lean for the first case and semi lean for the second) has widened compared to the halved service time cases.

2. Disproportionate percentages of the two categories

In order to investigate different schedules for the situation where the two patient categories have different percentages, the model is run for all the previous service times, with different arrival rates. 80% of the arrivals are assumed to be from category one patients with \((\mu_1, \sigma_1)\) service time characteristics and 20% from category two with \((\mu_2, \sigma_2)\) service time characteristics. Table below gives a summary of the average waiting time results.

<table>
<thead>
<tr>
<th>Service time</th>
<th>(\bar{W}) for Base schedule</th>
<th>(\bar{W}) for Semi lean schedule</th>
<th>(\bar{W}) for Lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,2)(10,8)</td>
<td>-</td>
<td>4.27</td>
<td>2.45</td>
</tr>
<tr>
<td>(10,8)(5,8)</td>
<td>9.20</td>
<td>9.05</td>
<td>9.11</td>
</tr>
<tr>
<td>(10,2)(5,2)</td>
<td>3.24</td>
<td>2.29</td>
<td>2.48</td>
</tr>
<tr>
<td>(10,2)(5,8)</td>
<td>4.58</td>
<td>4.29</td>
<td>2.61</td>
</tr>
<tr>
<td>(5,2)(10,2)</td>
<td>4.13</td>
<td>2.77</td>
<td>2.80</td>
</tr>
<tr>
<td>(5,8)(10,2)</td>
<td>10.82</td>
<td>10.4</td>
<td>8.21</td>
</tr>
</tbody>
</table>

The overall conclusions in terms of the best policy for different service time characteristics remain as the same as the previous case where the percentages of the two categories were the same. Except that the improvement gaps are now narrower, meaning that in each case the recommended best policy remains the same but the differences between the improved waiting times and the base schedule’s waiting times are slightly less.

In all cases the category with the increased percentage is the one with the bigger average service time, except the last row. One explanation can therefore be that the most decrease on the waiting times from lean and semi lean policy is on the less time consuming patients. Consequently, increasing the percentage of longer timed category will result in less overall reduction of waiting times. In the fifth and sixth row of the table where the dominant percentage is from the shorter service requirement category, the improvement gap is comparable with its corresponding service time characteristic in row 3 of table 1. For the (5, 8)(10, 2) case, as concluded in the previous section, we have allocated the first part of the clinic to the (10, 2) category with 20% of arrivals and the second part for the (5, 8) category with 80% of arrivals, in the model. The slight narrowing of the improvement gaps for the last two cases, compared to their correspondence in table 1, is due to an increase in the total number of patients. Since the average service times of 5 minutes have a higher percentage, this allows for more booking in the 3 hour clinic and the total
number of patients has increased from 24 to 29. The same argument holds for the first four cases, where in the first row the total number of booked appointments is 18 and for the second to fourth row, this number is 20.

3. Equal proportions of the two categories in a 6h clinic
The third factor under study in this model is the total clinic time. Here we double the total service time to 6 hours and keep the arrival rates of the two patient categories equal. For different service time distributions, table below gives a summary of the results.

<table>
<thead>
<tr>
<th>Service time</th>
<th>$\bar{w}$ for Base schedule</th>
<th>$\bar{w}$ for Semi lean schedule</th>
<th>$\bar{w}$ for lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,2)(10,8)</td>
<td>-</td>
<td>8.77</td>
<td>5.41</td>
</tr>
<tr>
<td>(10,8)(5,8)</td>
<td>14.74</td>
<td>14.03</td>
<td>14.46</td>
</tr>
<tr>
<td>(10,2)(5,2)</td>
<td>5.96</td>
<td>3.83</td>
<td>3.59</td>
</tr>
<tr>
<td>(10,2)(5,8)</td>
<td>10.78</td>
<td>10.61</td>
<td>6.65</td>
</tr>
<tr>
<td>(10,8)(5,2)</td>
<td>10.58</td>
<td>10.33</td>
<td>12.70</td>
</tr>
<tr>
<td>(5,2)(10,8)</td>
<td>12.37</td>
<td>10.29</td>
<td>6.77</td>
</tr>
</tbody>
</table>

The overall conclusions in terms of the best policy for different service time characteristics again remain the same as section 1, where the total clinic time was 3 hours. Except that the improvement gaps here are a bit wider especially the result from last row of the table is quite impressive. The lean policy in this case has resulted in 5.6 minutes reduction of the average waiting times.

4.1.2. Overbooked Clinics
In all of the above experiments, a simple assumption was made that all the booked patients would attend for their appointment. In reality though there is usually a percentage of patients who do not turn up for their appointment. This would result in some kind of waste in resources, and reduction in doctor’s utilization. Since the doctor is available and the clinic is running but due to the non attendance rate of the patients, the resources will not be used in all of their operational period. As it is not possible to require every patient to attend in the real world, most places have attempted a different approach to reduce the resource idle time.

The idea for overbooking appointments gets its support from the above argument. Clinics tend to book more than one patient for a time to cover the non attendees. The number of overbooked appointments throughout the whole clinic would depend on the non attendance rates for that clinic. The pattern for this however, varies between different places. Some would just overbook at the beginning of the clinic; some spread them in certain time intervals through the whole clinic. Whatever the pattern, this attempt will result in increasing the patient’s waiting times, together with its positive effect of increasing the doctor’s utilization. Since this strategy is
supported strongly in appointment scheduling systems, we will model this condition and look at the differences between a base, semi lean and lean in an overbooked schedule with a fixed non attendance rate.

Before going into details of this new model, one argument needs to be flagged up at this stage. Elimination of waste lies in the core of lean principles and within that frame, waste is defined from patient’s point of view only. In short terms, anything that does not bring value for the patient directly is some kind of waste. In our experiment, although the non attendance will result in some kind of waste, i.e. resource capacity, but since it is not a waste for the patients; lean will not consider it to be eliminated. Instead, waiting times have the priority. Therefore, within a lean framework overbooking would not be recommended. But since many would consider it as an inevitable consequence, we will study to see if lean can add any value to such a system.

4.1.2.a Model parameters

- Arrivals

The arrival patterns as for the three schedules are the same as discussed in the previous section, with two main differences. Here there is a fixed probability for patients not turning up for their appointment. We assume the probability is fixed at 8% for each patient category, which is close to the practical values of 10% in real cases and to have integer number of overbooked appointments we have set it to 8%. The overbooking is in a way where one is given at the beginning of the clinic and one at the middle of the clinic, this would cover the total average number of non attendees which is 2, for our case of 24 patients. Therefore the expected workload would be 100%, which can provide reasonable scope for comparison of the results with the non overbooked clinics. This for the lean schedule would mean that there is one overbooked at each part of the clinic for both categories of patients. Therefore the pattern for overbooking and the overall nonattendance rates are the same for the three schedules. The non attendees are determined base on a random number generation in the model. The percentages for the two categories of patients are assumed to be 50:50 and the appointments are spread through the whole clinic time which is set at 3 hours in this section.

- Assumptions

The assumption that patients are on time for their appointments, still holds here. This means that no individual can be seen by the doctor earlier than their appointment time, but due to the stochastic nature of the services, they are likely to have to wait to be seen.

- Other performance measures

In this section, apart from the average waiting times and the number of patients who have to wait more than 30 minutes, we look at an extra performance measure of doctor’s utilization. Since the whole idea of overbooking was to improve this performance, although it is not usually referred to in the lean health literature, we aim to observe the effect of different policies on this measure.
4.1.2.b Validation of the model

The extra parameter to be validated in this part of the model is the non attendance rates which are set at 8%. For this purpose we have collected data on the attendance rates for different schedules and checked that the average attendance rate is equal to 92%. The distribution for one of the scenarios is given as an example in appendix 1. The other validation steps for the model are the same as the previous section, for the non overbooked clinics.

The model runs for 500 iterations and T-tests are performed to test for significance of difference in the average waiting times for different schedules. The following idealistic assumptions and extra performance measures are made for this part of the study.

4.1.2.c Results

For the same service time characteristics as section 1, the model is run for the three appointment systems, with the fixed non attendance rates and overbooked clinics. The results for the average waiting times are given in the table below.

<table>
<thead>
<tr>
<th>Service time</th>
<th>( \bar{W} ) for Base schedule</th>
<th>( \bar{W} ) for Semi lean schedule</th>
<th>( \bar{W} ) for lean schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10,2)(10,8)</td>
<td>-</td>
<td>10.46</td>
<td>9.20</td>
</tr>
<tr>
<td>(10,8)(5,8)</td>
<td>13.99</td>
<td>9.99</td>
<td>10.22</td>
</tr>
<tr>
<td>(10,2)(5,2)</td>
<td>7.52</td>
<td>4.08</td>
<td>3.48</td>
</tr>
<tr>
<td>(10,2)(5,8)</td>
<td>10.47</td>
<td>8.56</td>
<td>6.06</td>
</tr>
<tr>
<td>(10,8)(5,2)</td>
<td>9.50</td>
<td>7.76</td>
<td>8.64</td>
</tr>
<tr>
<td>(5,2)(10,8)</td>
<td>11.07</td>
<td>7.81</td>
<td>6.19</td>
</tr>
</tbody>
</table>

As expected, the average waiting times have increased compared with the non overbooked clinic in section 1. In terms of the best policy with the minimum average waiting time, the results stay the same, apart from the (10,8)(5,8) case, whereas before there was no real difference between the policies but here, the semi lean policy results has substantial advantage over the base policy. The (10,8)(5,2) case has only been put to the table to point out again the difference between the priorities for these service time characteristics. As was concluded in section 1, this will not be the best arrangement and instead it would be replaced by (5,2)(10,8), which does have significantly lower average waiting time under the lean policy.

In all cases both lean and semi lean outperform the base system. Since the semi lean system would be more convenient for the patients, unless lean makes a huge amount of difference, the preference would be with the semi lean system. The figure below shows the difference in waiting times of these two policies. As can be seen from the graph, for the last two cases where the variances and the means in the service time are different for the two categories, the reduction of average waiting times under the lean policy is more than 1.5 minutes compared with the semi
lean. Although in practice other considerations need to be taken into account, but in theory the lean system of appointment could save more than 36 minutes in total in a 3 hour clinic with 24 patients, compared to the semi lean for these scenarios.

Figure 4.4. Difference in the average waiting times of the semi lean and lean system.

The doctor’s utilization in our model is defined according to the following formula.

\[
\text{Doc Utilization} = \frac{\text{Doc busy time}}{\text{Doc available time}} = \frac{\sum_{\text{all attendants}} \text{Service time}}{\sum_{\text{all appointments}} \text{Service time}}
\]

For the doctor’s utilization measure, under the three policies, there is no real difference and the average of this parameter to two decimal places, for all of the above cases is .92± .08.

For the number of patients who have to wait more than 30 minutes, the results for the average numbers are shown in the figure below. As can be seen from the figure, the lean and semi lean system result in smaller number of patients in comparison with the base schedule for all the service time scenarios. In comparison between the semi lean and lean policy, this measure to exactly the same recommendations as the average waiting times.
4.2. Conclusion
Our simulation model for the appointment scheduling of two patient categories, provides the opportunity to explore better appointment systems for different service characteristics, in order to improve different performance measures such as average waiting times for patients or number of patients having to wait more than a certain time. The model is flexible to test for different appointment schedules. However, the main objective in this chapter was to investigate the effect of streaming the patients and allocating separate time slots for each patient type. This schedule was compared with the base schedule of random allocation, giving every patient the same amount of time based on the average of the two category’s service time. A third alternative appointment system was based on giving each patient the time they require based on their average service time, but spread the two categories throughout the whole clinic without streaming them.

Three different factors were investigated in an ideal system with zero nonattendance rates for patients and no overbooked appointments. Firstly, by considering equal percentages for the two categories of patients, we found the following key results:

1. Variation rather than the average length of the service times causes the longer waiting times.
2. In the case where service times of both patient categories have equal big variances, streaming or different time slot appointments does not improve the waiting times much.
3. For cases where only one service time has big variation, streaming has a dramatic effect in reduction of the waiting times under the condition that the category with the smaller service time variation is given the priority and is given the first part of the clinic.
4. For service times with equal small variations, just giving each category the appointment
time slot based on their own service time average, can have the same effect as streaming,
with the benefit of convenience for patients.
5. Increase in the average and variance of the two service time categories, leads the same
conclusions with wider improvement gaps.

Secondly, by increasing the percentage of one category with respect to the other, it was
revealed that the improvement gap, by the streamed and semi lean policy from the base
policy gets narrower. This is due to the dominance of less improvement in the waiting time
of the longer service and most variant category.

Thirdly, by keeping the percentages of each category equal and doubling the total clinic time,
the reductions in the average waiting times under the lean system, nearly double, compared
to the shorter clinics.

Finally, in modeling overbooked clinics with constant and equal non attendance rates for the
two categories of patients, it was found that the overbooking does increase the average
waiting times. But in this case as well the semi lean and lean policies do reduce the waiting
times compared with the relevant base model. The following key conclusions can be pointed
out for this case:

1. For the cases of different variance for the two service time category, streaming under
the lean policy does reduce the waiting times considerably.
2. For the cases of equal variance in the service time, just allocating different time slots
for the appointment of the two categories of patients, without streaming them, does
reduce the waiting times nearly the same as complete streaming.

One point of argue is that the lean policy has the disadvantage of reducing the total range
of time available for each category of patients. Therefore in cases where the lean and
semi lean policies result in nearly same performance measures, the semi lean is
advantageous.

The three different scheduling policies do not perform much different in their measure of
doctor’s utilization. Regarding the number of patients who have to wait more than a
certain time, where the streaming and semi lean policy have been recommended by
average waiting times, this measure also outperforms. Meaning that wherever one
specific policy results in lower average waiting times, this same policy leads to the
minimum number of patients having to wait more than 30 minutes. This might have been
predictable as both measures although different, but are related to the same thing of
patient’s waiting times.

The analysis of a lean recommended policy in a simple appointment based clinic has
enabled us to think more in depth about similar issues. In our example case, the results
are in support of a lean recommendation. Therefore this at first might lead the reader to the conclusion that lean can be put into practice without the need for further analytical analysis and modeling. However, as our results indicate, the improvements made by a lean scheduling system, is not the same for every scenario under study in this chapter. For example, the reduction in the average waiting times from the lean schedule compared to the base schedule is .04 minutes for the (10, 8) (5, 8) scenario and 3.24 minutes for the (10, 2) (5, 8) scenario, in a 3 hour clinic with equal proportions of patients. This means a total waiting time reduction of .96 minute and 77.76 minute respectively, in the whole clinic. The additional consideration is from the practical point of view. The lean schedule requires the clinic time to be divided to two parts and each part to be allocated to one of the patient categories. Practically, this limits the time window that each category is given, but the significant improvement that it makes for the (10, 2) (5, 8) scenario, makes this limitation worthwhile, while for the (10, 8) (5, 8) scenario, it will not be a sensible change.

Also the result shows that increasing length of the clinic time will widen the range of improvements by the lean schedule. For the (10, 2) (5, 8) scenario, as discussed above, 43% of the time is saved in a 3 hour clinic, this increases to 55% for the same scenario in a 6 hour clinic. Similarly, the proportion of the different categories of patients is an influencing factor in the improvement range. These findings could not have been concluded from lean principles. In more complicated cases where the number of servers changes or there are more than two categories of patient, other influencing parameters such as staffing levels or combinations of some categories need to be taken into account.

Overall, the streaming policy under lean principles can be advantageous in reducing the waiting times in an appointment system. However, lean on its own does not clearly state how to distinguish between the two different value streams in this system. This report has provided valuable guidelines in recognizing different value streams under different conditions. As prior to modeling such system, one could presume that the patient types with different average service time can form different value streams. The modeling has revealed that streaming is wiser to be based on the service time variations rather than averages. Also it has provided guidelines on the priorities of the two categories, with the recommendation of giving the priority to the category with smaller variance in service time.

The simulation model developed for this study is quite flexible and provides the opportunity to investigate the effect of these three policies on other new service time characteristics, clinic hours and proportions of the two categories.
5. Conclusions

Lean provides a total system approach but is short on details, organizational structures, and analytic tools for diagnosis. In a large service system, due to the complexity and the interaction of sub-systems relying only on one specific strategy can result in limited managerial powers. Lean lacks the analytical tool and evidence, but there are also critical views on some of its principles.

The process mapping of Lean provides the tool to specify the root problem. This step is aimed to define the key processes and value added steps from the customer’s point of view. This is a limitation itself, since the needs of both server and customer needs to be taken into account. The process mapping is one of the operational management tools and has been widely used in problem solving and consultancy projects. Sometimes this is only applied from the project owner’s point of view, which in Lean is considered to be the customer. But in order to arrive at the optimal plan for the entire system the requirements of every stake holder needs to be taken into account.

Customer Critical-to-Quality needs are not front and center: In Lean, the person who creates the value stream map decides as to whether an activity is value added or not. Furthermore, the definition of waste in Lean is based only on the direct customer’s requirements. This, as already seems to have been realized by the Lean managers is not applicable in the system. For example the change in the inventory strategy of Toyota proved that although inventory is not a direct beneficiary for the customer, but in the long run it is not beneficial to eliminate it. Therefore Lean fails on itself to define the actual level of “wastes” which sometimes might be beneficial in the system.

In terms of putting the lean principles to practice in the healthcare system, some additional difficulties might be encountered due to the sensitivity and complexity of the system. A critical mind on a hypothesized lean health system might pick up on some of these challenges as discussed in our next section.

5.1. Problems with lean in the healthcare

In the lean framework one of the approaches in the reduction of non value adding activities is moving towards a “just in time” system by having more frequent services. However, implementation of such a strategy in all parts of the hospital and the entire NHS would require the professional members of staff to be available more frequently and create a lot of inconsistency in their job. In the long term it could have the negative effect of low performance and bringing other potential problem into the system due to lack of concentration and confusion for the resources. This especially for the more professional members of staff who are in direct contact with the patient could have life frightening consequences for the patient and lasting effects on the medical staff’s reputation. Therefore it is highly unlikely to be able to schedule such a system through the entire health care system due to the inconveniences for the resources.
In contrast with the automobile industry, the direct resources in the NHS system are doctors who always need to be in their best functional mode and cannot be faced with too much variability in short time intervals. For example it is more difficult for a surgeon to adopt himself in doing different kinds of surgery in a row than organizing different workers in the car industry for different assembly parts. In the latter, a manager could plan and allocate resources to different parts and most of the hand offs can be automated. Whereas in the health system, the customer is a patient and the product usually is a person’s life. In similar cases different doctors would have different opinions and recommendations, which makes it crucial for the patient to be under the care of one doctor and if the system was scheduled in a way that a doctor had to do frequent services in different places and spend much time in travelling and preparation for variable tasks, it could have serious effects for the patients.

The idea of short and more frequent services is based on the ‘Just in time’ strategy which was first introduced by the Toyota in the 1950s. This was the perfect principle to serve the automobile market for that time, since the demand was small and Toyota was the first company to suggest and implement this system. However, as more companies come to follow these principles, the manufacturers and the community will be faced with serious problems, some of which were discussed in chapter 2. One of these problems would be for the community and government to be faced with high congestions in urban areas due to high levels of delivery for the companies, since the aim with lean is to minimize the inventory. This eliminates waste in one area but created it elsewhere in the system where people have to spend a lot of time in traffic. Also Japan was faced with shortage of workers under this principle that companies needed to employ foreigners and this excessive attraction for labor force from other countries can have can have other negative social and economical effects for the community. Although lean was the answer for the automobile industry in that certain time and did help Toyota to become one of the leading car manufacturers in the world, its globalization for the entire system might not be possible. This would be due to the fact that the high demanding policy of just in time delivery eases the flow in one part of the system, while it puts more pressure to some other part. In a more complex system like health care, where the product is of infinite value and there is a limit on the number of trained professionals as the servers, this type of service will not be practical through the entire system.

Elimination of waste lies at the heart of a lean system. This, at first seems a logical approach. However, if we categorize any kind of queue and waiting time as a kind of waste, it would not be practical to eliminate it through the entire system. For example increasing the number of resource in the system or distributing uncategorized tasks for some parts in order to ignore queue will have other consequences, where it might be more beneficial for the customer to queue rather than to face an unreliable server. On the other hand in places where there is clerical work and document handling or department settings, implementing this principle can increase efficiency and there is much less risk involved in case of failure.
There is also this belief among some of the clinicians that having a waiting list protects them from a tide of unreasonable and untenable demand (Lendon, 2004). Although a lot of people within the system would disagree, one support for this argument was provided by one of the case studies which we came across in this study. In this case reduction of variation in capacity in the breast unit of a city hospital leads to reduction of the waiting times (Silvester, 2004). Thereafter an increase in the average demand from 29 to 36 per week occurs for no apparent reason. Whether this excessive demand can contribute towards improvement in the community’s health level or create waste and cost in the system, is another area of research.

The idea of a pull policy appears to be more practical in the manufacturing industry than healthcare. In the health system, in parts where patients enter the system according to an appointment system and have the ability to wait, it is possible to allocate everybody to parts where can exactly serve their requirement. However, in many walk-in centers such as A&E, where the arrivals are out of control to a great extent, it will not be affordable to implement a pull strategy. For example it is highly likely that the ward specializing in a particular patient’s requirements that needs to be discharged from A&E is full at the moment. An extreme pull system would keep them in until the relevant ward becomes available. The actual disadvantages of such a system in this department seem to overtake the advantages. Because not only the patient will receive less quality care in A&E than they would if they were admitted in ward, but also the A&E will be more likely to run out of resources. This could cause more frustration for the patients and staff, with the risk of endangering urgent cases whom face an overcrowded A&E. The situation here is not really comparable with a manufacturing industry, where waiting of artificial components in one station to be pulled from the next related station, might not actually have any harmful consequences.

Streaming of the different value streams is another approach in a lean system that is aimed to ease the flow for each one and let them move through the process in their own speed. As discussed in the A&E case studies in chapter 3, this policy has shown to lead to decrease of waiting time for the patients in this department, in practice. The simple experiment that we explored in chapter 4 was particularly designed in this study, to investigate the benefit of such streaming in an appointment based clinic. As discussed earlier, the nature of this policy is beneficial for the system. However, it is often hard or impossible to distinguish between different value streams, relying only on lean. In order to do so, modeling of the system and processes does provide a clearer picture of the situation and consequences of different changes. This then can help the decision makers to better define value streams and also provides the analytical support for implementation of any changes.

Apart from the above arguments which are related to the principles of lean and differences between a manufacturing industry and a health system, there is another concern among the managers and researchers on this area. They point out to the cultural differences in a society which can make implementation of lean in Japan, incomparable to its implementation in the rest of the world.
5.2. Lean failure due to different cultural structures

The lean production model has been one of the pillars of the highly productive Japanese economy during the postwar period. Its transference to other countries has been less successful. In the US, for example, workplaces employing such techniques apparently witness little significant productivity improvement (Freeman, 2000).

Several explanations have been advanced in the literature that could account for the different performance of lean production approach in Japan and the US. Some analysts argue that productivity improvements stem from the way in which workplace features are packaged and not from the adoption of one or another. The most notable aspect of lean production that US have not adopted is the promise of lifetime employment. Ichniowski (1997) gives evidence that supports the idea of this being one of the main reasons for lean’s less success in the US.

Cultural differences also have been advanced as a possible explanation for differences in productive efficiency across countries with similar organizational features. For example, self-interested behavior is not as acceptable in Japan as it is in the US. Hence lean production in Japan may result in greater improvements in productivity and product quality, compared with the US (Fairris, 2002).

These cultural differences are more related to the acceptance of such systems by the work forces. One might be able to oblige the workers in a factory to obey a certain policy as this class of people may be less powerful in terms of their job maneuver. In a hospital though, the resources are doctors, nurses and more qualified people. The workforces in a healthcare system actually are more authoritative than the managers of the system. For example if a doctor does not agree with a plan change, in contrast with a factory worker, he/she is in a position to choose to not to go ahead with it.

5.3. Discussion

Overall, in making any modifications in the health care services it is crucial to do extensive research and have strong evidence and proof for improvement beforehand. There has been examples of improvement in the services where have put the lean principles into practice. However it cannot be concluded that lean would be the perfect solution for the entire system.

For simple cases where patients are not directly involved and recommendations such as restructuring and data handling are recommended to improve the system, relying only on lean can actually bring value to the system. This helps to make rapid improvements without the need for modeling. For more complicated issues, lean does have simple rules. However, their implementation without detailed analysis sometimes will do more harm than good for the system.

Separation of different value streams is one of the lean approaches in easing the flow which we studied in detail in this report. Our own analysis of this strategy in an appointment based clinic proved that the capability of lean in making improvements in such systems is dependent on the
system’s characteristics. For example the streaming policy can save up to 55% of clinic’s time by reducing waiting times for some scenarios while for some other scenarios it does not make any considerable improvements.

Similar arguments hold for the lean example case on the A&E department that was discussed in chapter 3. The practical implementation of lean in the Flinders medical center leads to the separation of two categories of patients in terms of their waiting areas and resources. As a result improvements were made in terms of times the patients spend in the department and so on. In light of our own analysis and the results and arguments that we made earlier in chapter 4, we can say that it will be impossible to reach the best design for the flow of patients in such complex system as A&E, relying only on lean principles. A lot of factors like staffing levels, further divisions of categories, arrival rates and waiting areas for different categories play influential roles in the operation of the system. It also seems that the range of improvements is dependent on characteristics of individual cases. Therefore, similar results could not be expected for other A&E departments.

Lean does provide a guideline for changes that can lead to improvements. It can be a framework to work under, it lacks the analytical tools. In the healthcare system, where every clinic, hospital or department in different regions has different characteristics, similar approaches are expected to result in quite different outcomes. Modeling is a powerful tool which can be considered as a complementary to lean and helps the decision makers to see the potential consequences of their actions beforehand.

The lean principles were originally designed within the manufacturing industry. Healthcare system is different in terms of its service operations and goals from the manufacturing industry. This brings some of the limits of lean for the design of healthcare system while others refer to the concepts themselves. The latter has been argued in chapter 2 and the expected obstacles of lean in healthcare have mostly been pointed out in chapter 4 and 5. This all does not mean that it should be completely ignored for the healthcare system because in many cases the lean principles have the potential of improving the system. Since a lot of focus in lean is on elimination of waste and ease of flow it especially can be useful in the queuing issues.

In conclusion the lean principles propose some changes or redesigns which can improve the system, especially where queuing issues are present. This report has provided guidelines on the use of lean for queues within appointment based systems. Not always these approaches can be effective and their success depends on the circumstances. Since lean does not have analytical tools, itself its implementation in the system needs to be supported by other analytical methods. Simulation modeling and queuing modeling which have long history in the study of queue issues in the health system, are seen as most powerful tools for this purpose.
5.4. Further work

This report did provide valuable insights on lean principles and their approach to queue issues in healthcare. The idea of providing guidelines for implementation of lean on every different case seems rather unfeasible now. This is partly due to time limitations of the project. Also most of the health queue issues seem to be dependent on some characteristic factors of the particular practice or department. As our simple example proved, changing some factors like the parameters of a service time or length of a clinic can influence the extent of improvements by changes. However, the experiment did provide some guidelines on use of streaming policy in an appointment system. This cannot be generalized to other systems where arrivals have different pattern. Although we did refer to the reported achievements of this in practice, modeling of such policy in a walk-in centre is also believed to be valuable in this context. Exploration of the pull principle does seem to be beyond the scope of this project. As it involve strategic decision makings and a dynamic simulation modeling that takes into account interaction of different departments within the system.
6. Appendix

This appendix covers a brief explanation on how the model operates and some of the detailed analyses in the process of model validation. The figure below illustrates a schematic presentation of the simulation model.

**Figure A.1. A schematic presentation of the simulation model in Excel.**

<table>
<thead>
<tr>
<th>Category 1: $\mu=10$, $\sigma=8$</th>
<th>Category 2: $\mu=5$, $\sigma=8$</th>
<th>Ratio: 50:50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>Semi lean</td>
<td>Lean</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7.5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>22.5</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>37.5</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>52.5</td>
<td>52.5</td>
<td>55</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>67.5</td>
<td>67.5</td>
<td>70</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>70</td>
</tr>
</tbody>
</table>

**Base**
- $\mu=2.030301$
- $\sigma=0.046338$
- $\text{Avg} = 29.37958$
- $\text{StdDev} = 24.00328$
- $\text{Max} = 4.628958$
- $\text{Median} = 0.516741$
- $\text{IQR} = 10.95163$
- $\text{Min} = 2.557886$
- $\text{Max} = 9.72315$
- $\text{Median} = 6.872203$
- $\text{Max} = 2.768106$

**Semi lean**
- $\mu=8.542376$
- $\sigma=2.751497$
- $\text{Avg} = 0.099678$
- $\text{StdDev} = 10.74922$
- $\text{Max} = 3.678497$
- $\text{Median} = 3.721656$
- $\text{IQR} = 7.069602$
- $\text{Min} = 0.012529$
- $\text{Max} = 13.73738$
- $\text{Max} = 25.90612$

**Lean**
- $\mu=20.94928$
- $\sigma=12.17754$
- $\text{Avg} = 14.44832$
- $\text{StdDev} = 16.79885$
- $\text{Max} = 23.22024$
- $\text{Median} = 17.7055$
- $\text{IQR} = 20.74922$
- $\text{Min} = 0.012529$
- $\text{Max} = 13.73738$
- $\text{Max} = 25.90612$

**Base visit time**
- 0 10.57268 13.37051 42.84977 77.17055 83.52116 91.5716 105.2985 109.7644 119.5 131.1565
- 0 10.57268 15 42.84977 77.17055 83.52116 91.5716 105.2985 109.7644 119.5 131.1565
- 0 29.49166 44.4207 58.9687 86.08505 98.55592 111.711 121.5559 133.6646 137.3559 155.8776

**Semi lean visit time**
- 0 29.49166 44.4207 58.9687 86.08505 98.55592 111.711 121.5559 133.6646 137.3559 155.8776
- 0 29.49166 44.4207 58.9687 86.08505 98.55592 111.711 121.5559 133.6646 137.3559 155.8776
- 0 29.49166 44.4207 58.9687 86.08505 98.55592 111.711 121.5559 133.6646 137.3559 155.8776

**Lean visit time**
- 0 20.94928 33.12683 47.57515 64.374 75.12322 80.74463 87.81424 98.01491 101.6938 115.4311
- 0 20.94928 33.12683 47.57515 64.374 75.12322 80.74463 87.81424 98.01491 101.6938 115.4311
- 0 20.94928 33.12683 47.57515 64.374 75.12322 80.74463 87.81424 98.01491 101.6938 115.4311

**Base waiting time**
- 0 3.072677 0 20.34977 47.17055 46.02116 46.5716 52.79854 49.76437 52.00005 56.15654
- 0 3.072677 0 20.34977 47.17055 46.02116 46.5716 52.79854 49.76437 52.00005 56.15654
- 0 3.072677 0 20.34977 47.17055 46.02116 46.5716 52.79854 49.76437 52.00005 56.15654

**Semi lean waiting time**
- 0 24.49166 34.4207 43.9687 61.08505 68.55592 76.71103 76.55595 83.66457 82.35594 90.87762
- 0 24.49166 34.4207 43.9687 61.08505 68.55592 76.71103 76.55595 83.66457 82.35594 90.87762
- 0 24.49166 34.4207 43.9687 61.08505 68.55592 76.71103 76.55595 83.66457 82.35594 90.87762

**Lean waiting time**

**Average**
- Semi lean waiting = 109.9201
- Semi lean wait>30 = 19
- Average Lean waiting = 22.8973
- Average Base waiting = 70.95628
- Average Lean wait>30 = 5
- Average Base wait>30 = 17
The first three lines of the model represent the arrivals which are based on the three appointment schedules of base, lean and semi lean as explained in chapter 4. The next three lines represent the service times for each arrival. Service times are generated from a Gamma distribution and different mean and variances are implemented to investigate the effect of different schedules on different situations. For the base, the parameters of distribution are set randomly from the characteristics of the two categories of patient. For the semi lean and lean, the parameters of the distribution for each patient are determined, based on their appointment time slot. The variable defined as time in the model, is set to keep the track of time when each patient is due to be visited. It is defined as the cumulative sum of all the previous service times at each point of appointment. The visit time for each patient is set by comparison between their appointment time and the system time at that point and is defined as the maximum of the two. The waiting time for each patient is then defined as the difference between their appointment time and visit time.

Part of the model validation as discussed in chapter 4, involves justification of the input distributions. The following paragraph covers the explanations for this part.

The figure below, gives the distribution on the total number of patients for the semi lean schedule for which the time slots are chosen based on a random number generation for the two categories of patients. This figure is for the scenario where the percentage of two categories is 50:50, with average service times of 5 minutes and 10 minutes for the two categories of patients. Over a 180 minute (3 hour) clinic, the expected number of patients would be 24 as calculated below and the figure provides the support for the appropriate distribution of arrivals.

\[ 10(12) + 5(5) = 180 \]

In the next step, to test for the right sampling of the service time distributions, we test for the mean difference between a service time distributions with one of Gamma distributions with the
same parameters. For the example illustrated in figure 3, the p-value of .97 provides strong
evidence to accept the null hypothesis of no difference to 5% significance level. This leads us to
trust the model of adequate sampling for the service times.

Figure A.3. A statistical test of the mean difference on the service time distribution with a test data
distribution.

Figure A.4. The attendance rates for the base schedule on the (10,8)/(5,8) scenario.

For the overbooked clinic, one extra step toward validation of the model, remains regarding the
non attendance rates for the arrivals. In the model, the appointment slots are set the same as
previous section apart from two overbooked appointments for the beginning and middle of the
clinic. Then by use of random number generation, a non attendance rate of 8% is set throutout the
clinic for all the three schedules. To confirm that this process is running as expected, we collect
data on the actual attendees to the clinic. One example of this data is illustrated in the figure
below which is for the attendance rate of the base schedule for the (10,8)/(5,8) scenario. The
average attendance rate of .919=.92 is the proper figure as expected by the model scenario. This
same result is obtained for the other schedules and scenarios.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9.997417</td>
</tr>
<tr>
<td>Variance</td>
<td>65.34782</td>
</tr>
<tr>
<td>Observations</td>
<td>500</td>
</tr>
<tr>
<td>Hypothesized Mean Difference</td>
<td>0</td>
</tr>
<tr>
<td>df</td>
<td>499</td>
</tr>
<tr>
<td>t Stat</td>
<td>0.032517</td>
</tr>
<tr>
<td>P(&gt;</td>
<td>t</td>
</tr>
<tr>
<td>t Critical one-tail</td>
<td>1.647913</td>
</tr>
<tr>
<td>P(&gt;</td>
<td>t</td>
</tr>
<tr>
<td>t Critical two-tail</td>
<td>1.964729</td>
</tr>
</tbody>
</table>
7. References


King's Fund (2004). *Has the government met the public's priorities for the NHS?: A King's Fund briefing for the BBC 'Your NHS' day*.


Silvester, K. a. (2004). Case study to demonstrate the principles in the pape r'Reducing waiting times in the NHS. is lack of capacity the problem?'. *Clinician in management*.


Souza, L. a. Launching Lean Thinking in a UK NHS Trust., (pp. 41-53).


