Managing Lateness to Improve Process Flow in Primary Care

A Major Qualifying Project Report submitted to the Faculty of
Worcester Polytechnic Institute
In partial fulfillment of the requirements for the Degree of Bachelor of Science

March 23, 2016

Submitted By
Amy Stevens

Advisor
Sharon Johnson

Abstract

Primary care clinics are facing increased patient volume, which can contribute to delays and dissatisfied patients. This project examined root causes of delays and potential countermeasures in a large primary care clinic that is part of an integrated health system. The team gathered data and created simulation models to explore opportunities for improvement. Based on the simulation analysis, reducing variation in appointment durations and earlier patient arrivals had significant positive effects on the timeliness of appointment starts and the total visit time.

Acknowledgements

I would like to express my sincere and overwhelming gratitude to the project liaisons at the sponsoring clinic. Their enthusiasm, knowledge, and guidance made working on this project an exceptional experience. Additionally, I would like to acknowledge every member of the clinic staff who gave us information or allowed us to observe them in the clinic. The information we gathered from them was of the utmost importance as we began the project.

I would like to thank my project advisor, Professor Sharon Johnson, for her high expectations, which motivated me to work hard and produce my best possible work. Last, but most certainly not least, I would like to express my immeasurable gratitude for Sarah Abell, who was a vital part of the project team for the first portion of the project. As part of an independent study, Sarah worked with me to develop a qualitative and quantitative understanding of the clinic's current state, including a significant contribution to data collection and analysis. She is a co-author on the background, methodology, and results from data analysis sections.

Executive Summary

A primary care physician is a patient's first step in preventing, diagnosing, or treating any health issue, and primary care practices provide services including, but not limited to, vaccinations, prescriptions, and referrals to higher-level specialists (Schulte, 2013). Most Americans visit their primary care physician annually, but many patients have more frequent visits for acute health issues or to monitor chronic conditions. This, coupled with increases in both population and ease of access to healthcare in recent years, means that primary care practices nationwide are facing higher-than-ever patient volumes. However, the number of physicians choosing primary care as a career is not increasing at the same magnitude, possibly because primary care physicians make, on average, less than doctors in other specialties (Bodenheimer et al., 2007). A high proportion of patients to doctors makes it imperative that primary care practices operate as efficiently as possible.

Because primary care practices are currently reimbursed on a per-encounter basis, they have a major incentive to reduce costs and schedule as many appointments as possible in order to generate revenue (Bodenheimer & Pham, 2010). However, scheduling more appointments means more opportunities for delays, and "no healthcare organizations or practicing clinicians are immune from the simultaneous pressures to improve quality and lower costs" (Toussaint). In many primary care clinics, issues exist that impact patient flow, and therefore, patient satisfaction. In many cases, the most important determinants of patient satisfaction in a clinic are how long the patient waited to be seen and the overall smoothness of the process (Clinic Physician, personal communication). Delays can stem from a variety of root causes on both the patient side and the physician side. Studies have indicated that patients who arrive late for appointments tend to wait longer to see their physician and spend less time with them overall (Onisuru et al, 2008), which may lead to lower patient satisfaction ratings for the clinic.

This project aimed to assist a primary care clinic that is part of an integrated health system to understand the effects of lateness and delays on the system and to identify the most promising ways to decrease delays in the clinic. The project team first gained a qualitative understanding of the clinic's operations through informal discussions and observations. Quantitative data was obtained from data collected automatically in IT systems as well as through manual data collection from both the patient and the physician perspective. This information was used to develop a simulation model using Arena that represented the clinic's starting state. Alternate models were created and tested against the base model in order to evaluate the effectiveness of different approaches for the clinic to achieve its goal of reducing delays to improve patient experience. The following areas were explored:

- Patient arrival patterns
- The staff schedules of the Patient Care Associates (PCAs)
- The duration of appointments relative to the scheduled duration
- Asking late patients to reschedule their appointments for another day.

To validate the base case model and to compare the results from the alternate simulations to the base case model, two metrics were used. First, the model measures the difference between an appointment's simulated start time and when it was scheduled to start. It also records the total time patients spend in the system, beginning when they complete the registration process and ending after they have seen their primary care provider. The metrics from the model were compared with the same metrics taken directly from the clinic in order to ensure that the model was a representative approximation of the clinic's activities.

All simulation models represent reality imperfectly, due to assumptions made to simplify modeling and data limitations. Once generated, simulation results must be evaluated in the context of the real system. The current project was performed with a narrow set of parameters that might not mirror a particular practice and real world circumstances. The simulation model indicated that patient arrival patterns can have a significant effect on the way the clinic operates on a given day. Increasing the number of PCAs in the clinic did not appear to have a statistically significant effect on the metrics, but this is likely due to the limitations of the simulation model. Reducing variation in appointment durations (so more appointments have measured durations closer to the durations they were scheduled for) as well as asking patients who arrive after a set lateness threshold to reschedule their appointments also proved to be very effective in improving the two metrics.

Based on the information from the data collection and the simulation study, the clinic should focus their efforts on reducing the gap between scheduled appointment duration and measured duration. The clinic should also investigate ways to adjust patient arrival patterns. Reducing the number of extremely late patients had a positive effect on the metrics within the simulation model. One possibility is to explore the introduction of a lateness threshold rule. More research may be needed to formulate a policy that satisfies the clinic staff or to determine possible long-term effects, but the preliminary results from the model indicate that such a policy has the potential to have a positive effect on the clinic's functioning. Finally, due to model assumptions that overestimate the availability of PCAs, adding PCA staff is likely to improve clinic performance more than the model indicates and should be investigated further.

Table of Contents

Table of Figures

Table of Tables

1.0 Introduction

A primary care physician is a patient's first step in preventing, diagnosing, or treating any health issue, and primary care practices provide services including, but not limited to, vaccinations, prescriptions, and referrals to higher-level specialists (Schulte, 2013). Most Americans visit their primary care physician annually, but many patients have more frequent visits for acute health issues or to monitor chronic conditions. This high level of demand, coupled with increases in both population and ease of access to healthcare in recent years, means that primary care practices nationwide are facing higher-than-ever patient volumes. However, the number of physicians choosing primary care as a career path is not increasing at the same magnitude, likely because primary care physicians tend to make, on average, less money than doctors in some other specialties (Bodenheimer et al., 2007). A high proportion of patients to doctors makes it imperative that primary care practices operate as efficiently as possible.

Such is the backdrop where Lean, a technique of process improvement originally developed by Toyota for automobile manufacturing, comes into play. In addition to making mistakes that result in negative health repercussions for patients, many practices struggle with bringing timely and high-quality care to their patients, which not only increases frustrations and risks for patients, but can also create an unsatisfactory work environment for providers. Lean is "a quality improvement philosophy and set of principles that, when well executed, transforms how an organization works and creates an insatiable quest for improvement" (Toussaint). Using techniques such as process mapping, simulation and root cause analysis can help to define key issues in a process's current state that can ultimately be improved in order to save time and money and decrease the occurrence of preventable mistakes in healthcare.

Lean techniques have been used successfully in primary care. Many primary care visits are for the treatment of common, non-chronic illnesses or conditions, or for referral to a higher-level specialist (Schulte, 2013). Every person could benefit from access to primary care to diagnose and treat conditions as early as possible. However, there is a pronounced shortage of primary care physicians in the nation, which contributes in part to several problems with primary care, most notably long wait times and unsatisfied patients (Bodenheimer & Pham, 2010).

Adding to the problem is a combination of tightly scheduled appointments, late patients, late doctors, and several patients arriving simultaneously, all of which can result in shortened appointments and negative effects on the quality of care patients receive from a clinic (Onisuru et al, 2008; Bard et al., 2014; Graig & Perosino, 2011). If any part of the clinic is behind schedule, the whole clinic can be affected for the remainder of the day.

The goal of this project was to analyze lateness and delays at a large primary care clinic that is part of an integrated health system– its causes, its frequency, and its effects – and to perform a simulation study to determine actions that might be effective for addressing delay. The health system offers medical services in several capacities including emergency medicine, inpatient and outpatient care, hospice, homecare and rehabilitation. As stated above, the project focuses only on the large primary care clinic. The current project was performed with a narrow set of parameters that might not mirror a particular practice and real world circumstances.

In order to achieve this goal, the project group first performed manual data collection over the course of one week in the clinic. Using the data that was collected during that period as well as some automatically recorded date from the clinic's systems, a simulation model was created to represent the clinic and the current state of its appointment process. A series of alternative models was then created, each simulating a possible change the clinic could make to how it operates and how it handles late patients.

The report is organized as follows. The common problems found in primary care clinics nationwide are described in Chapter 2. Chapter 3 describes the clinic that was the focus of the project in more detail. A summary of the goals and objectives of the project, as well as the methods used to fulfill them, can be found in Chapter 4. Chapters 5 and 6 discuss the findings from the data analysis and the simulation study, respectively. Conclusions and recommendations are detailed in Chapter 7.

2.0 Literature Review: Primary Care and Lean

Primary care is a patient's first step in preventing, diagnosing, or treating any health issue, and provides services like vaccinations, prescriptions, and referrals to higher-level specialists (Schulte, 2013). In the United States, millions of people see their primary care physicians annually or biennially, and many visit their doctors more often. With such a high patient volume and a limited number of physicians, it is important that primary care clinics operate efficiently.

2.1 Trends in Primary Care

There have been countless advancements in medical technology in the last two centuries. In 1842, Dr. Crawford Long used ether as anesthesia for the first time; the ability to diminish patients' pain allowed surgery to be more widely practiced and allowed earlier intervention for injuries and illness (Schulte, 2013). Since then, humans have developed incredibly advanced and intricate systems for medical treatment. Diseases that were a death sentence a century ago, like diabetes, have become extremely manageable with technologies such as blood sugar test strips and personal insulin pumps. Additionally, some surgeries can be performed non-invasively and patients can often go home within a few hours of surgery (Milestones in Medical Technology, 2012). These advancements in medical treatment have made medicine a more accessible option for those living with ailments. As a result of President Barack Obama's efforts to offer healthcare to all Americans as well as population increases in the last several years, the number of patients seeking care has increased substantially (Schulte, 2013).

While the number of patients making use of primary care clinics is growing, the number of primary care physicians is not increasing at the same rate (Bodenheimer & Pham, 2010). Many medical students and residents are dissuaded from pursuing a career in primary care, considering the relatively low pay of primary care physicians compared to specialists (Bodenheimer et al., 2007). Because primary care physicians are paid on a "fee-for-service basis" (Bodenheimer & Pham, 2010) and these fees are low from both Medicare and private insurance companies, primary care clinics need to serve more patients in order to earn more money (Bodenheimer et al., 2007). Since there has been a vast increase in the number of diagnostic and imaging procedures done over the years, specialists can be paid at higher rates than primary care physicians (Bodenheimer et al., 2007). Thus, primary care physicians must face increasingly busy schedules, further exacerbated by inefficiencies within their practices and the importance of balancing the teaching of residents with research and patient care (Bodenheimer & Pham, 2010). Additionally, physicians and their practices are not financially compensated for referrals and other administrative tasks completed before patient arrival or after patient departure (Bodenheimer & Pham, 2010).

2.2 Quality of Patient Care

Despite recent medical successes, medicine is still plagued with problems similar to those that humanity has struggled with for centuries. Although we understand how infections spread, it is often still the case that patients develop an ailment completely unrelated to the medical problem they started with (Schulte, 2013). Whether this stems from a preventable mistake such as being given an incorrect dosage of medication or from simply developing an infection, health care still has a long way to go. Mistakes can be made quickly if procedures are not properly followed – medication may be given in the wrong quantities or to the wrong patient altogether, surgeries may be completed on the wrong limb or organ, or instruments may be left inside a patient after their surgical incisions are closed (25 Most Shocking Medical Mistakes). Beyond costing hospitals thousands of dollars or more when insurance companies refuse to cover the cost of fixing preventable errors, these mistakes can cause harm or even death to patients. Other preventable conditions include pressure ulcers from spending extended periods of time in bed and infections picked up from being exposed to pathogens in the hospital; these are often attributed to unnecessarily long lengths of stay.

Though the desire to improve patient care quality is widespread, "no healthcare organizations or practicing clinicians are immune from the simultaneous pressures to improve quality and lower costs" (Toussaint). In addition to the already high financial costs of healthcare, "a clinician's insensitivity, a delayed operation, and a stress-increasing medical facility can be meaningful nonmonetary burdens to patients in their assessment of value" (Toussaint). Patient satisfaction is an extremely important factor in evaluating a clinic's care quality, and much of patients' perception of quality depends on how smooth the process flow is and how long they wait to be cared for (Clinic Physician, personal communication). In order to improve flow, eliminate mistakes, and increase patient satisfaction, there have been extensive applications of process improvement in healthcare in the last few years.

In many primary care facilities, patient flow issues may cause wait times to be longer than ideal. Increasing timely patient access and decreasing wait times would increase satisfaction for both patients and clinic employees. There are several key performance indicators that must be explored in order to get a better idea of important issues. Throughput is defined as the number of patients "processed within a given time" (Dictionary.com). When throughput is high, delays can easily develop. Additionally, there are two main types of service delays within primary care. The first is appointment delays, or the length of time between the requested and scheduled appointment dates (Bard et al., 2014). A long appointment delay suggests large-scale issues within a facility, encompassing problems with overall strategy and management of a practice. The second type of delay is real-time delays, or how long a patient waits, starting at the scheduled appointment time and ending at the time they receive service. While appointment delays are measured in days, the latter is measured in minutes or hours and focuses on shorter delays, caused by general service inefficiency, patient/provider earliness or lateness, mismatch between the average appointment duration and the scheduled time slot, and financial or insurance problems (Bard et al., 2014). This project focuses on real-time delays.

As was aforementioned, in recent years, lack of quality has become an important concern in healthcare. In a primary care setting, this can stem from several causes, including a low number of primary care physicians, tight schedules, a rise in administrative requirements and the delicate balance between several responsibilities required of physicians in teaching hospitals (Bodenheimer & Pham, 2010).

2.3 Lateness

Delays can affect more than just the flow of patients through the clinic. Lateness also has an effect on the quality of patient care. A 2008 study performed at a urology clinic in Chicago, Illinois compared the care that late patients received compared to the care received by patients who were early or on time for their appointments. Of the 306 outpatient visits the study team observed, 63.2% of the scheduled patients arrived on time. 22% of patients arrived after their scheduled appointment time, and 14.8% of patients failed to show up at all (Onisuru et al, 2008). The researchers measured time spent waiting in the exam room for the physician, time spent with the physician in the exam room, and exam room utilization for the two groups of patients. A comparison of the two groups can be found in Table 1:

Table 1: Early vs. late patients (Onisuru et al, 2008)

Performance Metric	On Time Patients	Late Patients
Wait time in exam room (minutes)	14.8 ± 9.2	11.0 ± 8.4
Time spent with physician (minutes)	10.7 ± 6.0	8.9 ± 5.8
Exam room utilization	0.31	0.37

The data in the above table shows that patients who arrived on time waited, on average, about four minutes longer to see a physician than patients who arrived late. In general, though, on-time patients were able to spend about two more minutes with their physician (Onisuru et al, 2008). While two extra minutes may seem insignificant at first glance, those two minutes could be the difference between a patient being satisfied with his or her visit and the patient feeling that their appointment was rushed or impersonal.

According to a primary care physician at the clinic, one late arrival can have an enormous effect on the clinic's schedule for the rest of the day (Clinic Physician, personal communication). To illustrate his point, he used the example of a patient checking in 15 minutes late for an appointment scheduled to last 20 minutes. Patients who arrive late are given the option to wait for their physician or see another physician who may be available sooner. The dilemma this physician would face if the late patient requested to wait for him is which patient to see first: the one who arrived late, or the on-time patient who is scheduled for the 20-minute slot starting in five minutes. He could also choose to reduce the appointment duration for one or both patients in order to lessen the effect on his schedule (Clinic Physician, personal communication). In either case, one or both patients may experience unsatisfactory levels of waiting, or they may be subjected to a shorter, possibly rushed, appointment with their physician.

While it is often a major contributor, patient lateness is not the only reason for delays in primary care. If a physician starts running behind schedule (due to an appointment lasting longer than scheduled, or something else entirely out of his control), the rest of his on-time patients for the remainder of the day will suffer as well (Clinic Physician, personal communication). This delay would have the same effect on the remaining appointments as a late arrival would. Other things that can contribute to real-time delays include insufficient support staff, long check-in lines and complex patients – patients whose aliments are more acute or complicated than expected.

2.4 Lean Process Improvement

Lean is a commonly used continuous process improvement method that can be effective when applied to healthcare. A Lean system is free of waste and produces exactly what its customers want at exactly the time they want it (Toussaint). In a healthcare setting, Lean involves reducing seven kinds of waste: confusion, motion, waiting, overprocessing, inventory, defects, and overproduction. Lean processes strive to create products that have value, defined according to the wants and needs of the customer, by streamlining processes and reducing activities that do not directly add value to the customer's product. The underlying idea in Lean culture is that small changes over time will eventually compound and result in a major improvement in the process. Lean also encourages every member of an organization to play an active role in problem solving and continuous process improvement, placing an "emphasis on finding solutions instead of assigning blame" and "focusing on the problem; not the person," allowing every member to contribute to the growth and success of an organization. From a healthcare perspective, Lean means that each patient receives timely, appropriate care that is customized to his or her specific needs. This is especially the case in primary care, which is often an ill or injured patient's first step for seeking treatment (Jimmerson, 2010).

2.5 Lean Case Studies

The desire for higher quality care and patient satisfaction has led many healthcare facilities, from emergency rooms to small family practices, to reconsider existing processes and make improvements. Studies have been performed in various clinics across the nation, using a wide variety of techniques to measure and improve healthcare processes. This section examines two studies conducted in primary care clinics, focusing on patient flow and scheduling.

2.5.1 Example Case 1

A study published in July of 2011 involved three different Federally Qualified Health Centers (FQHC) in low-income neighborhoods throughout the continental United States. At each of the centers, the researchers from the Altarum Institute organized three one-day events to apply Lean thinking to the primary care clinics: the first to map the current state, the second to determine an ideal future state, and the third to create a to-do list and an action plan to move from the current state to the ideal future state (Graig & Perosino, 2011).

The study identified that the main issues at hand in the clinics included patient flow, the registration and vaccine processes, cycle time, and high volume of walk-in patients. To address these problems and arrive at the ideal future state, each task on the to-do list (varying from simple things like purchasing equipment to more complicated tasks like creating a new check-in and check-out process) was assigned to a specific person and given a deadline. The team participated in monthly check-ins with a Lean coach to discuss the progress being made and to keep the project moving successfully toward the goal. Outside of these monthly meetings, clinic staff received additional Lean trainings. In each of the three cases, several key problems were resolved, saving time, money, and frustration for the patients and the staff (Graig & Perosino, 2011).

2.5.2 Example Case 2

A second example comes from a simulation study performed on the Family Health Center (FHC) in San Antonio, Texas. The center is made up of about a dozen individual clinics, each with its own specialty, but the study focuses on the family health continuity clinic only. In this clinic, patients register and check in upon arrival. Then, their vital signs are checked by a medical assistant (MA) before they are seen by their provider, either an attending physician or a medical student working under the supervision of a physician (Bard et al., 2014).

The study began with a period of observation to establish a baseline to which future results could be compared. At the start of the study, patients waited, on average, 18 minutes to get their vital signs checked by a nurse and an additional 27 minutes to see a provider. The average total elapsed time, starting at the patient's scheduled appointment time and ending when they left the clinic, was about 80 minutes. This number does not take into account any time that a patient waited before their scheduled appointment time (if they arrived early). In the initial state, more than half of the time a patient spent in the clinic was spent waiting, and patients waited much longer to see a provider than for anything else. In this case, the provider was the bottleneck. Utilization rates for providers in the initial state were above 90%, much higher than the rate for any other resource. It was also observed that 21.8% of scheduled patients did not show up for their appointments (Bard et al., 2014).

After a baseline was established, the study entered a simulation phase. A detailed model of the clinic was created using Arena, and three sets of alternate scenarios were compared to the baseline model. The researchers tested a number of revised appointment schedules, including adjusting the length of scheduled appointment blocks and the distribution of patients in each block. When compared to the baseline, the revised schedules did not have a significant impact on the process, due to the fact that the provider was the bottleneck in the process. The second alternative tested involved reducing the no-show rate by 50%. It was found that this had a negative effect on the process, increasing the average length of stay by 14%. Completely eliminating no-shows would require many hours' worth of overtime in order to clear all the patients out of the waiting room. This dramatic effect is because patients who do not show up for their appointments ease the burden on the clinic and reduce the number of patients a provider must see. The baseline model had enough delays present as it was, and adding more patients only strained the clinic more. The scenario that proved to be the most effective in improving throughput and reducing patient wait times was a reduction in the simulated time spent with the provider to 75% of the observed time. Reducing the variation in provider time had a minimally significant effect. This simulation study illustrates that many of the delays in a primary care facility are a result of a bottleneck in the providers, and that the most effective tactic to improve patient flow is to reduce the amount of time patients spend with providers so that the time spent with providers matches more closely with the scheduled appointment duration (Bard et al., 2014).

2.5.3 Summary of Key Results from Example Cases

The two studies above show the variety of opportunities and Lean solutions available to the healthcare industry. Below is a summary of the methods and results of the two studies.

Table 2: Summary of results from example cases (Bard et al, 2014; Graig & Perosino, 2011)

Problem	**Method(s)**	**Results**
Inefficient vaccine administration process	• Standardized step-by-step protocol and color-coded vaccine storage bins (Graig & Perosino, 2011)	• Streamlined vaccine process (Graig & Perosino, 2011) • Saved time with associated improvement in patient flow (Graig & Perosino, 2011) • Reduced waste (Graig & Perosino, 2011) • Improved staff (Graig & Perosino, 2011)
Poor communication between care teams during patient visit	• Visual management - installation of color-coded flags outside each patient exam room (Graig & Perosino, 2011)	• Enhanced staff communication and reduced interruptions (Graig & Perosino, 2011) • Improved cycle time (Graig & Perosino, 2011)
Patients failing to arrive for scheduled appointments (no-shows)	• Revised scheduling - book extra patients to offset those who are no-shows (Graig & Perosino, 2011) • Simulated 50% reduction in number of no-shows, with no change to scheduling (Bard et al, 2014)	• More appointment slots filled (Graig & Perosino, 2011) • Improved access to care (Graig & Perosino, 2011) • Decrease in throughput (Bard et al, 2014)
Provider bottleneck	• Simulated reduction in provider length and variation to 75% of original levels (Bard et al, 2014)	• Reduction in measured length of patient visit (Bard et al, 2014)

These two studies demonstrate that there are many factors that affect the length and quality of a patient's visit to their primary care clinic. They also illustrate that there are multiple ways to successfully improve a clinic's performance – both Lean techniques (such as process mapping) and simulation studies have been applied effectively in situations similar to the one this project focused on.

3.0 The Primary Care Clinic

This project focuses on a large primary care clinic that is part of an integrated health system. The system offers medical services in a number of areas, including emergency medicine, inpatient and outpatient care, hospice, homecare and rehabilitation.

3.1 Process Flow in the Primary Care Clinic

This large primary care clinic serves between 100 and 200 patients every day, including 30 new patients each week. The staff consists of 26 attending physicians, who supervise about 30 resident medical students, 8 Patient Care Associates (PCAs) and 2 nurses, as well as several support staff including telephone operators, administrative assistants, and registration desk staff. These support staff perform a variety of functions which encompass checking patients in and out, booking appointments, and other administrative tasks.

When patients arrive at the clinic, they proceed to the registration desk to check in for their appointment and ensure that the clinic's records are up-to-date. Upon completion of the registration process, the front desk marks the patient as "Arrived" in the scheduling system, which then flows into the electronic medical record (EMR) system. Depending on which exam rooms their physician is covering that day, the patient is sent to one of three sections of the waiting area. A Patient Care Associate (PCA) calls the patient out of the waiting room and checks his or her vital signs including height, weight, and blood pressure. The PCA also asks the patient some questions about their general health - for example, if they have any pain or if they smoke cigarettes. In many cases, a disconnect exists between the clinic's electronic and paper systems. For instance, at some of the PCA stations, a patient's vital signs can be logged electronically, and automatically flow into the EMR system with a timestamp. However, many PCAs choose to enter the information manually. In the rest of the clinic, inputting the information must be done manually, so the timestamp can be inaccurate by the difference between the time the PCA measures the patient's vital signs and the time they are inputted into the system.

After the vital signs check, if the patient's scheduled exam room is ready, the PCA escorts the patient there to wait for the provider. If the exam room is not yet ready, the patient is sent back to the waiting area until the room is prepared. Once in the exam room, the patient waits until their physician enters the room, at which point their appointment begins. Once patients have seen their physician, they proceed to the check-out desk. Some patients are then free to leave the clinic, but some will need to receive vaccinations or have a procedure done in the lab services department. Since vaccinations are generally not administered in the exam room, a patient who needs a shot must wait in the main waiting area until the designated room is available. If the patient needs a laboratory procedure, he or she proceeds to the lab services waiting area (adjacent to the primary care waiting areas) and waits for an available provider, who may be an attending physician, a nurse practitioner or a medical resident. Providers have busy schedules consisting of several responsibilities, one of which is seeing patients at the clinic. Members of the staff have expressed their dissatisfaction with the clinic's operations, citing issues such as constantly high activity levels and inconvenient placement of some patient services. There are several opportunities to reduce waiting in the clinic and improve the overall process flow.

A 2012 study of the primary care clinic conducted by local high school students revealed that waiting time made up a significant portion of overall time that patients spent at the clinic. Although their sample size was low, encompassing 17 patients or less at each step of the process, they measured the total time at 10 different process steps (patient parking, registration, waiting time, time spent with medical assistant, time waiting in the exam room prior to provider arrival, time spent with provider, check out time, time waiting for blood draw, time with phlebotomist, and the length of time from last stop to car). Overall, this yielded an average amount of time spent at the clinic of an hour and twelve minutes for a twenty minute appointment. Almost half (48.8%) of that average time was allocated to waiting, whether in the waiting room or the exam room. This data accounted for 6.05 minutes at registration, 13.5 minutes in the waiting area, 2.8 minutes with the medical assistant, and 21.3 minutes waiting in the exam room before the provider arrived. In total, patients spent an average of 43.5 minutes in the clinic prior to seeing the provider. Because the clinic considers the appointment to start when the patient is in the exam room with the provider, this data indicates that appointments are starting on average 43.5 minutes late, and also shows that only 5% of patients would start their appointments "on time". Because this study had a small sample size and is a few years old, the collected data cannot necessarily be used directly; however, this study is still helpful in providing an initial understanding of the magnitude of the lateness problem within the clinic. Additionally, it emphasizes the need to dig deeper into waiting times and patient flow within the clinic with the ultimate goal of uncovering the root causes of these issues and implementing effective solutions, instead of simply treating the symptoms of inefficient processes.

Figure 1 is a swim-lane diagram summarizing the appointment process, starting when the patient arrives at the clinic and ending when the appointment with the provider begins:

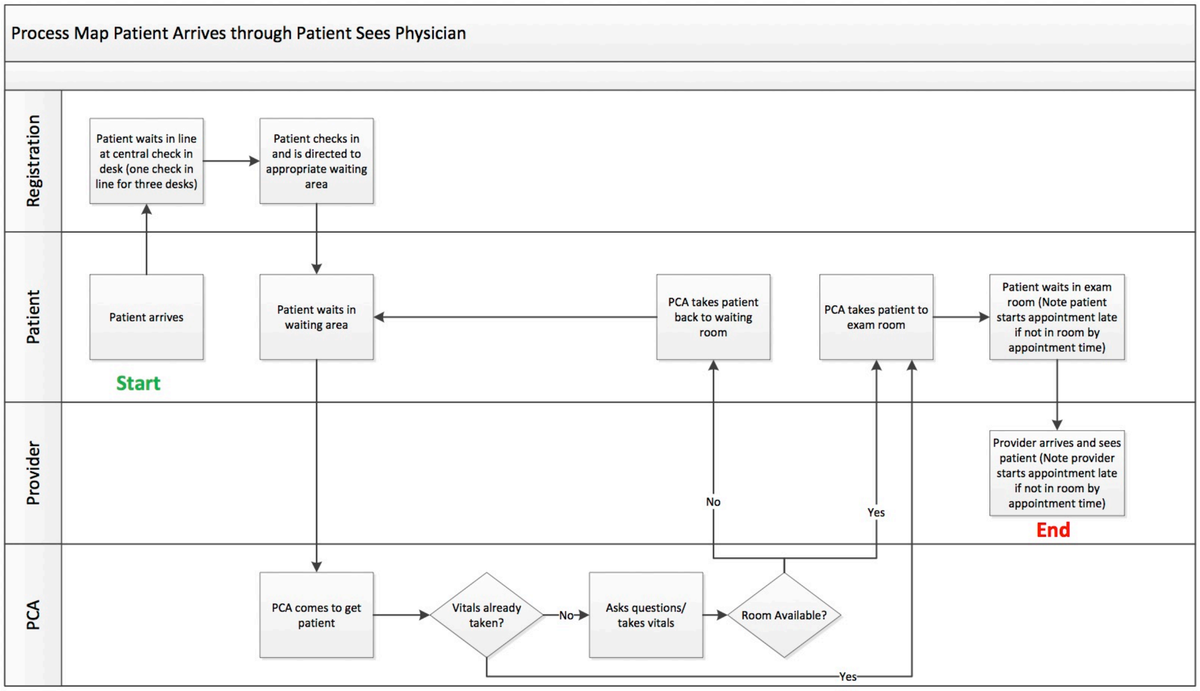


Figure 1: Swim lane diagram of flow through clinic

3.2 Process Improvement at the Integrated Health System

The primary care clinic that was the focus of this project strives to be a Lean environment. The clinic staff aims to continuously improve processes throughout the entire facility. One of the most widely used tools throughout the organization is the A3 report, a summary of the problem and the actions taken to analyze, correct, and monitor it. An A3 for this project, based on the organization's preferred A3 report style, can be found in Appendix A

The health system offers several levels of Lean training to its employees. The first level, the White Belt, is an introduction to Lean principles and to some popular tools used in Lean workplaces. Employees who have completed the White Belt program are aware of the main points of Lean, such as defining and identifying value and waste, the basics of root cause analysis, and the PDSA cycle. The next level, Yellow Belt, presents a more in-depth look at the A3 problem solving process. In addition to the training they provide, process improvement implements process improvement projects all throughout the organization.

4.0 Methodology

This project, focused on a large primary care clinic belonging to an integrated health system, aims to analyze lateness at the clinic, its effects on the staff and patients, and the way that the clinic responds to patients who are late or providers who are running behind schedule. In order to achieve this goal, the project team focused on the following objectives:

1. Collect qualitative and quantitative information about clinic operations
2. Develop a simulation model to represent the clinic and use it to test possible changes to the clinic's operations
3. Analyze the model outcomes and make recommendations to the clinic in order to improve performance.

The methods used to complete these objectives are detailed below.

During the first portion of the project (background research and data collection/analysis), the project team consisted of Sarah Abell and Amy Stevens (the author of this report). For the second portion of the project (simulation modeling and analysis), Amy was solely responsible for the project.

4.1 Objective 1: Collect Qualitative and Quantitative Data about Clinic Operations

4.1.1 Interviews and Observations

The first step in any Lean process improvement initiative is to observe the system in question and speak to the people who know it best. It is crucial to gain an understanding of the process before attempting to make changes. At the beginning of the project, we attended several meetings with two attending physicians in the clinic and a process improvement specialist employed by the health system. They gave us a tour of the clinic and provided us with previous studies of the flow within the clinic and other materials, including a map of the clinic and role descriptions for the different jobs within the clinic. The data allowed us to establish a baseline of knowledge about the layout and inner workings of the clinic. We used Microsoft Visio to compile this knowledge into a swimlane diagram (see Figure 1 in Chapter 3), which depicts the steps a typical patient would follow during their time in the clinic.

Next, we spent time at the clinic observing the process in action. During this period, we were primarily located in one of the areas in which the Patient Care Associates (PCAs) take patient vital signs, and thus, were able to watch the process from when the patient is taken from the waiting room for vitals to their placement in an exam room. We also had the opportunity to speak with the PCAs, registration desk staff and Ambulatory Service Representatives to gather their concerns, preliminary ideas for improvement, and additional information that would allow us to better understand the current state. At this point, we were able to fine-tune our swimlane diagram.

4.1.2 Timestamp Data

Next, we were provided with data pulled directly from the EMR system that contained information about patient arrivals. The data covered the period starting on October 1, 2014 and ending on October 8, 2015. For each appointment during that time period, the data included the scheduled date and time, the provider, and the time that the patient completed the registration process. We referred to this time as the patient's arrival time, because at this point, the patient is marked as "arrived" in the EMR system and is ready to continue with the next step of their appointment. Using Microsoft Excel and Sigma XL, we organized the data into pivot tables and made control charts that depicted the difference between patient appointment and arrival time, and quantified lateness. We also analyzed the data in terms of what percentage of patients were late to appointments and broke down lateness by physician and appointment type. In this way, we discovered that there was an unusually high incidence of lateness in patients coming to the clinic to receive a flu vaccination (approximately 70%). Further investigation on this extremely high number showed that appointments labeled "flu visit" are actually part of a walk-in flu clinic. Because of the walk-in nature of the flu clinic, patients do not have an assigned appointment time, but a time is listed in the record in the EMR system. For this reason, we excluded flu clinic patients from our data analysis.

4.1.3 Manual Data Collection

At this point, most of the data and information we had was qualitative, and the majority of the quantitative information we had did not cover the entire process – it was concentrated on the arrival and registration parts of the process. In order to get a more complete quantitative picture of how the clinic functioned, we decided to collect data manually about the rest of the process. We chose to collect data from both the patient perspective and the physician perspective, because lateness is a problem that can originate from many sources. Patients' arrival times can affect how the clinic functions, but physicians' actions can also contribute to the clinic being behind schedule. After some preparation, our data collection plan was approved by WPI's

Institutional Review Board (IRB). Our data collection period consisted of five days: each weekday between Wednesday, November 18th and Tuesday, November 24th, 2015.

Patient Data Collection

Because we wanted to track a large number of appointments, and based on both our own observations and a review of some data collected by high school students in 2012, we understood how difficult and time consuming it would be to follow individual patients through their appointments. We concluded that having the patient track the appointment themselves would provide satisfactory data. This method yielded significantly more data than if we were to shadow patients through the whole process, and it gave us baselines for each step of the appointment process, starting at the registration line and ending when the patient's time with his or her provider came to an end.

During the data collection period, one or both of us was stationed at the clinic entrance between 7:30 AM and 4:30 PM, distributing blank sheets to patients who entered the clinic. We explained to each one that we were performing a study with the end goal of reducing wait times in the clinic and asked if they would be willing to anonymously track their progress through their appointment by writing down what time each process step started and ended. The form we distributed can be found in Appendix B. There were some time periods during which neither of us was able to be in the clinic, and the clinic staff helped to fill in these gaps as much as possible.

There are several significant limitations to collecting data this way – first and foremost, we were unable to capture every patient who walked into the clinic, for various reasons. During the times when there was nobody distributing forms, which in many cases corresponded to the highest-volume periods of the day, no patient data was captured. At some points, patients were entering the clinic at a faster rate than we could approach them. Additionally, some patients simply ignored us or refused to participate in the study. The second limitation of this sort of data collection is that it relies entirely on self-reported data from patients and is completely optional. Since we could not ensure that every patient we approached would participate fully in the study, our data may not be an entirely accurate snapshot of the clinic's flow. However, we believe that it provides a good initial qualitative understanding of the way the clinic operates.

As a result of these two limitations, the data sheets we received at the end of the collection period had varying degrees of completeness. Because it was completed manually and there are often many other things going on during an appointment, some of the forms have fields left blank. This may have resulted from confusion, forgetfulness, distractions or lack of a timekeeping device. Some patients also left unfinished sheets in the waiting room or the exam room, either because they forgot about them or because they decided to end their participation in the study.

Another limiting factor is the demographics of the clinic's patients. Because the clinic is located in an urban area, not all of its patients speak English fluently. Patients who have limited proficiency in English often use interpreters during their appointments so that they can communicate effectively with clinic staff. Using an interpreter essentially doubles the length of an appointment, since everything that is said must be translated and repeated by the interpreter. Such long appointments could be a major contributor to delays in the clinic since patients who need interpreters are not scheduled in longer appointment slots than their English-speaking peers. We contacted the health system's interpreter services department and were able to procure two translated versions of our data sheet – one in Spanish and one in Vietnamese. This proved to be more challenging than we expected because the patients who needed these translated sheets the most were the ones that we could not communicate well with. We suspect that some patient data was lost due to the language barrier.

More than a quarter of the patients who had scheduled appointments during that week participated in the study, but the actual rate of participation is higher when the fact that not every appointment that was scheduled actually occurred (due to cancellations or no-shows) was taken into consideration. A total of 314 patients participated in the study. We filtered out appointments with residents as well as urgent care and psychology, because those appointments were outside of the scope of our project. As can be seen in Figure 2, 196 of 280 in-scope participants filled out all of the fields on the form. An additional 24 failed to fill in physician name, date and/or appointment time; 49 were missing at least one timestamp data point. Eleven more were double counted because patients were missing both a data point and a time/date/physician name field, so ultimately 70% (196) of the forms were completely filled out.

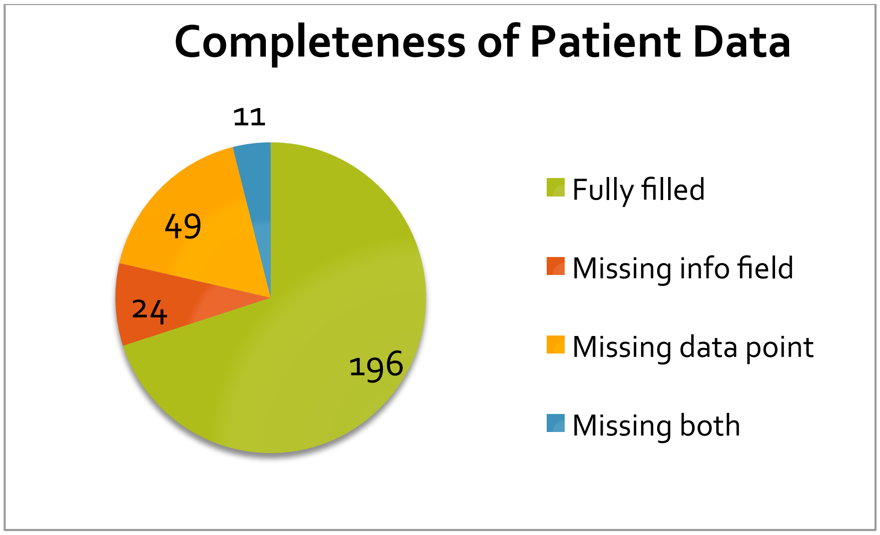


Figure 2: Completeness of patient data sheets

Physician Data Collection

In addition to the data recorded by patients at the clinic, we also asked physicians to record some data from their perspective. Each willing physician started with a copy of his or her schedule for the day – listing each appointment's scheduled start time, scheduled duration, and appointment type. As they went about their day, the physicians were asked to record what time each appointment actually started and what time it actually ended, as well as any relevant notes that might explain why an appointment started late or lasted longer than it was scheduled to.

The main limitations of this data collection was similar to that of the patients – completeness and credibility. Some of the sheets that were returned had spaces left blank or were missing important information about the appointment (such as its scheduled start time).

Thirty-seven providers of various types (attending physicians, medical residents, and nurse practitioners) participated in some capacity during the data collection period. Some provided us with data from every day during the period that they were working in the clinic, while others only gave us data from one clinic session. In total, 1,028 appointments were recorded by providers in the clinic.

Data Analysis

After entering all of the data that we collected into Excel, we began to closely examine the data to ensure that it was usable for the project. The most significant part of this was to compare the patient data to the provider data to make sure that there was some degree of consistency between the two sets of data. Since the commonality between the physician and patient data was the physician name, date, and scheduled appointment time, the goal of this exercise was to determine if providers and patients reported similar appointment lengths. Beginning with the 314 patient data sheets and 1,028 physician appointments, we first sorted out all residents in addition to urgent care and psychology appointments. Then, we sorted out any form with blanks in the information that we were using to compare (appointment date/time, physician name, and start and end time). After this step, we were left with 221 patient sheets and 605 physician appointments. Of these, only 56 appointments matched (when looking at date, time and physician name). We then calculated the appointment length for each, and compared them. If the difference in appointment lengths was less than or equal to five minutes, any difference was attributed to a difference in timekeeping device or an acceptable level of forgetfulness in marking down the time. If the difference was greater than five minutes, it was considered a bad match. Ultimately, 71% of the matched appointments were "good" matches and 29% were "bad" matches. This exemplified that although the sample size was low, there was some degree of accuracy between the two data sets. Because we had no way of knowing which of any set of matching data points were accurate, or if the points that we had no way of cross referencing were accurate, we did not throw any points out of the study based on this examination.

4.2 Objective 2: Develop a Simulation Model to Represent the Clinic

After discussions with the sponsoring physicians about how best to approach the problem we faced, we decided to create a simulation model using Arena. A simulation model allows investigators to experiment with changes that may be costly or organizationally difficult to implement, rather than testing changes on the physical system itself. It also allows us to be confident in the recommendations we make to the clinic – we can be reasonably sure that they will have the effect we intend. Because we already had a significant amount of quantitative data about the current state of the clinic's operations, it would be relatively straightforward to use it to create a computer model of the clinic. Data collected from the patient data sheets was used to create and validate the model, in conjunction with the data from the scheduling and EMR systems for the specific days we collected patient data. A detailed description of the model and the validation process can be found in Chapter 6.

4.3 Objective 3: Explore Opportunities for Improvement

Once the base case model was complete, three main areas for improvement were tested, based on discussion with the sponsoring physicians. More details about the alternatives and the results of simulation modeling can be found in Chapter 6. The areas of improvement that were tested include:

- Changes to patient arrival patterns
- Increasing the number of available PCAs in the clinic
- Reducing the gap between scheduled appointment length and the actual length observed in the clinic

In addition, the policy of not processing patients who arrived to the clinic after a certain lateness threshold was tested; in particular, by examining the effects of having patients arriving more than 15 minutes late reschedule their appointments for another day. Policies like this are not uncommon in clinics in the United States – and are even used in some subspecialty clinics in the health system.

Alternate models were created for each of the scenarios, and compared to the base case model using two calculated statistics – the average difference between the scheduled appointment start time and the actual start time, and the patients' total time spent in the system (from the time they finished registration to the time their appointment with their provider ended). Independent t-tests were to determine if the alternate models caused a statistically significant change in either of the statistics. After analyzing these results, I compiled a set of recommendations, which are discussed in detail in Chapter 7.

5.0 Results and Findings from Data Collection

This chapter summarizes the findings from the clinic data we analyzed. This includes both the data from the scheduling system that we received, as well as the data that we collected ourselves in November 2015.

5.1 The Majority of Patients Arrive Before their Appointment Time

Between October 1, 2014 and October 8, 2015, the primary care clinic served about 51,000 patients. Of these patients, 31% completed the registration process (and were marked as "arrived" in the clinic's computer systems) after their scheduled appointment time. (Clinic Data, October 2015). While about two thirds of patients arrived early to their appointments, in most cases, patients who arrive extremely early cannot be accommodated before their appointment times due to other patients in the clinic. The majority of patients (about 55%) arrived within 15 minutes of their appointment time (on either side). It is noteworthy that only 6% of patients (about 3,700 over the entire year) arrived more than 15 minutes late (Clinic Data, October 2015). Figures 3 and 4 below represent this arrivals data visually in the form of pie charts.

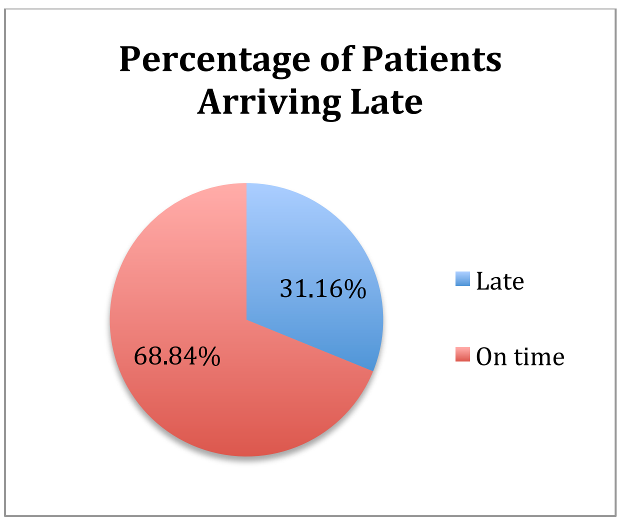


Figure 3: Percentage of late arrivals

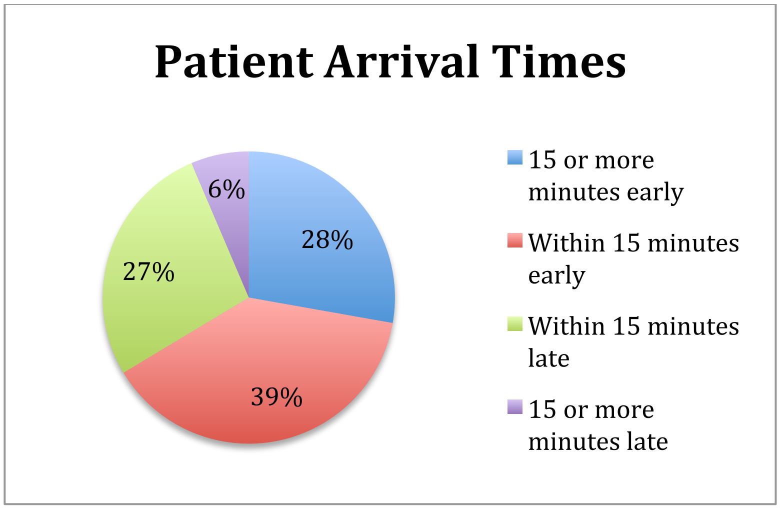


Figure 4: Detailed pie chart of patient arrival times

While late arrivals to the clinic can cause severe flow issues within the clinic, there are also large numbers of very early patients, so the average time difference between arrival time and appointment time is -11 minutes, or 11 minutes early to an appointment. In this case, an average is not the ideal metric to use because an overall average does not reflect the variation in arrivals to the clinic – though the average patient arrives early, there is significant variation in the data that can have an effect on the process. A histogram, such as the one shown in Figure 5, can provide a more accurate depiction of the range of early and late arrivals within the primary care clinic. In the histogram, 0 (represented by a vertical red line) represents that a patient registered exactly at his or her appointment time. Negative numbers (on the left side of the red line) indicate that the patient registered before his or her appointment, and positive numbers (on the right side of the line) represent patients who registered after their appointment times.

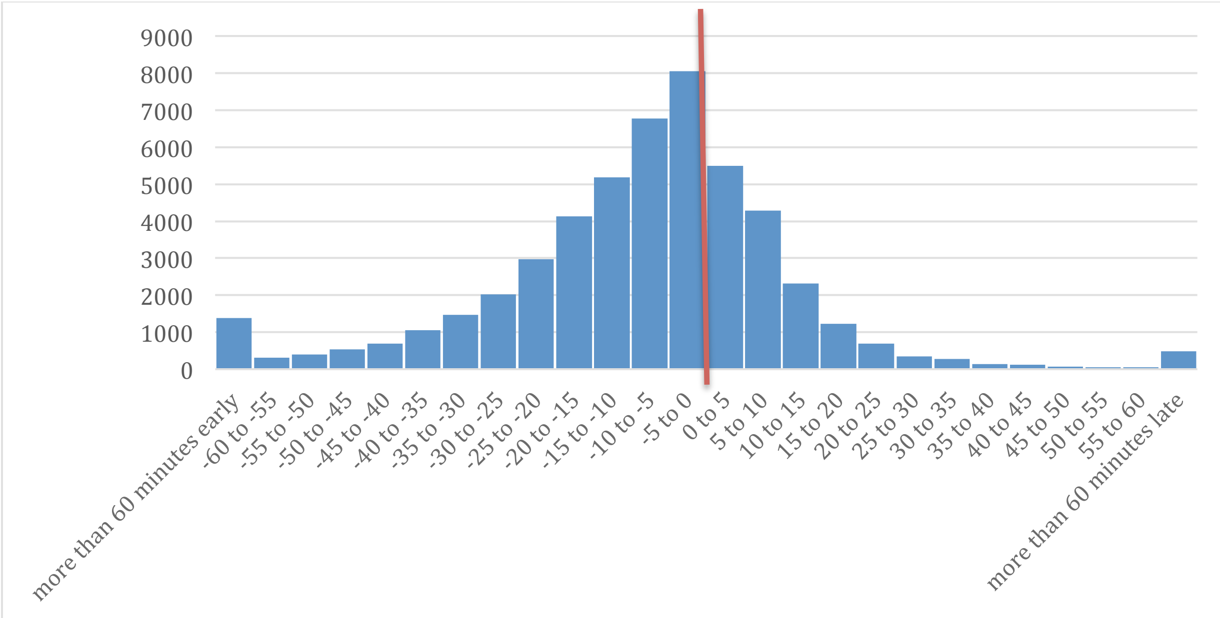


Figure 5: Patient arrival times from scheduling system data

The histogram in Figure 5 depicts that arrivals to the clinic follow an approximately normal distribution, with the peak of the distribution just before the appointment and a considerable amount of variation in arrival times relative to appointment times. The majority of patients arrive early for their appointments, and most of those who arrive late are not extremely late.

5.2 More than Two-thirds of Appointments Start Late

This section examines the proportion of appointments in the clinic that start after their scheduled start times. It explores the difference between the proportion of patients arriving late and the proportion of appointments starting late, using data collected by physicians and by timestamps in the EMR system. It also examines the proportions from a patient perspective, and discusses the reasons an appointment may begin late, based on comments from physicians.

5.2.1 Lateness throughout the Process

Although most patients arrive early, many patients do not get their vital signs measured before their appointment time and even fewer appointments with physicians actually start on time. The histogram in Figure 6 shows the distribution of patient arrival (registration) times compared to appointment times during our data collection period. The red bar is positioned at 0 –

anything to the left of this bar (a negative number) indicates an early arrival, and anything to the right (a positive number) indicates a late arrival.

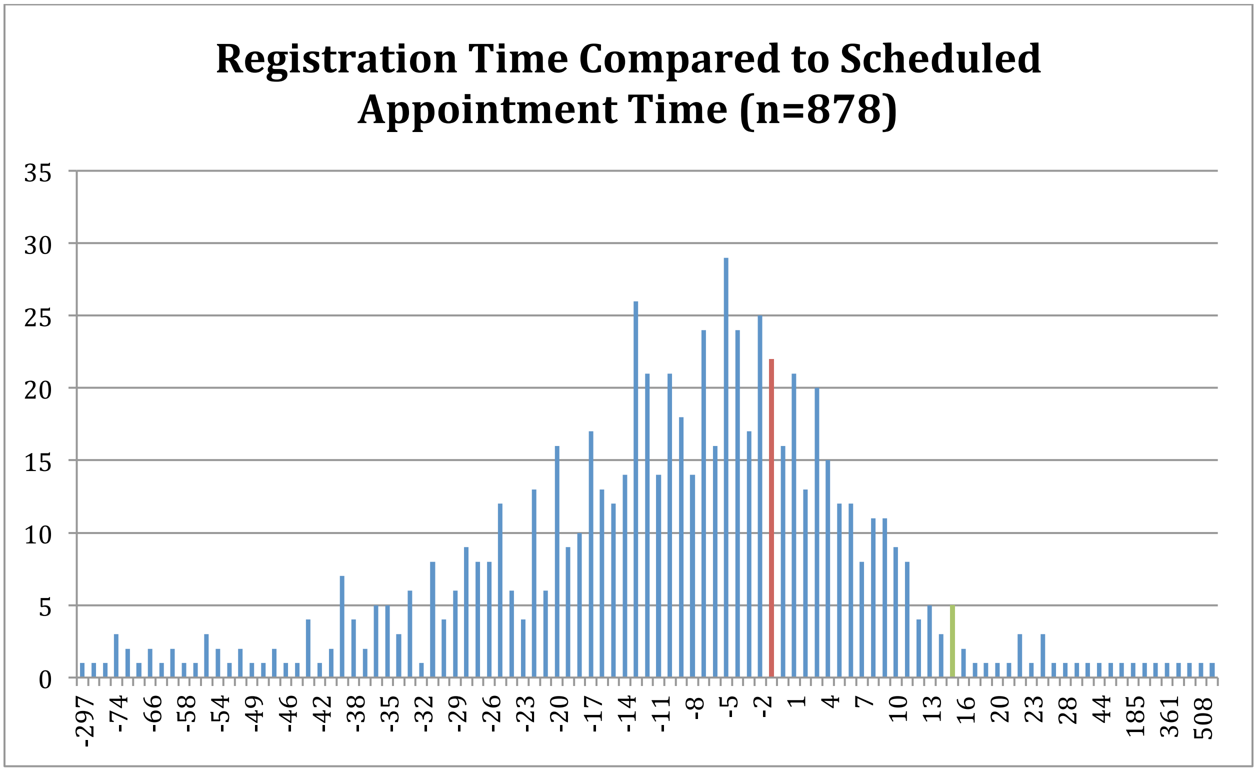


Figure 6: Histogram of registration time minus appointment time

This distribution is consistent with the distribution from the scheduling system data from October 2014 – October 2015 (discussed in Finding #1). It shows that the majority of patients during our data collection period arrived before their appointment times, with very few patients arriving more than 15 minutes after their scheduled appointment time (represented by the green bar).

Figure 7 displays the difference between the time a patient's vital signs were measured and his or her scheduled appointment start time. Again, the red bar represents a difference of 0. As above, points to the left of the red line represent that the vital signs step happened before the scheduled appointment time, and points to the right of the red line represent that it was completed after the scheduled appointment time.

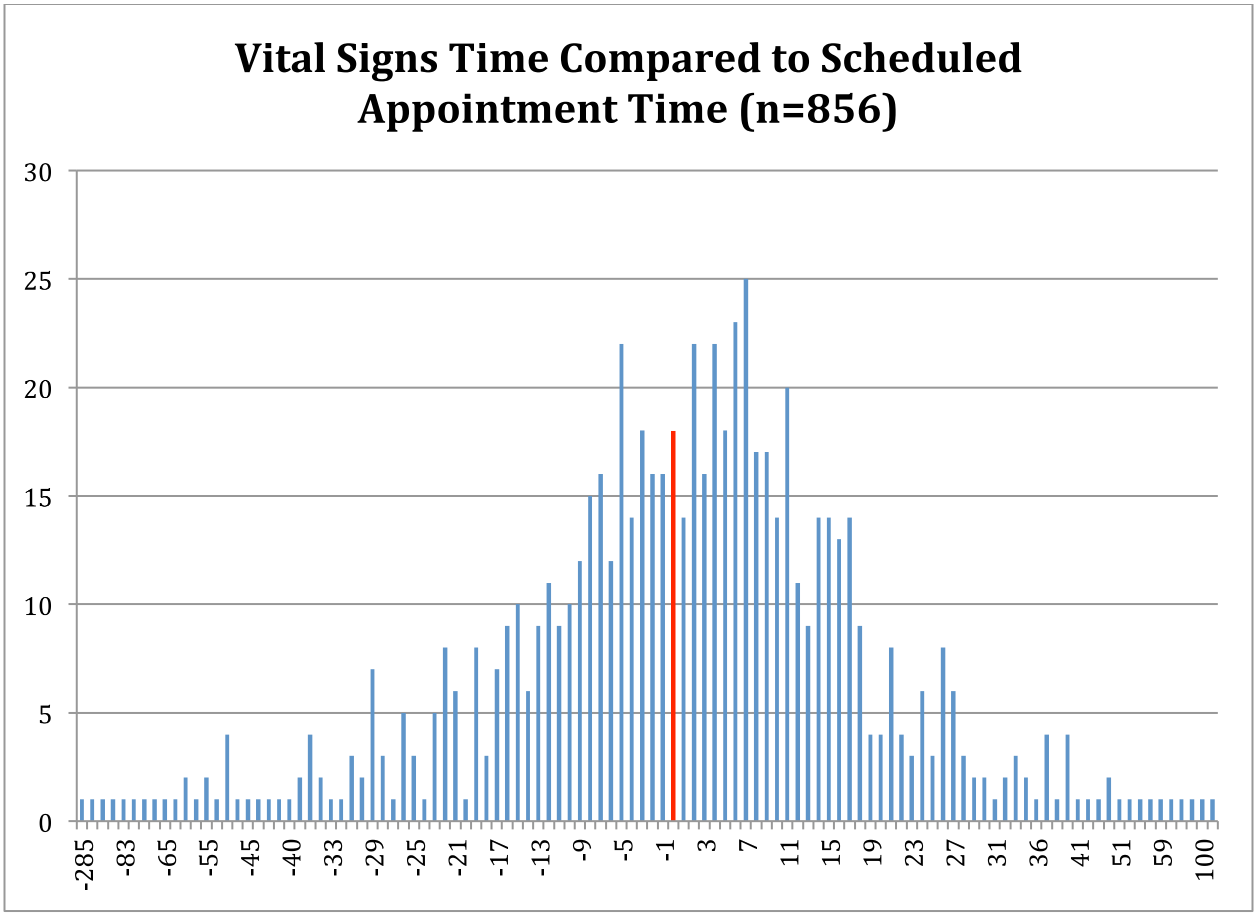


Figure 7: Histogram of vital signs time minus appointment time

When compared to Figure 6, it can be seen that a higher percentage (appearing to be roughly half) of patients complete the vital signs step after their appointment start time. The most probable cause for this difference is believed to be insufficient PCA staff. The PCAs are responsible for measuring patients' vital signs, performing certain tests requested by physicians or residents, and also for cleaning exam rooms when appointments end. All of their responsibilities mean that they are usually very busy. The PCA staff that we spoke with mentioned that they were constantly moving around in the clinic performing various tasks, which may contribute to delays in this process step.

The histogram in Figure 8 shows the difference between the time the provider entered the exam room (considered to be the time the appointment started) and the appointment's scheduled start time. The trend continues – even fewer patients start this step on time.

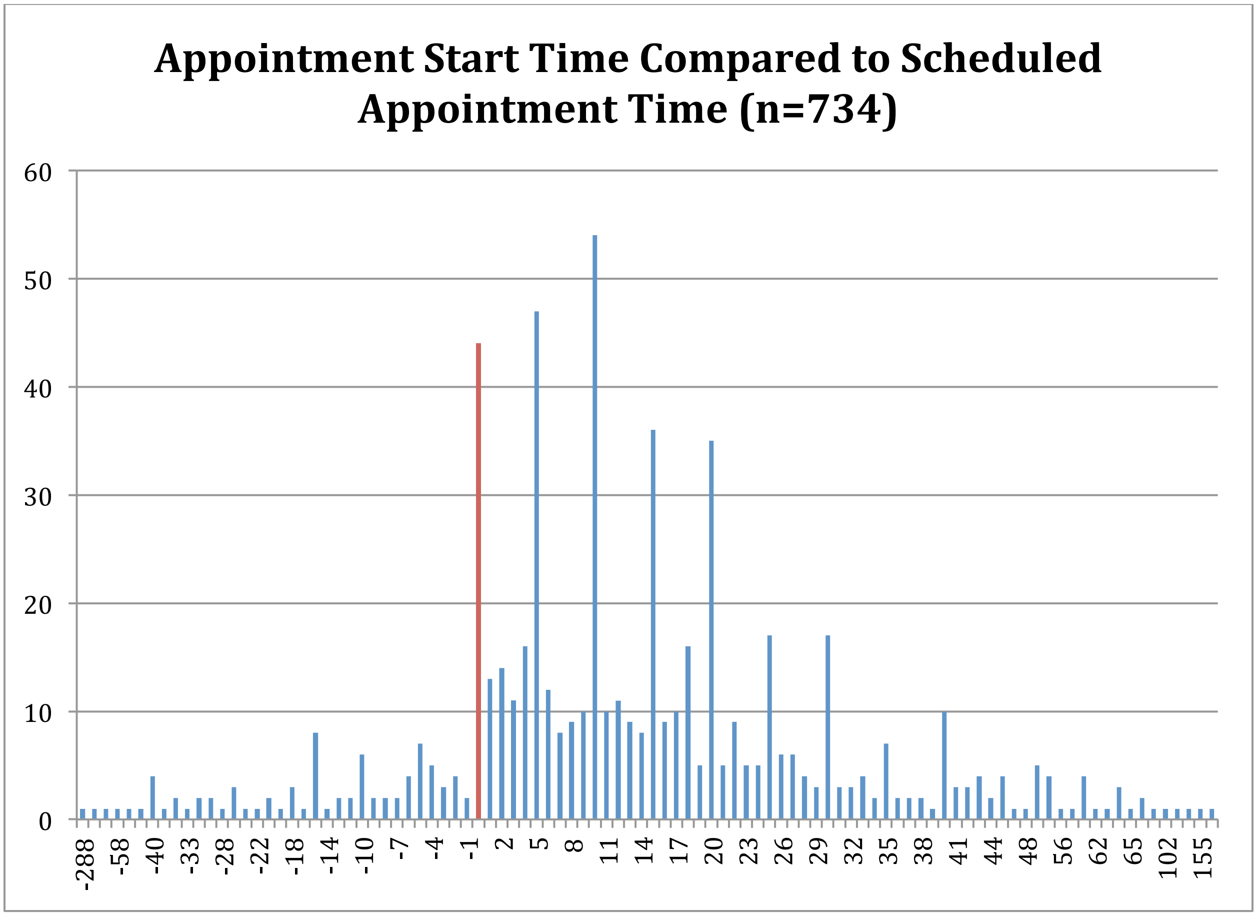


Figure 8: Histogram of appointment start time minus appointment time

Again, the red bar represents appointments that started exactly on time. The majority of appointments are to the right of the zero mark, which indicates that they started after they were scheduled to. This histogram has some very definite peaks on numbers that are divisible by 5. Our best guess about these peaks is that physicians rounded their recordings to the nearest 5 minutes (for example, if they entered an exam room at 9:08 AM, they might have written down 9:10 as an approximation). We concluded that these approximations would not impact our data significantly, so we did not investigate them further.

5.2.2 Consistency between Data Collected by Patients and Physicians

Next, we calculated the percentage of appointments started on time. Figure 9 depicts the timeliness of appointment starts based on both the patient (left) and physician (right) data. While the percentages are slightly different between the two samples, and the physician sample is nearly three times as large (605 patients compared to 221), it is clear that the majority of appointments are not starting at the scheduled time. From both sets of data, 15% of appointments started on time. From the patient data, a quarter of appointments started early or on time, while 35% of the appointments recorded by physicians started on time.

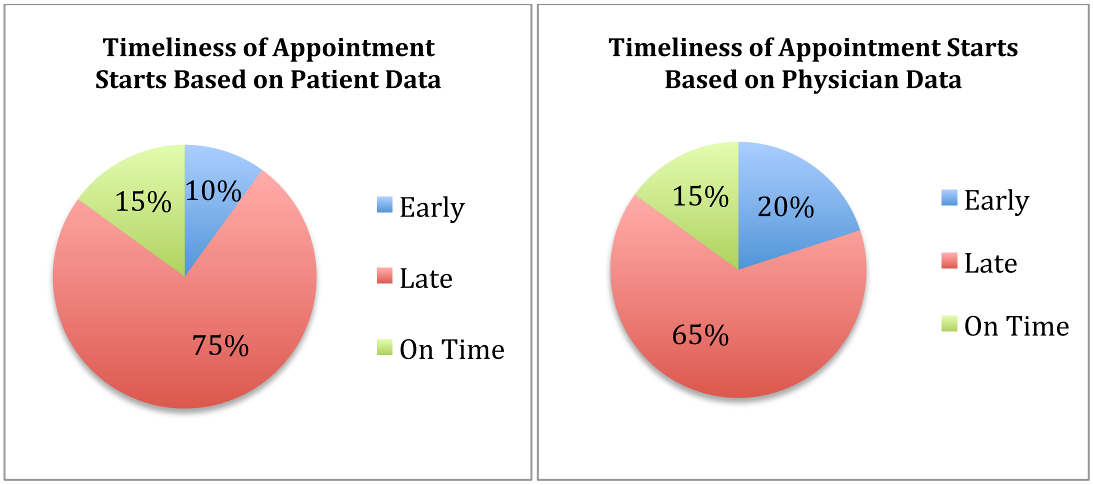


Figure 9: Timeliness of appointment starts

In these pie charts, “Early” means that the appointment started before it was scheduled to (For example, a 9:00 AM appointment that started at 8:58 AM). “On Time,” in this case, means that the appointment started close to when it was supposed to. After some discussion with the physicians in the clinic, we assigned a five minute cushion time (a 9:00 AM appointment that started at 9:03 AM would fit into this category). A “Late” appointment started sometime after that five minute buffer period (for a 9:00 AM appointment, this would encompass start times of 9:06 and beyond). These charts emphasize the discrepancy between the proportion of patients who arrive before their appointment time and the proportion of appointments that actually begin on time. The histogram in Figure 10 examines the appointment start times in more detail – organizing them in more categories depending on how early or late they began compared to their scheduled start time. It illustrates the variation in appointment start, and shows that the majority of appointments that begin late are beginning between 15 and 29 minutes late.

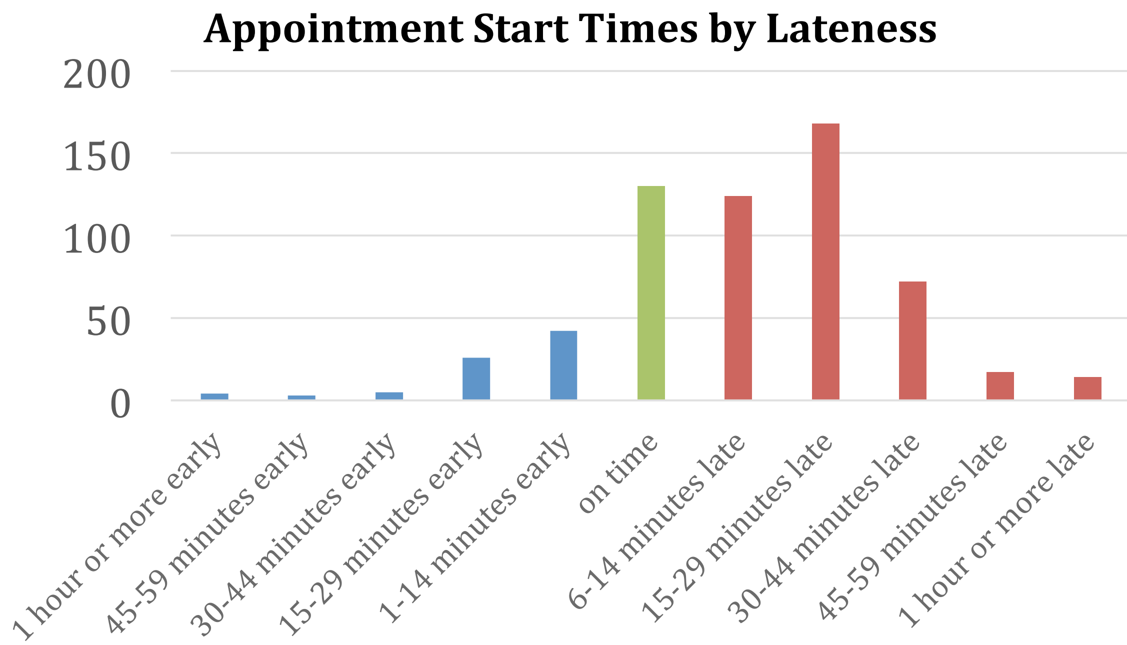


Figure 10: Evaluation of appointment starts

5.2.3 Reasons for Lateness

In addition to the timestamps, a vital portion of our manual data collection process was the comments physicians provided regarding why they may have been running behind schedule. One of the sponsoring physicians aided us by analyzing 236 legible comments written by physicians. She provided us with the chart in Figure 11. From her chart, we determined that events outside the physician's control cause the majority of delays. The top three most common reasons for falling behind schedule stemmed from a complex patient, delayed rooming, or a late patient arrival. Complex patients may include patients who required extensive discussion regarding new diagnoses or several medical conditions, or those who were booked in appointment slots that were shorter than they needed to discuss their issues. Delayed rooming could be caused by a PCA being unavailable to put a patient in a room because they were busy with something else, such as patient testing (EKG), or could also be a result of EMR system errors or patient lateness. The exact reason for a patient being put in an exam room later may be difficult to determine from a physician perspective, so the two categories could be difficult to tell apart.

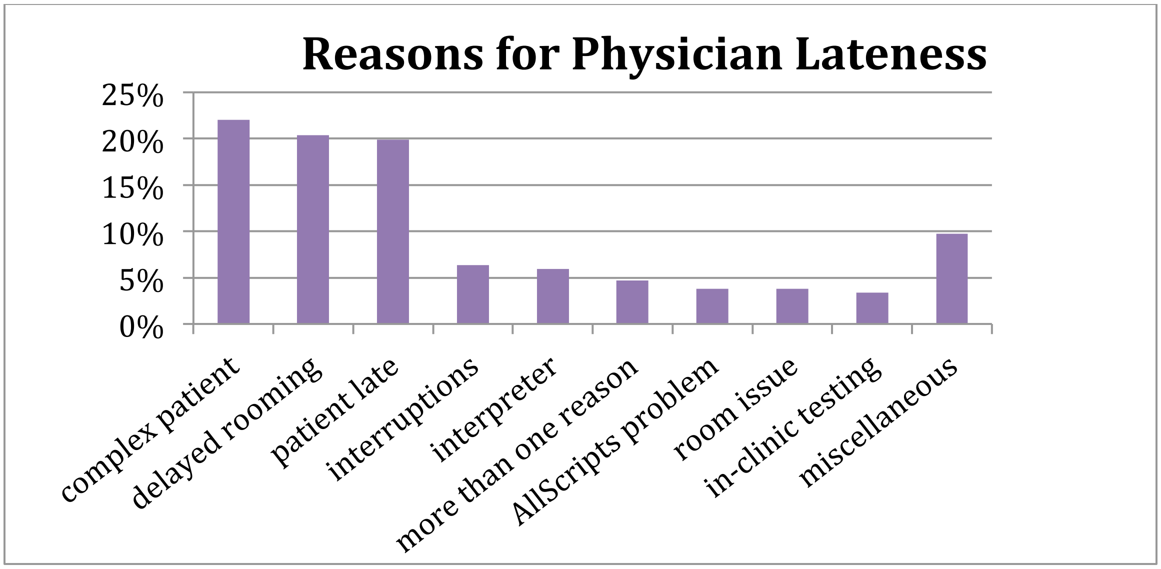


Figure 11: Reasons for patient lateness

Other delays were caused by interruptions (which could be anything from needing to leave the room to use a printer or a physician getting a personal phone call about a family emergency). Interpreters not being available right away, as well as the extra time required to converse through an interpreter, also contributed to lateness. Occasionally, issues with the EMR system also contributed to delays – in a few instances, patients' arrivals were not reflected in the system, but the patient was in the clinic waiting to be seen.

5.3 Appointments Contain a Large Amount of Non-value-added Time

One of the main underlying motivators of this project was to make changes that would maximize the value-added portion of the patient's visit. From a Lean perspective, value-added time consists of the activities a customer, in this case a patient, would willingly pay for. Figure 12 is a box-and-whisker plot that depicts how much of a patient's total time in the clinic (from getting in line at registration until the end of the appointment) is spent on value-added activities (defined here as the time spent on vital signs and the time with the provider, because these are the times when clinical activities are being performed). It is important to note that the patient and the clinic may have different perceptions of value-added time. For example, the patient may not see the vital signs process as a valuable part of his or her appointment, but the clinic perceives it as a valuable use of time that benefits the patient.

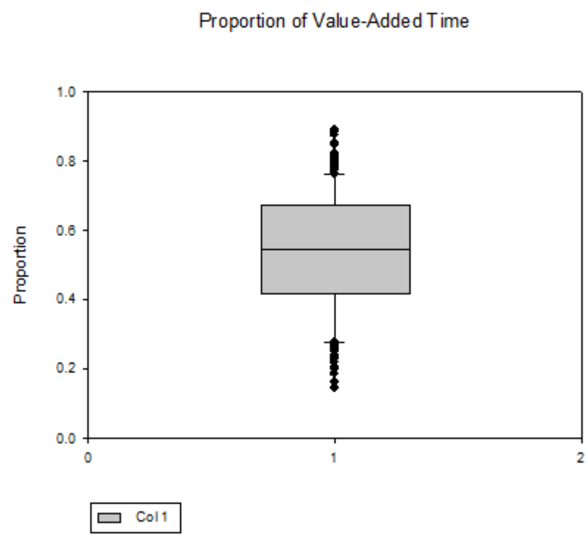


Figure 12: Box plot of proportion of value-added time

In Figure 12, the middle 50% of patients (contained in the box) spend between 40% and 67% of their total appointment time on value-added activities. The best-case scenario entailed spending approximately 77% of one's appointment on value-added activities, and the worst-case scenario involved less than 20% value-added time. We did not however, track the length of check-out, lab services or additional tests in our study, which would have changed the percentages. Additionally, there are some business non-value added activities (such as registration and travel time between stations) that cannot be completely eliminated from the process, even if the patient would not consider them completely necessary or valuable. Activities like this would be classified in the non-value added category, even though they are necessary for the clinic's operation.

To break the data down further, we looked at the proportion of the patient's time in the clinic (again, from registration line to physician leaving exam room) that was spent with the provider. From Figure 13, below, it can be seen that patients spent up to 70% of their appointment with their provider. However, the middle 50% of patients only spent between 35% and 49% of their total appointment time with their physician, and some patients spent very little time with their physician.

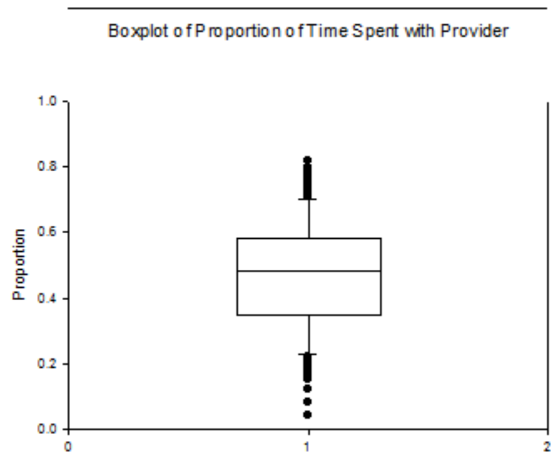


Figure 13: Box plot of proportion of time spent with provider

Finally, we evaluated how much of a patient's total time in exam room is spent waiting for his or her provider. This is depicted below in Figure 14. We found that the middle 50% of patients spend between 10% and 40% of their total time in the exam room (time waiting plus time spent with physician) waiting for their physician arrival. Most patients spend no more than 55% of their exam room time waiting for their physician.

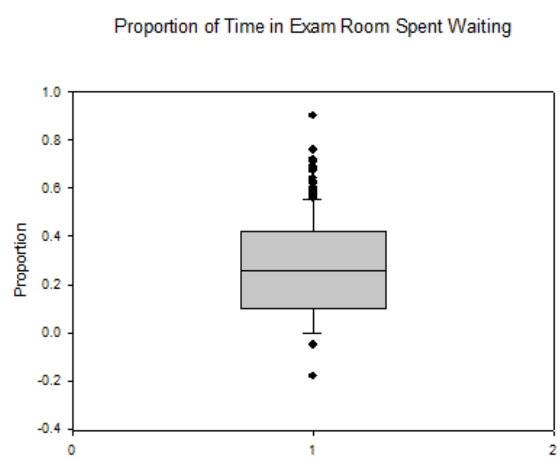


Figure 14: Box plot of proportion of time in exam room spent waiting

5.4 There is often a Gap between the Scheduled and Actual Appointment Durations

While one focus of our data analysis was on the number of appointments that began late, evaluating the length of appointments was a vital piece of the study, because it quantified the issue of delays due to physicians being held up in unexpectedly long appointments. We determined that there were several possible appointment lengths including 15, 20, 30, 40 and 45 minutes. In our data, we had a limited number of 15 and 45 minute appointments, so we focused on the appointment lengths that had large enough sample sizes to provide us with beneficial information (20, 30, and 40 minute appointments in that order). We calculated the difference between the physician's arrival and departure from the exam room, and showed that in a histogram for each appointment length (20, 30, 40 minutes). Extreme outliers may be attributed to human error in data collection or an instance of unexpected tests.

For the 20-minute appointments during our data collection period, the vast majority lasted between 16 and 40 minutes, as seen in Figure 15. The length, as well as the variation in lengths, is a concern because if a physician has a schedule full of back to back 20-minute appointments scheduled in a day, it would be extremely easy to get behind schedule if a large portion of them are lasting longer than scheduled.

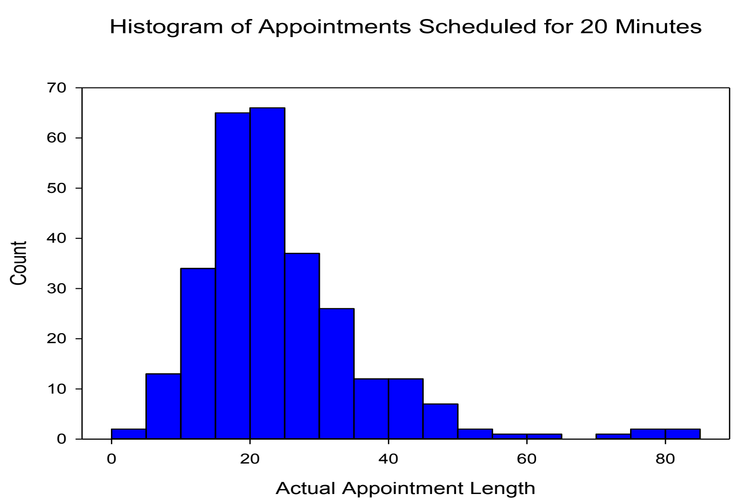


Figure 15: Histogram of appointments scheduled for 20 minutes

Of the 30-minute appointments, most fell between 21-25 minutes or 31-35 minutes. Figure 16 illustrates this. These were of less concern than the 20-minute appointments because, depending on how the appointments lengths fall between physicians and days, some of the shorter appointments could make up for the longer ones.

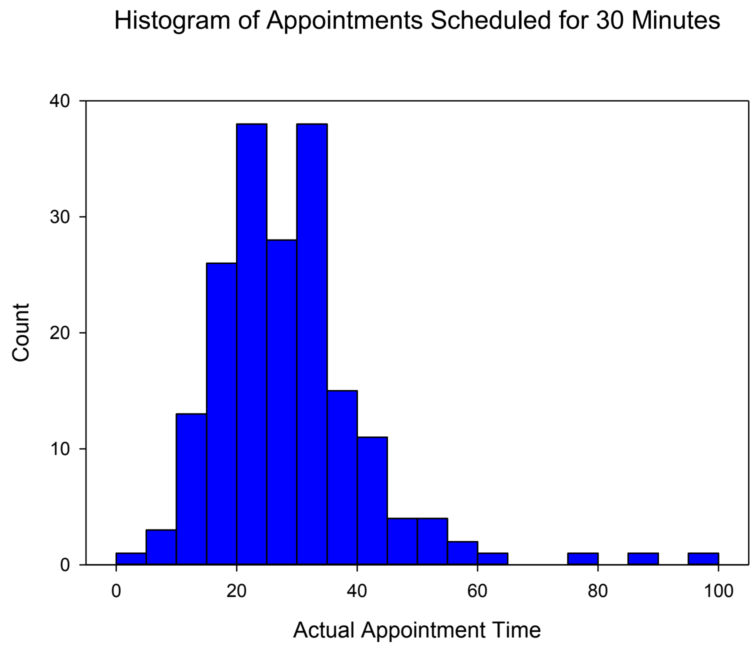


Figure 16: Histogram of appointments scheduled for 30 minutes

The majority of the 40-minute appointments, on the other hand, lasted less than 40 minutes, as shown in Figure 17.

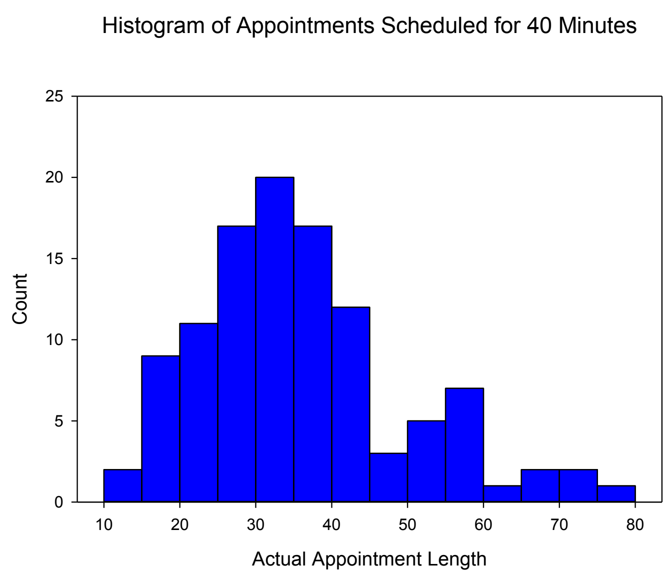


Figure 17: Histogram of appointments scheduled for 40 minutes

This analysis emphasized the importance of evaluating the reasons that such a high percentage of appointments are lasting longer than scheduled. We enter this study shortly after the physicians in the clinic changed appointment lengths, concerned that the previous appointment lengths were causing the issues in the clinic. We understand that simply extending appointment lengths can only add a limited benefit, especially noting how full physician's schedules generally are. Fitting in a 40-minute appointment to a physician's schedule can be difficult unless it is scheduled far in advance.

5.5 Cancellations and No-shows Can Help the Clinic Absorb Lateness

The data analysis for cancellations and no-shows was performed using the original scheduling system data that spanned from October 1, 2014 to October 8, 2015. After removing weekends, flu visits, and duplicates (in the case where an appointment was booked in a time slot, then cancelled, and the slot was booked again by another patient), we discovered an average of 28 no shows and 58 cancellations a day. The control chart in Figure 18 depicts the average number of cancellations and no shows (added together) per day over the course of October 1, 2014 to October 8, 2015.

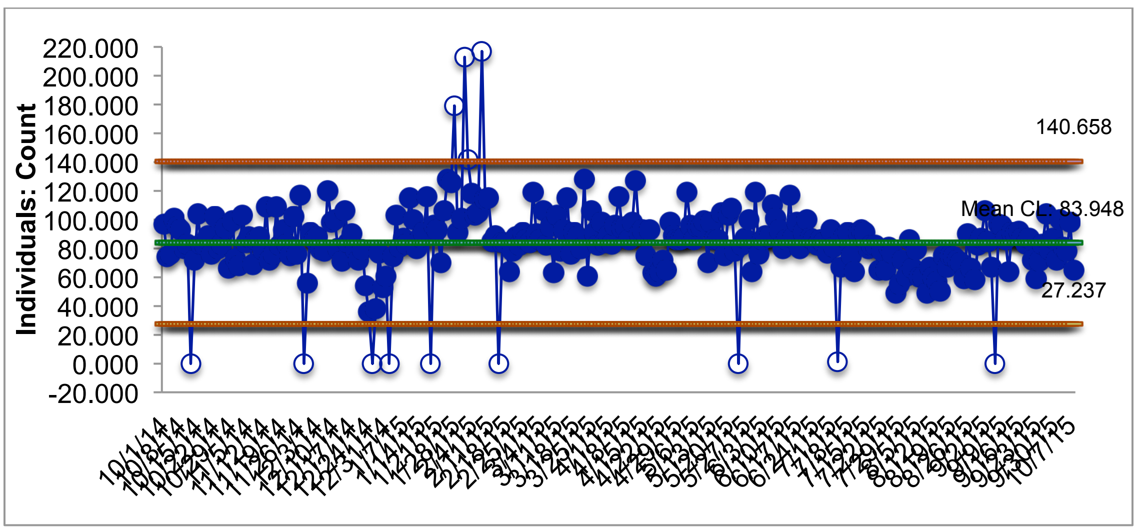


Figure 18: Control chart of cancellations and no shows per day

About 20% of those cancellations, or approximately 11 per day, happen within 8 hours, and according to clinic staff, it becomes very hard to fill an appointment slot within 8 hours of its start time, so these slots tend to remain open. Therefore, the 28 no shows and the 11 last minute cancellations increase the clinic's capacity to absorb lateness in the clinic throughout the day.

There are generally about 30 physicians and residents working in the clinic per day. As a result, those no shows and cancellations may equate to roughly one per physician per day. Figure 19 shows the cumulative percentages of cancellations, organized by time in advance of the appointment.

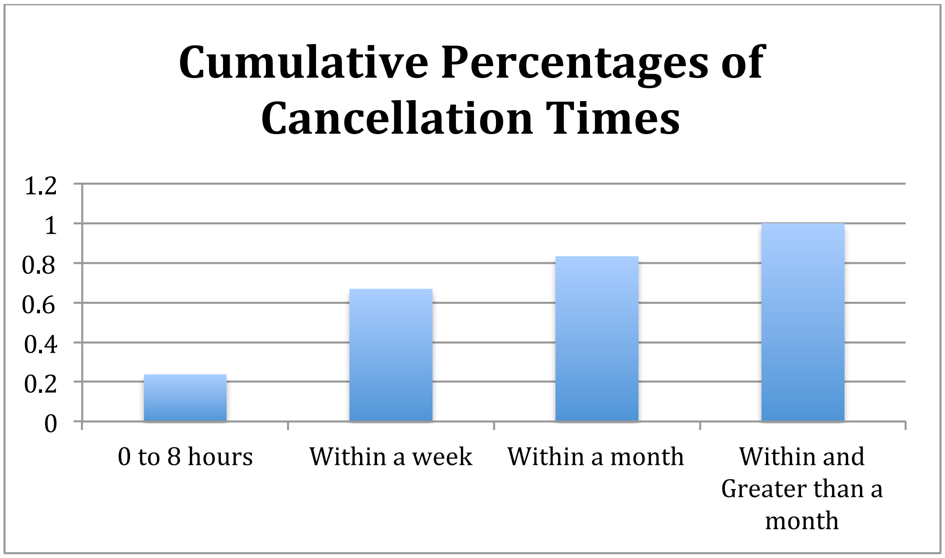


Figure 19: Cumulative percentages of cancellation times

The Pareto Chart in Figure 20 categorizes all cancellations within the period (10/1/2014 through 10/08/2015) and organizes them by frequency. The blue line depicts the cumulative total percentage of cancellations. For instance, 45% of cancellations can be attributed to unspecified personal reasons, while 73% are attributed to a combination of transportation/schedule issues and personal reasons. The thirteen remaining reasons account for only 27% of cancellations.

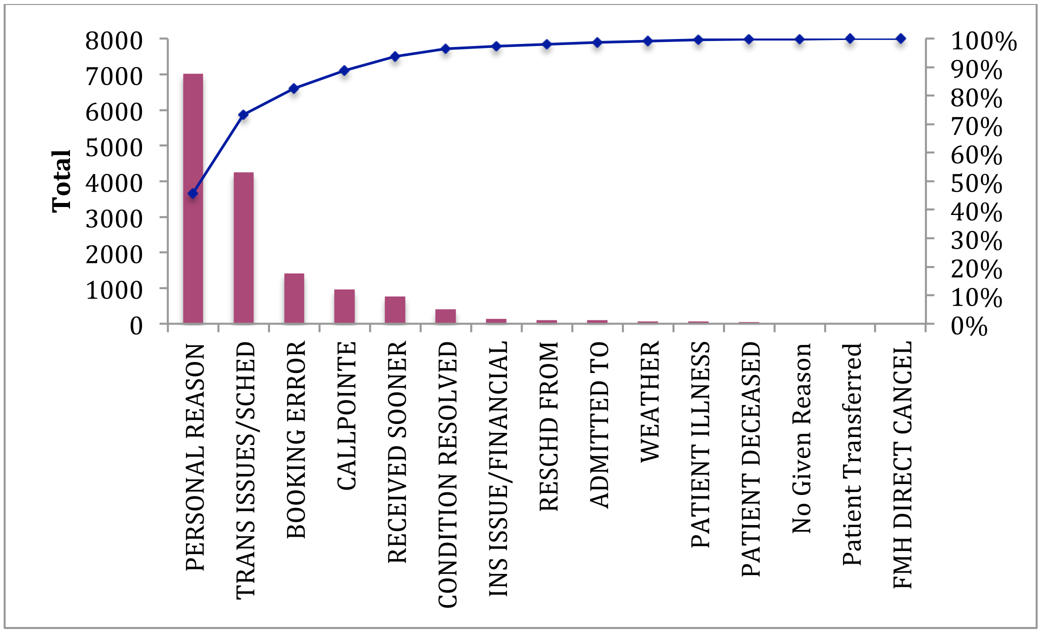


Figure 20: Pareto chart of cancellation reasons

Some of the above codes, however, are relatively new, having been added partway through the time period covered by the data (February 6th, 2015), so the percentages may be slightly inaccurate. For example, "weather" may have fallen under the transportation issues or personal reason (for example, if schools were closed and people were required to stay home with their children) categories prior to February 6th. That said, each new code accounted for less than 1% of cancellations between February 6th and October 8th 2015. The frequency of cancellations citing personal reasons did decrease very slightly in percentage during this time period as compared to the first part of the year (October 1, 2014 to February 5, 2015), accounting for 0.6% less of cancellations.

6.0 Results and Findings from Simulation Model

This chapter describes the simulation models that were created to explore opportunities for process improvement in the clinic. The first model is the base case, which represents the clinic's current state. All of the alternate models were created by making adjustments to this base case model.

6.1 The Base Case Model

6.1.1 Introduction and Model Basics

The model simulates the appointment process at the clinic, starting after the patient has completed the registration process and ending after the exam ends, specifically when the face-to-face interaction with the provider is over.

The model starts with a master schedule for the day ahead. The schedule includes a list of appointments that are scheduled as well as the staff members who are scheduled to be in the clinic. Before a day in the clinic starts, there is no way to tell if patients will arrive early or late, or how long each appointment will take. The model simulates different ways the day could go, based on past data from the clinic. The patient arrivals as well as patient service times are determined using statistical distributions.

6.1.2 Excel File for Arrivals

The Excel spreadsheet attached to the model contains a possible scenario based on existing data about the clinic. It contains a combination of actual data (the appointment schedule taken directly from the clinic) and random data (data generated based on distributions that is comparable to actual data).

The file has seven columns, four of which (B, D, F, G) are used in the model. These four columns are also Named Ranges in Excel – Arena requires this in order to use the data. The first few rows of the spreadsheet are shown in Figure 21 below.

	A	B	C	D	E	F	G
1	7:01 AM	1.03	7:00 AM	1.0000	2	10	30
2	7:09 AM	1.16	7:30 AM	1.5000	-20	9	30
3	7:15 AM	1.25	7:30 AM	1.5000	-15	15	40
4	7:21 AM	1.35	8:00 AM	2.0000	-39	12	20
5	7:22 AM	1.37	7:50 AM	1.8333	-28	11	40

Figure 21: Screenshot of Excel data file

Column A contains the patient's arrival time to the model – the time they completed the registration process and left the check-in desk. **Column B** (named "ArriveTime") contains the same information in a form that is usable by Arena. Here, the arrival time is expressed in hours after the model's start time.

Column C contains the patient's scheduled appointment start time. **Column D** (named "ApptTime") contains the same information in a form that is usable by Arena. The appointment time in Column D is expressed in hours after the model's start time.

Column E contains a randomly generated number that represents how early or late the patient was for the appointment. This lateness number was calculated using the formula in Figure 22, and was used to calculate the patient's arrival time relative to their appointment time (Columns A and B). In this data, a negative number indicates earliness and a positive number indicates lateness. The mean and standard deviation for this formula were calculated with Arena's input analyzer. All of the actual arrival times during the data collection period in November (minus a few extreme outliers on either side) were used, and this was the distribution that best fit the data. Note that, on average, patients arrive a few minutes early to their appointments. The distribution is centered at -9.91, or about 10 minutes before the appointment time, but there is considerable variability in the distribution of lateness – the standard deviation of the distribution is 17.3 minutes.

random lateness	formula		=NORM.INV(RAND(),K2,K3)		
	mean	-9.91			
	stddev	17.3			

Figure 22: Calculating lateness for the model

Column F (named "ProviderCode") contains an integer that refers to which provider the patient is scheduled to see. Attending physicians were assigned an integer number (starting at 1), and all out-of-scope providers (urgent care, residents, and the clinic's psychologist) were assigned 0.

Column G (named "ApptDuration") contains the scheduled length of the patient's appointment in minutes.

As an example, from the first row of the spreadsheet in Figure 21, it can be read that a patient arrived at 7:01 AM (~1.03 hours after the model started at 6:00 AM) for an appointment at 7:00 AM (1 hour after the model started). The patient's appointment is with provider 10, and it is scheduled to last 30 minutes.

6.1.3 Model Structure

Entity Creation and Arrival

When the simulation starts (Model time = 0.000, equivalent to 6:00 AM), a single entity is created. This is the only entity that will be created for the duration of the simulation. The entity then passes through a series of ReadWrite modules that read, from the Excel spreadsheet described in Figure 21, the first patient's arrival time, scheduled appointment time and duration, and the provider the patient is scheduled to see. The information is attached to the entity as a series of attributes. This process is displayed in Figure 23.

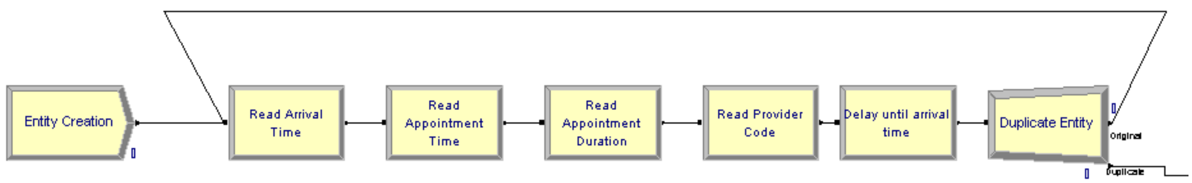


Figure 23: The beginning of the model

The entity then enters a Delay module, the details of which are shown in Figure 24. The entity is held in this module until the arrival time that was read from the file. The length of the delay is calculated by subtracting the current model time (TNOW) from the arrival time.

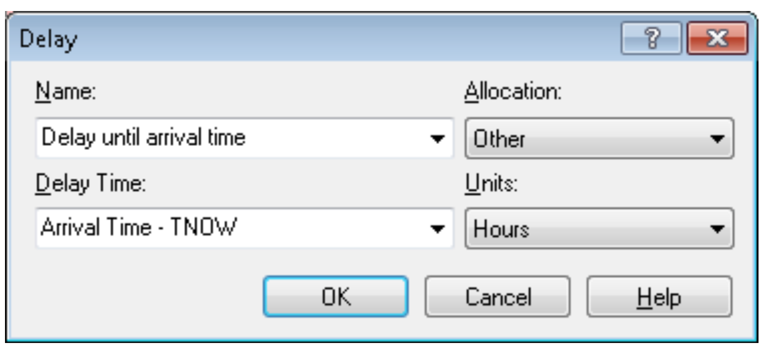


Figure 24: Details about the delay module

After the entity is released from the Delay module, it enters a Separate module. This module creates a copy of the entity, sends the duplicate into the next step of the model, and sends the original entity back to the beginning of the loop. The original entity travels through the ReadWrite modules again, and the attributes assigned for the previous patient are overwritten with new ones from the next row of the spreadsheet. This process is repeated until all rows of the file have been read, at which point the original entity stops moving and the rest of the model continues to run.

The Vital Signs and Appointment Processes

After the arrivals loop shown in Figure 23, the duplicate entity enters the next part of the model (pictured in Figures 25 and 26). First, a counter called "PatientsInClinic" is incremented. This counter is used because the way Arena defines WIP (work in process) is different from the way it is defined in the clinic. The patient then travels to the waiting room (using a Route module to travel to the Waiting Room Station).

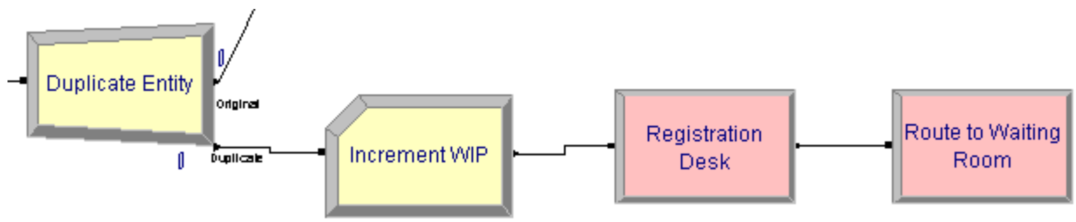


Figure 25: The patient enters the rest of the model

Figure 26: The vital signs process

When a PCA is available, the patient seizes a PCA resource and travels to the vital signs area (this time, using a Leave module to travel to the Vital Signs Station). The processing times at the vital signs station are assigned based on the Weibull distribution

$$0.5 + \text{WEIB}(2.83, 1.88)$$

This distribution was determined using Arena's Input Analyzer tool and is based on a best-fit analysis of the vital signs length collected by patients during our data collection period. An image of the distribution plotted over a histogram of the data is shown in Figure 27.

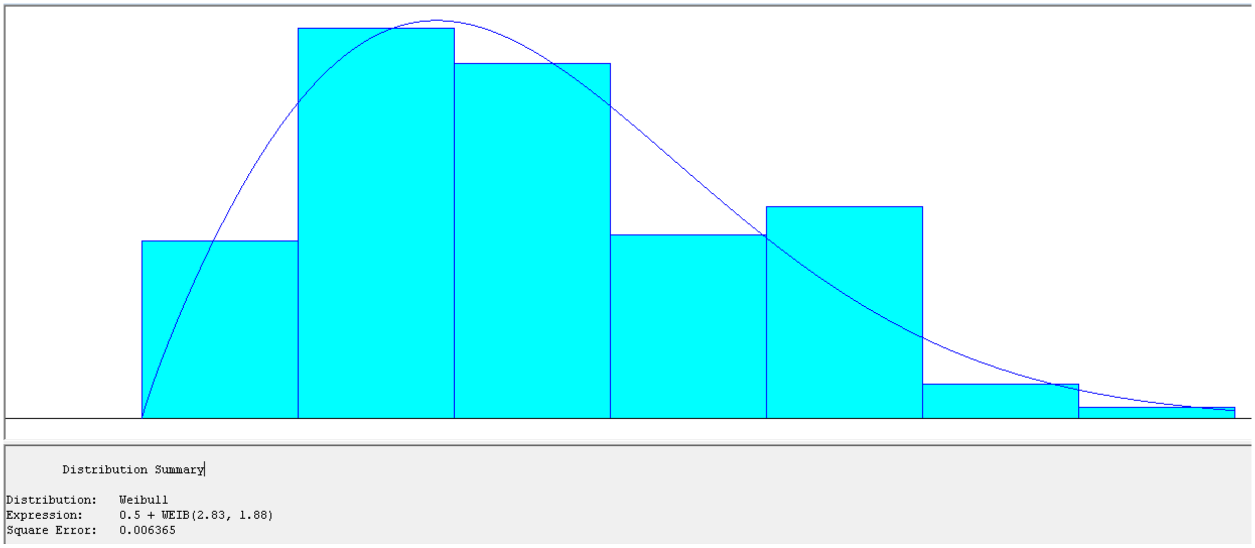


Figure 27: Screenshot of Arena's Input Analyzer

After the vital signs process is complete, the entity moves to a Decide module, which sorts the patients based on the provider they are scheduled to see. Out of scope patients (those with a provider code of 0) are sent on to a branch of the model with slightly different parameters than the rest, while in-scope patients continue to a series of Decide modules that sort them into different queues based on the provider they are scheduled to see. Part of this sorting mechanism is pictured in Figure 28.

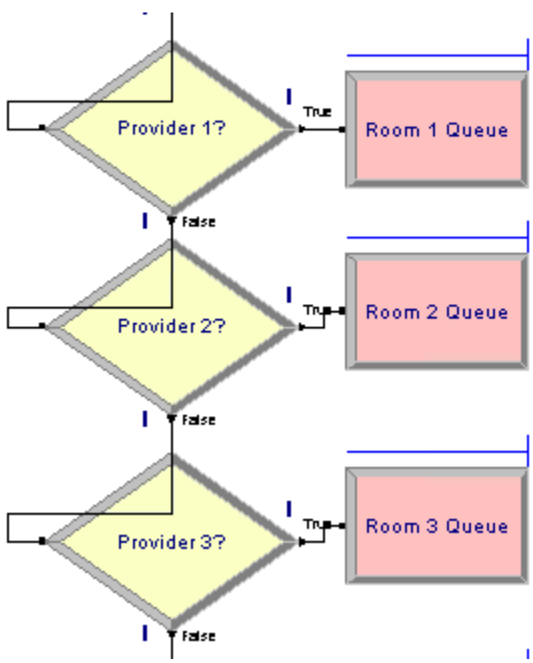


Figure 28: Part of the model's patient sorting mechanism

Each of the Decide modules has two exit points: one on the rightmost corner (for patients who satisfy the module's condition), and one on the bottom corner (for patients who do not satisfy the condition). Patients who satisfy the condition are passed to a queue to wait for an available exam room (equivalent to being sent back to the waiting room). Each provider has two rooms, both of which are covered under the "Room X Queue" module. When a room belonging to their assigned provider becomes available, the patient seizes the room and continues on to the part of the process being scrutinized most closely, the time spent with his or her provider. This portion of the model is pictured in Figure 29.

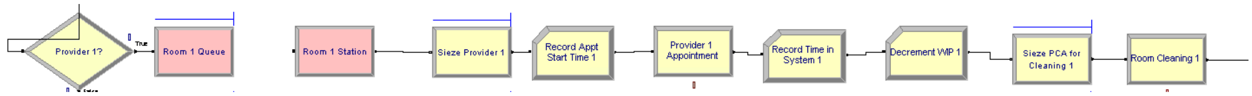


Figure 29: The appointment process

After the patient seizes an exam room, it enters a queue to wait for the provider. If their provider is available immediately, the appointment can start right away, but if not, the patient must wait (equivalent to waiting in the exam room for the provider to come in).

Once the patient seizes his or her provider, the appointment starts. The start time is recorded (expressed as the difference between the scheduled start time and the actual start time – a negative value indicates that the appointment started ahead of schedule, and a positive value indicates that it is behind schedule). The duration of the appointment is determined using the following formula:

ABS(NORM((1.1*ApptDuration),(ApptDuration/3)))

The distribution is normal, with a mean (center) of 110% of the scheduled appointment duration, and a standard deviation of 1/3 of the scheduled length. This standard deviation was chosen based on the actual measured appointment lengths for each scheduled length, and also because 99.7% of points below the normal curve fall within 3 standard deviations of the mean. It is important to note that, for ease of modeling, this distribution was developed with aggregate data for appointments of all lengths, as opposed to creating separate distributions for each scheduled appointment length. This generalization may result in some unrealistic service times being generates by the model.

Because the normal curve is unbounded, it is possible for a negative value to be generated, but due to the parameters, it is extremely unlikely. In the rare case that a negative process time is generated, Arena will use the absolute value of the time (for example, if a -4 is produced by the normal distribution, Arena will use an appointment time of 4 minutes).

Out-of scope appointments are simulated the same way, but with a slightly different appointment length assignment. This part of the model was adjusted toward the end of the project, and creating an entirely new model and set of data files was not feasible within the time constraints. A compromise was reached – to determine the processing time for all the out-of-scope appointments, the average of the scheduled appointment lengths was determined and that time was assigned to each appointment with provider 0.

After the appointment has been simulated, the provider resource is released (so the provider can care for another patient). The patient enters a Record module, which records the total time spent in the system – from their arrival time until the end of their time with their provider. Then, the counter "PatientsInClinic" is decremented, because after the appointment ends, the patient is free to leave the clinic.

After each patient, the exam room must be cleaned. The entity remains in the system (even though the patient has left the scope of the model), seizes the next available PCA, and there is a delay, which follows a triangular distribution with minimum 3 minutes, mode 7 minutes, and maximum 10 minutes, after which the room and the PCA are released and the entity leaves the system. Though in most cases, cleaning a room takes no more than 5 minutes, the extra time built into the distribution accounts for the possibility that the patient needs additional tests or the PCA has other tasks to complete before moving on to the next patient.

Utilizing Data from the Clinic

Two identical versions of the model were created with different inputs to represent different days. Both models have the same logic and process patients in the same way, but they use different inputs (for example, the staffing schedules and the list of scheduled appointments). The first version represents the highest-volume day during our data collection period (Monday, November 23rd, 2015), and the second represents the lowest-volume day during the same period (Thursday, November 19th, 2015).

The models have the same structure and function in the same way, but their resource availability is different. For example, on the high-volume day, there were 23 attending physicians in the clinic (not including residents, urgent care, or psychology), but on the lower-volume day, there were only 15. The two days also had slightly different PCA staffing schedules.

Both versions of the model have a PCA staffing schedule that mirrors the actual staffing schedule used in the clinic on that day. The schedule from November 23rd is below.

- 6:45 AM – 3:15 PM: 1 PCA
- 6:45 AM – 3:45 PM: 1 PCA
- 8:00 AM – 4:30 PM: 1 PCA
- 8:30 AM – 5:00 PM: 1 PCA
- 9:00 AM – 5:30 PM: 1 PCA
- 9:30 AM – 6:00 PM: 1 PCA
- 10:00 AM – 6:30 PM: 1 PCA

Somewhere in the middle of their scheduled shift, each PCA is assigned to one of four 45-minute lunch slots. On November 23rd, the lunch schedule was as follows:

- Lunch 1 (11:15 AM – 12:00 PM): 1 PCA
- Lunch 2 (12:00 PM – 12:45 PM): 3 PCAs
- Lunch 3 (12:45 PM – 1:30 PM): 2 PCAs
- Lunch 4 (1:30 PM – 2:15 PM): 1 PCA

Similarly, the shift and lunch schedules from the 19th are below:

- 6:45 AM – 3:15 PM: 1 PCA
- 6:45 AM – 3:45 PM: 1 PCA
- 8:00 AM – 4:30 PM: 1 PCA
- 9:00 AM – 5:30 PM: 2 PCAs
- 9:30 AM – 6:00 PM: 1 PCA
- 10:00 AM – 6:30 PM: 1 PCA

- Lunch 1 (11:15 AM – 12:00 PM): 2 PCAs
- Lunch 2 (12:00 PM – 12:45 PM): 2 PCAs
- Lunch 3 (12:45 PM – 1:30 PM): 2 PCAs
- Lunch 4 (1:30 PM – 2:15 PM): 1 PCA

These shift hours and scheduled lunch breaks allow all of the PCAs to be in the clinic during the busier periods, and coordinate the breaks with the hours during which fewer appointments are scheduled.

Above, it was mentioned that out-of-scope providers' appointments are handled slightly differently from the rest of the appointments. On November 23rd, the processing time for these appointments was calculated to be 23 minutes, and on November 19th, it was 30 minutes.

6.1.4 Model Assumptions

When simulating, it is impossible to model a system exactly as it functions in real life. While creating this model, several assumptions and generalizations were made, due to limitations of the modeling software or as a way to simplify the modeling process. As a result, the model is not an exact representation of the clinic, but rather an approximation.

The first major generalization has to do with the availability of providers in the model. Even though some providers only have half-day clinic sessions (morning or afternoon), their schedules were created so that they would be available for the entire duration of the model. This reflects the fact that there is no universal start or end time for a morning or afternoon session. A morning session may start at 7:00 AM, or it may start at 9:00 AM. Rather than assign each provider a window of availability, the appointment times scheduled for each provider were assumed to do an appropriate job of dictating the provider's schedule.

Second, the distribution of exam rooms was simplified. Though it is usually the case that a provider is assigned to two exam rooms, during some sessions, some providers may only have one room, or two providers might share a room. For simplicity's sake, each provider was assigned two rooms for the duration of the day. This could lead to a slightly higher number of rooms in use in the model than in real life, but since it is such an infrequent occurrence, it was decided that it would not have a significant effect on the results from the model.

Third, the PCAs in the model only have two tasks – measuring patients' vital signs, and cleaning rooms once appointments are finished. In reality, PCAs have significantly more responsibilities than just those two. They are responsible for escorting patients between stations, overseeing the overall status and function of the clinic, and in some cases serving as a liaison between patients and physicians. As a result of this assumption, the PCA utilization may be lower than it is in reality.

The model also does not take into account the clinic's geography or the fact that it functions as three smaller sub-clinics. In the model, if there is a patient that needs to be seen by a PCA and an available PCA, the PCA will help the patient. But, in reality, the patient needing assistance and the available PCA might be located in different areas, so they would not interact. Because of the PCA tasks and inaccurate representation of the clinic's geography, the PCA availability is overestimated in comparison to the clinic.

6.1.5 Base Case Verification and Validation

In order to be confident in a model's performance and the recommendations made as a result of a simulation study, it is important to verify that the model is structured correctly and validate its performance (whether or not it is representative of the system being modeled).

Verification

After constructing the model based on my own knowledge of the clinic, I consulted with the sponsoring physicians to verify that the process steps in the model mirrored the process steps in the clinic. I described the model step-by-step, including which data was used at each step, and made any necessary adjustments based on their feedback. When all three of us were satisfied with the model, I considered it to be verified.

Validation

For each of the two versions of the model (one representing each of the days), 10 different arrival scenarios were created, using the distribution for lateness developed from the actual arrivals during our data collection period. Each scenario is contained in a separate Excel spreadsheet, and they all follow the same format described in section 6.1.2. Each model was run for each of its 10 scenarios, for 5 replications each.

The two main metrics used for validation were the total amount of time a patient spent in the clinic (from the time they registered until the time their appointment ended) and the average lateness of appointment starts (the difference between the actual and scheduled start times). I compared the metrics using 95% confidence intervals (calculated using Excel for the past data, and generated by Arena for the simulated data). A confidence interval is based on the calculated mean of a set of data, and describes, with a certain level of confidence, a range in which the "true" value of the statistic falls. For example, if a 95% confidence interval came out to be 2 ± 0.25, it means that there is a 95% chance that the true value of whatever is being measured falls between 1.75 and 2.25. A smaller confidence interval (which could be caused by using a large sample size or a relatively low confidence level) indicates a higher degree of certainty.

Since each day in the clinic can be very different, and no two days are exactly alike, two identical schedules could produce very different results. I concluded that the model is representative of the clinic because the confidence intervals all overlap. In some real-life cases, there may be a statistically significant difference between the simulation's mean and the measured mean from the clinic that day, but since patient arrival times can have an enormous effect on the efficiency of the clinic, two days with identical schedules could perform in statistically different ways. The means and confidence intervals for all scenarios on both model days are displayed below in Figures 30 and 31. Another means of confirming that the model produces representative results is using the average of the scenario averages, pictured at the bottom of each image. The values generated by the simulation are close to (within the confidence interval of) the actual values from that day. For example, the left half of Figure 30 contains data about the average total time in system from the base case model representing November 23rd. The average of this statistic during our entire data collection period was 0.8375 hours (just over 50 minutes), and the actual measured average on November 23rd was 0.8274 hours (just under 50 minutes). The value of this statistic from each of the base case scenarios is listed, and each average value falls within the confidence interval based on the measured data from November 23rd (0.8274 ± 0.04 hours), confirming that the model performs similarly to the way the clinic does when faced with a similar scenario. The average of the 10 scenarios is 0.82235 hours (about 49 minutes), which is also well within this confidence interval.

	all scenarios were run with 5 replications					
Total Time in System (REG - PROV OUT)				Actual - Scheduled Appointment Start		
Source	average (hours)	95% half-width (hours)		Source	average (hours)	95% half-width (hours)
actual data (overall)	0.8375	0.0301		actual data (overall)	0.21	0.0214
actual data (this day)	0.8274	0.04		actual data (this day)	0.1686	0.06
scenario 1	0.7919	0.03		scenario 1	0.1355	0.02
scenario 2	0.8073	0.05		scenario 2	0.1883	0.04
scenario 3	0.8241	0.07		scenario 3	0.19	0.06
scenario 4	0.8263	0.03		scenario 4	0.1695	0.02
scenario 5	0.8169	0.02		scenario 5	0.1806	0.02
scenario 6	0.8577	0.02		scenario 6	0.1667	0.02
scenario 7	0.8636	0.04		scenario 7	0.2156	0.02
scenario 8	0.7933	0.03		scenario 8	0.1424	0.02
scenario 9	0.8136	0.02		scenario 9	0.1651	0.01
scenario 10	0.8288	0.05		scenario 10	0.1802	0.04
	average of scenarios				average of scenarios	
	0.82235				0.17339	

Figure 30: Confidence intervals for November 23rd

all scenarios were run with 5 replications

Total Time in System (REG - PROV OUT)			Actual - Scheduled Appointment Start		
Source	average (hours)	95% half-width (hours)	Source	average (hours)	95% half-width (hours)
actual data (overall)	0.8375	0.0301	actual data (overall)	0.21	0.0214
actual data (this day)	0.8659	0.07	actual data (this day)	0.2351	0.06
scenario 1	0.8772	0.06	scenario 1	0.2268	0.04
scenario 2	0.8589	0.05	scenario 2	0.1727	0.04
scenario 3	0.8331	0.04	scenario 3	0.1784	0.03
scenario 4	0.8788	0.04	scenario 4	0.2209	0.02
scenario 5	0.8251	0.02	scenario 5	0.1807	0.02
scenario 6	0.8255	0.04	scenario 6	0.1716	0.03
scenario 7	0.8704	0.05	scenario 7	0.2068	0.03
scenario 8	0.8519	0.04	scenario 8	0.1797	0.02
scenario 9	0.8767	0.04	scenario 9	0.1886	0.03
scenario 10	0.8789	0.06	scenario 10	0.2513	0.04
	average of scenarios			average of scenarios	
	0.85765			0.19775	

Figure 31: Confidence intervals for November 19th

The value that is furthest from the actual measured value, though it is still within the confidence interval, is the degree of appointment lateness from the lower-volume day (the 19th). This discrepancy could result from the fact that the model does not simulate the doctors' travel time or other activities between appointments. Some physicians choose to dictate details of each appointment between appointments, while others wait until the end of their clinic session. Overall, though, the testing demonstrated that the model is representative of the clinic's operations, so that any scenarios tested will produce valid results that can be used to make recommendations to the clinic staff.

6.2 Alternative Models Tested

After verifying and validating the base case model, I spoke with the sponsoring physicians and we identified 3 areas of focus – Patient arrivals, PCA staffing, and appointment lengths. Changes in each of those three areas were tested against the base case model. I also experimented with a lateness threshold – turning away patients who arrive to the clinic more than a specified number of minutes after their appointment time.

6.2.1 Reducing Variation in Patient Arrivals

To test the effects of patients' arrivals on the clinic's operations, four new sets of Excel data files were created, each with a distribution that differed from the original. The parameters of the alternate distributions are displayed in Table 3.

Table 3: Alternate distributions for patient lateness

Description	Mean (minutes)	Standard Deviation (minutes)
Base case	-9.91	17.3
Reduced variation	-9.91	8.65
Shift arrivals 5 minutes earlier	-15	17.3
Shift arrivals 5 minutes later and reduce variation	-5	8.65
Every patient is 10 minutes early	-10	0 (no variation)

Each set of files was used to drive the base case models to better understand how patient lateness affects the process. The results of each alternative are shown in Table 4. Instances in which the result from the alternate model was statistically significant from the base case model are shaded.

Table 4: Effects of patient arrivals

	Low-Volume Day (November 19th)		High-Volume Day (November 23rd)	
	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)
Base Case Model	11.865	51.459	10.403	49.341
Reduced variation	11.2632	50.0502	7.7598	46.5678
Shift arrivals 5 minutes earlier	8.0484	51.5754	5.5314	48.948
Shift arrivals 5 minutes later and reduce variation	15.324	49.1976	12.5598	46.5162
Every patient is 10 minutes early	11.1	50.298	8.376	47.466

Simply reducing the variation in lateness (changing the standard deviation from 17.3 to 8.65 minutes) had a significant effect on the average total time in the clinic on both model days. This is likely due to the fact that fewer patients arrived extremely early or extremely late, which can cause delays in the clinic. For example, if a patient arrives 30 minutes late for a 20 minute appointment, their physician is likely already seeing another patient, and will have to fit the late patient into their schedule among the rest of their appointments that day, which will cause significant delays. If patients arrive closer to their appointment times, they are more likely to be seen in order and less likely to cause delays for themselves or for other patients. A similar effect was seen in the distribution in which patients arrived, on average, slightly later, but with less variation. The conclusion I drew from this phenomenon is that patients who arrive either very early or very late are often seen out of order, which can lead to significant delays for their provider and for other patients in the clinic.

In both of the scenarios that involved shifting the center of the arrival distribution (one to -15 and one to -5), the average lateness of appointment starts was significantly impacted on both days. This is a logical effect; if all patients arrive a few minutes earlier (or later), but their processing times within the system do not change, it follows that appointments will start a few minutes earlier (or later) as a result.

In the situation in which every patient was registered 10 minutes before their arrival time, the effects were significant on the high-volume day, but not on the low-volume day. In general, the observed effects from all alternatives were of greater magnitude on the day with the higher volume, likely due to the higher patient volume and PCA utilization rates (since the maximum number of PCAs on the schedule was the same on both days). From this set of alternatives, it is clear that patient arrivals can have an impact on the way the clinic functions and on patient experience, especially on days with high patient volumes.

6.2.2 Adjustments to PCA staffing levels

To determine the effects that the number of PCAs in the clinic have on the way that the clinic functions, two adjustments to the PCA staffing levels the clinic currently uses were tested. First, an extra PCA was added to the schedule on both days. I assigned the new PCA to the 8:00 AM – 4:30 PM shift and to lunch slot 2, which starts at 12:00 PM and lasts until 12:45 PM. This shift time and lunch break were chosen because this would allow the extra PCA to be in the clinic during the busiest times, and out of the clinic when the patient volume is slightly lower due to fewer appointments being scheduled. Then, I tested the effects of having an unlimited PCA staff for the entire day. The results from these simulations, compared to the base case, are displayed in Table 5.

Table 5: Effects of PCA staffing levels

	Low-Volume Day (November 19th)		High-Volume Day (November 23rd)	
	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)
Base Case Model	11.865	51.459	10.403	49.341
One Extra PCA	11.6862	51.1908	9.891	48.8556
Unlimited PCAs	11.3652	50.9052	9.4974	48.3066

Neither of the altered PCA staffing levels resulted in any statistically significant changes to the metrics. This was an unexpected result. From our observations of the clinic and conversations with staff members, all of the PCAs seemed to be constantly busy – escorting patients between stations, measuring vital signs, cleaning exam rooms, performing tests, and a multitude of other tasks, but the model does not simulate all of these responsibilities. It only simulates the PCAs measuring vital signs and the room turnover process. In the high-volume model, the average PCA utilization was about 0.5808, or 58.08%, and in the low-volume model, it was about 0.4788, or 47.88%. Because the models do not simulate all of the tasks the PCAs are responsible for, the PCA utilization is likely much lower than it would be if it were to be measured by observing PCAs during a day in the clinic. Because the PCAs in the model are not constantly busy, and also because the model treats the clinic as one large clinic rather than three smaller sub-clinics, adding more PCAs did not have a significant effect. It is possible that adding a PCA to the clinic, where the PCA utilization rate is likely higher than in the model, would have a more pronounced effect than the model suggests.

6.2.3 Reducing Variation and Increasing Appropriateness of Appointment Lengths

As discussed in section 5.4, many of the clinic's appointments are not the same duration as scheduled, for various reasons. It could be that the patient booked the appointment on relatively short notice and was forced to take a slot shorter than what was needed, or it could be the case that the patient simply had a lot of complex medical issues to discuss with the physician or needed additional tests performed. In either case, the measured appointment length can differ significantly from the scheduled length. I chose to test what would happen if the clinic were able to increase the appropriateness of appointment lengths – in other words, ensuring that appointments are booked in long enough slots. First, I tested reducing the standard deviation of the model's appointment length distribution by half, so the new distribution was:

ABS(NORM((1.1*ApptDuration),(ApptDuration/**6**)))

Next I tested a more extreme scenario – eliminating all variation in appointment length. In this scenario, all appointments lasted exactly as long as they were scheduled to, simulating a perfect matching of patient needs to appointment time slots. Table 6 summarizes the results from these simulations. As above, statistically significant changes are shaded.

Table 6: Effects of variation in appointment lengths

	Low-Volume Day (November 19th)		High-Volume Day (November 23rd)	
	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)
Base Case Model	11.865	51.459	10.403	49.341
Reduced Variation in Appointment Length	10.6782	50.3478	9.3792	48.501
Appointments Length Exactly as Scheduled	7.047	44.7354	5.2206	42.117

Simply reducing the variation (by halving the standard deviation of the distribution) in appointment lengths appears to have an impact on the system, but at the 95% confidence level, it is not a statistically significant change. If the variation were to be eliminated completely, leading to all appointments lasting exactly as long as they were scheduled to, the change in the metrics becomes significant – on average, appointments start earlier and patients spend less time in the system, presumably because physicians do not get held up in appointments that last too long. This set of alternatives illustrated the magnitude of the cascade effect that can occur when multiple appointments in a row last longer than they are scheduled to.

6.2.4 Introducing a Lateness Threshold

Although the primary care clinic tries to accommodate every patient every day, I was interested in what would happen if they enforced a rule in which patients who arrive late for their appointments are asked to reschedule and are not processed on that day. Understanding that this may elicit mixed feelings from the clinic staff, two different lateness thresholds were tested – a relatively short 15 minutes, and a more lenient 30 minutes. Table 7 describes the effects of these two thresholds compared to the base case. Instances in which there was a statistically significant difference from the base case model are shaded.

Table 7: Effects of lateness thresholds

	Low-Volume Day (November 19th)		High-Volume Day (November 23rd)	
	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)	Average of Appointment Lateness (minutes)	Average of Total Time in System (minutes)
Base Case Model	11.865	51.459	10.403	49.341
15 Minute Threshold	6.650	48.377	4.276	45.908
30 Minute Threshold	11.320	51.607	9.378	48.838

In both the high-volume and low-volume models, the 15-minute lateness threshold had a significant impact on both appointment start times and patients' total time in the clinic. This is likely due to the fact that patients who were extremely late (the outliers) were eliminated. However, the way this was done was by simply removing them from the simulation and not processing them at all, therefore lowering the volume by 7-8%. Lower patient volume leads to fewer delays. The effects of this threshold were more pronounced on the high-volume day, presumably because a larger number of patients were eliminated from the simulation, essentially turning the high-volume day into a relatively lower-volume day. The 30 minute threshold did not have a significant effect on the system, due to the low number of patients arriving to the clinic more than 30 minutes late (about 1%).

6.3 Summary of Simulation Findings

To visualize the data presented in section 6.2, I created charts that compare the base case model to each of the alternatives. In each chart, the dots mark the averages for each scenario, and the black lines on either side of each one represent the 95% confidence interval for each one, calculated by Arena. The dot for the base case average is green, and the alternatives that had statistically significant impacts on the model when compared to the base case are red. Alternatives that had no significant impact are blue. Figure 32 shows the average lateness of appointment starts on the low-volume day (November 19th).

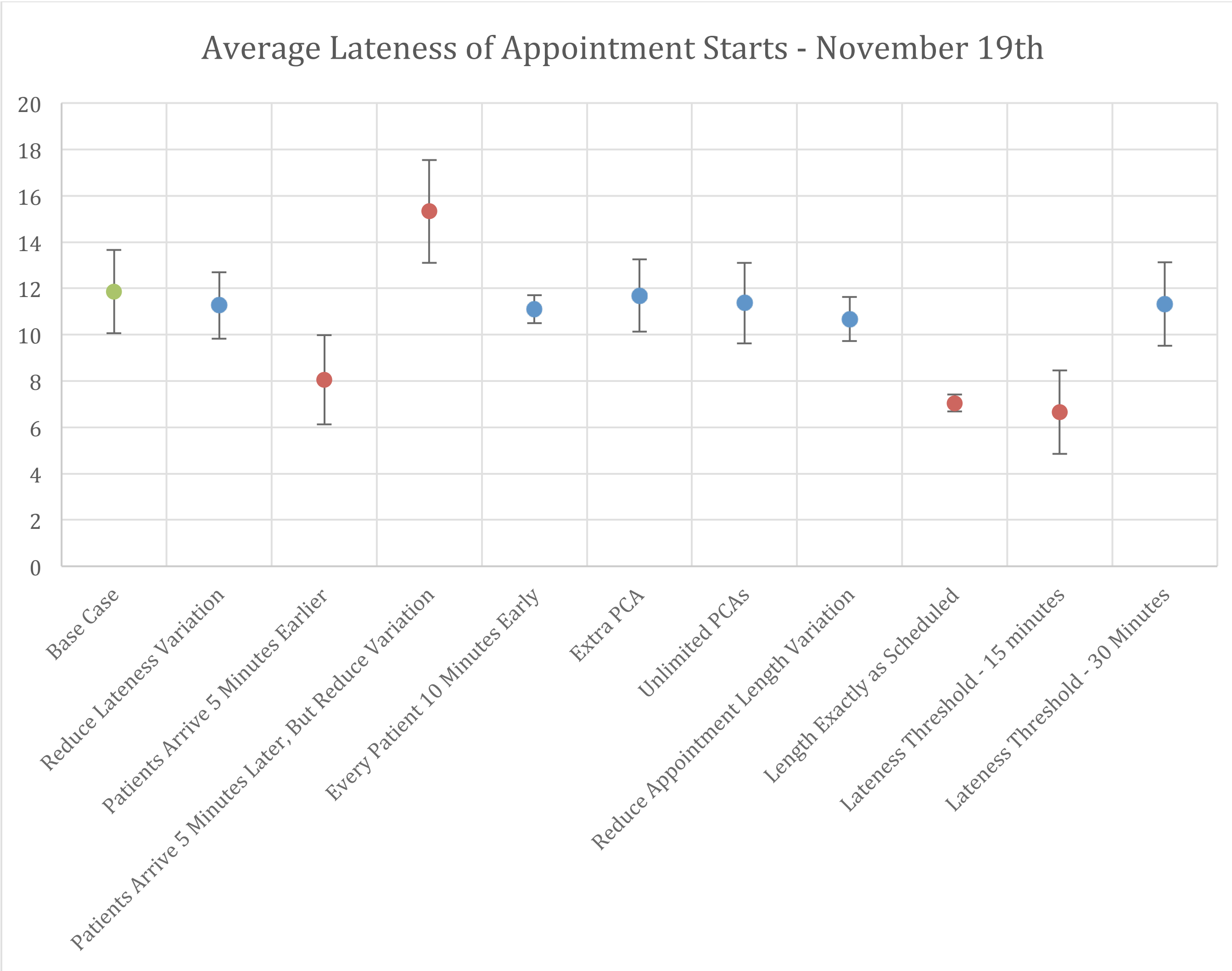


Figure 32: Changes to average lateness of appointment starts on November 19th

Figure 33 displays the averages and confidence intervals for patients' average total time in the clinic on November 19th. It follows the same format and color coding as Figure 32.

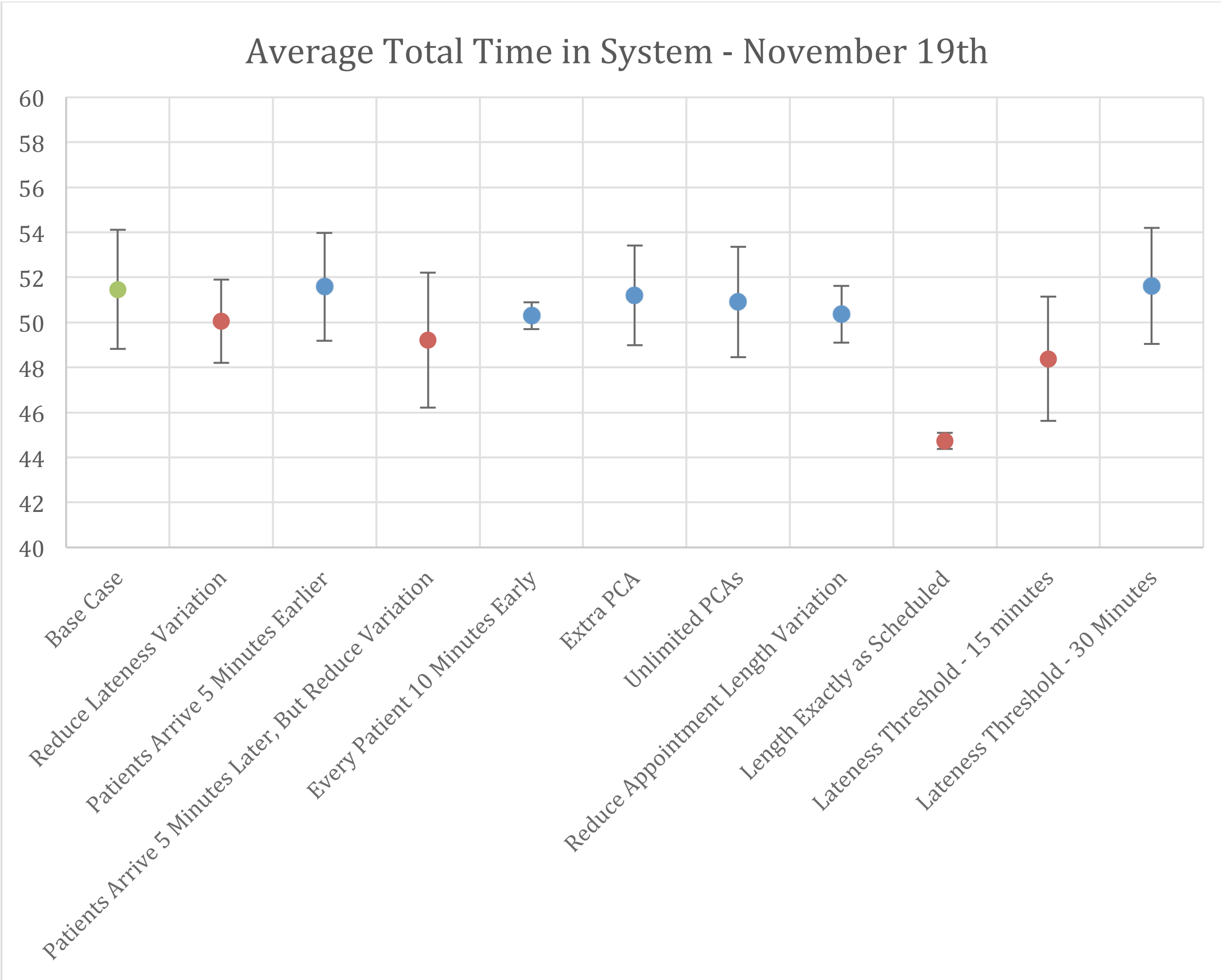


Figure 33: Changes to patients' total time in system on November 19th

Figure 34 shows the means and confidence intervals for the average lateness of appointment starts on the high-volume day, November 23rd, using the same format and colors as the previous charts.

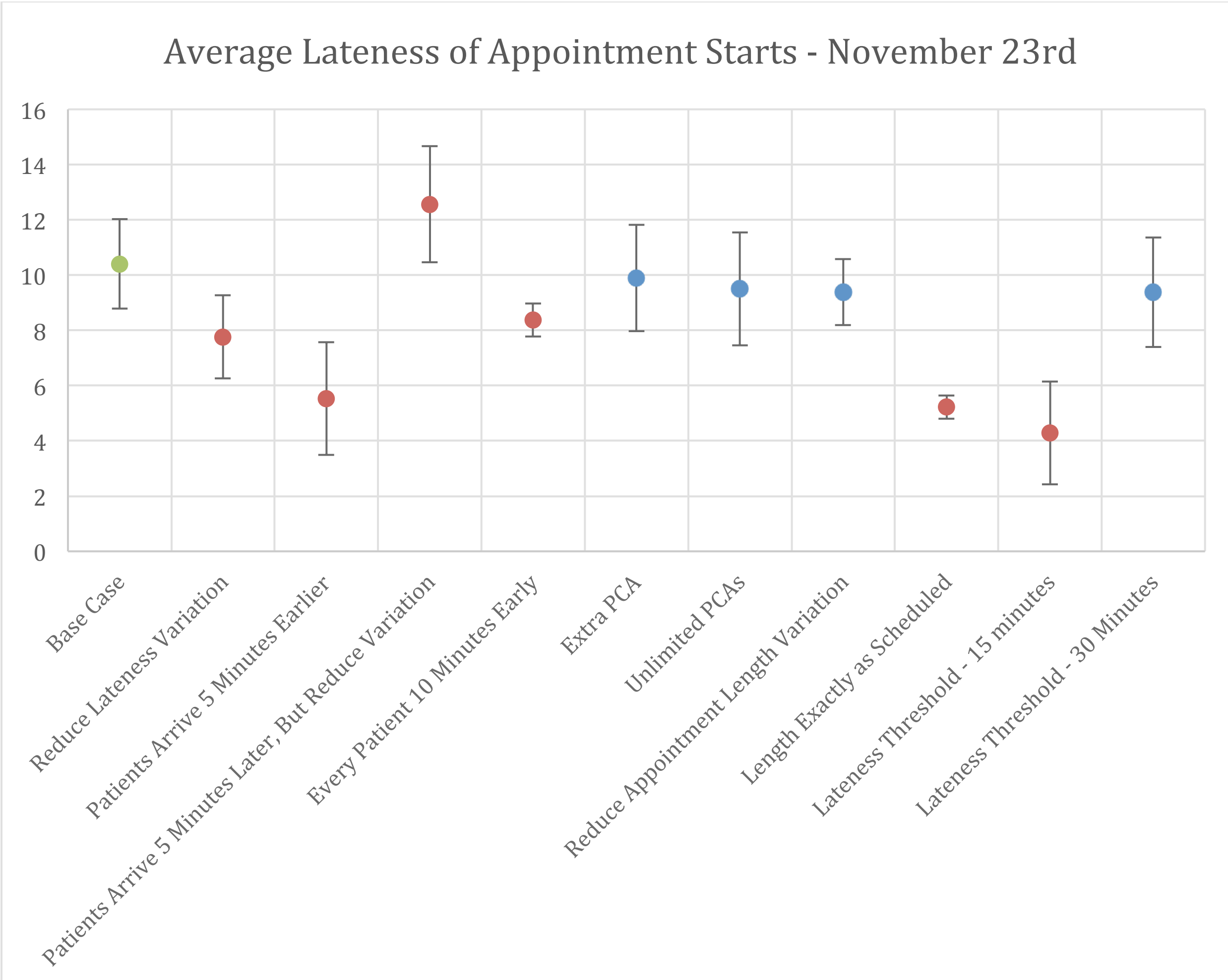


Figure 34: Changes to average lateness of appointment starts on November 23rd

Finally, Figure 35 shows the effects of each alternate scenario on patients' total time in the clinic on November 23rd (the high-volume day).

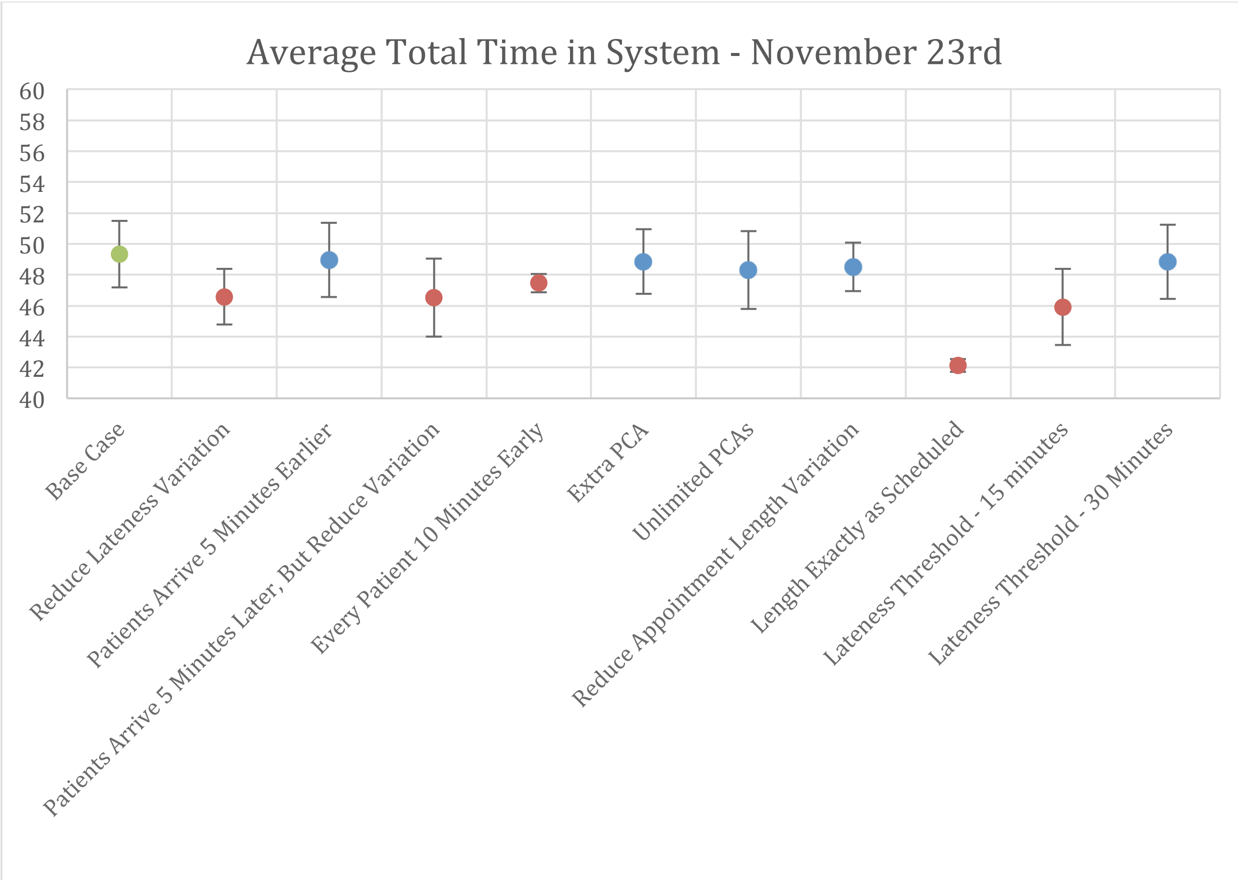


Figure 35: Changes in patients' average time in the clinic on November 23rd

Based on these results, overall conclusions are discussed in Chapter 7, along with some recommendations for the clinic about where to focus their efforts in order to have the most positive effects on patient experience.

7.0 Conclusions and Recommendations

The primary care clinic that was the focus of this project, like many others nationwide, is seeking to reduce delays in its appointment process. After data collection and analysis, a simulation model was created to represent the clinic and changes were tested in four areas to determine which could have the greatest effect on the clinic's process and the overall patient experience. The current project was performed with a narrow set of parameters that might not mirror a particular practice and real world circumstances. Although the model is an approximation of the clinic, it provided some valuable insights to the clinic staff about where their process improvement efforts might be most effective in their constant quest to improve patient experience. Some promising opportunities could not be adequately evaluated with the existing model, and should be studied further. Based on the results of the data analysis and simulation study, this chapter provides a number of recommendations for the clinic staff to consider as they continue their process improvement initiative. This work provides in-depth information on the current state that will support further process improvement.

First, the clinic should make an effort to decrease the difference between the scheduled length of appointments and the actual length, focusing on reducing the number of appointments that last longer than expected. While an appointment that ends 5 minutes before it was scheduled to end will not cause delays in the clinic (and it may even help to keep a physician on schedule), the same appointment ending 5 minutes after it was scheduled to end will cause the physician to be behind schedule for the rest of the day unless they can make up those 5 minutes somewhere else. Lengthening the duration of appointment slots has many fiscal and access implications and would need more study to best "right-size" appointment slots to the expected visit duration. There are other ways the clinic could go about decreasing the size of this gap, including adjustments to the appointment scheduling process or further investigation into the value-added portion of each appointment.

Second, the clinic should continue to encourage patients to arrive earlier. Although the average patient arrives to the clinic about 10 minutes before his or her appointment time, the lateness is so varied that about 30% of patients complete the registration process after their appointment times. Although the clinic cannot hope to completely control when patients arrive, patient lateness patterns had a significant impact on the efficiency and effectiveness of the clinic's operations in the simulation study. Reducing the number of extremely late patients will increase the number of appointments that start on time and cause patients to spend less time in the clinic overall.

The simulation results also suggest that the clinic should further explore the possibility of introducing a lateness threshold. Although the clinic makes every effort to accommodate every patient every day, the simulation model demonstrated that asking extremely late patients to reschedule their appointments significantly reduced delays. A policy like this may also encourage more timely arrivals in the future, which has been shown in this model to have positive effects on the clinic's performance. More investigation is necessary because the models tested were basic – they only use two possible cutoffs, and they do not provide the opportunity for patients to be fit into a different appointment on the same day. The clinic may feel that adjusting the cutoff time, giving patients the chance to be rescheduled to another open appointment slot on the same day, or even allowing for exceptions (for example, if the patient's provider is running ahead of schedule or if the patient is showing urgent symptoms) would make more sense and be more aligned with the clinic's philosophies.

Finally, because the model makes several assumptions and generalizations related to the geography of the clinic and includes only some tasks the PCAs are responsible for, the availability of the PCA staff is overestimated in the model. Making changes to PCA staffing is likely to improve performance more than is indicated in the model results because of these limitations. Based on discussions with the clinic, adding PCA resources appears to be a promising option; more data collection and detailed modeling would be necessary to be confidently predict the effects of adding an additional PCA.

After significant research and data collection and analysis, building a simulation model to represent the primary care clinic allowed me to investigate how changes to the clinic would affect the efficiency with which the clinic processes its patients. The model's results identified which areas of focus would yield the most significant results, allowing me to advise the clinic staff on what kind of improvements might lead to the most improved patient experience.

Reflection on Design

The engineering design process is used to create systems or products that meet the needs of society or the customer. The process starts with defining the objectives and major needs of the project, then moves on to analysis, construction, and evaluation of the results. Industrial engineers design processes, techniques, or models, as opposed to building machines or parts. To design the simulation model used for this project, an iterative design process was used. The original model was created, discussed, and critiqued by the project team and the sponsoring physicians. The feedback was integrated into the next iteration of the model, and this cycle was repeated until the involved parties were satisfied with the resulting model.

An important consideration in the engineering design process is the understanding of and accounting for constraints. The construction of the models took into consideration some of the constraints of the clinic, most notably the staffing schedules and the allocated exam rooms for each care provider. The models' results must also be evaluated with the constraints in mind – for example, though scheduling all appointments in longer slots would decrease delays in the clinic, it would also lead to fewer appointments being scheduled. This has both fiscal and access implications, and may create other problems. This is an important financial constraint to consider. Using simulation to model the system also introduced constraints – though a simulation model is intended to represent a real-life system, it is impossible to model reality with 100% accuracy. Because simulation models operate on assumptions and generalizations, they must be evaluated as such.

Reflection on Lifelong Learning

Working on this project allowed me to extend my education beyond the traditional classroom experience by giving me the opportunity to interact with an external organization. It required me to balance the expectations of WPI and my advisor with the wants and needs of the sponsoring organization. The simulation model I created for the project was more detailed than any I had worked with previously, which allowed me to expand my knowledge of simulation modeling beyond the concepts taught in class. Over the course of this project, my presentation skills improved significantly; I am much more comfortable speaking in front of large groups and for longer periods of time than I was at the beginning of this academic year. I believe the professional skills and experience I gained through working on this project have prepared me for my professional career after graduation.

When I started work on the project, the clinic had already implemented a number of countermeasures to improve their performance. Similarly, though my work on the project has come to an end, it is clear that the clinic's quest to improve patient experience is never-ending. My work on the project has identified potential areas of focus for the clinic's future endeavors, but it is in no way the end of the clinic's pursuit of quality. This project emphasized to me that continuous improvement is truly continuous – though it may involve multiple sub-projects aimed at different areas, they are all working toward the common goal of improving overall experience for those involved in the process. The process may seem slow-moving, but over time, small improvements can compound and lead to dramatic transformation across an organization. I learned an immeasurable amount from this experience, and, like any improvement process, the learning process is also a continuous one. In order to be successful in my future endeavors, I will need to be constantly learning and improving myself – by networking, researching, or gaining a variety of experiences to make me a well-rounded and high-performing engineer.

References

25 Shocking Medical Mistakes. (n.d.). Retrieved October 19, 2015, from http://www.cnn.com/2015/09/22/health/gallery/25-shocking-medical-mistakes/

Bard, J., Shu, Z., Morrice, D., Wang, D., Poursani, R., & Leykum, L. (2014). Improving patient flow at a family health clinic.

Bodenheimer, T., Berenson, R.A., & Rudolf, P. The Primary Care–Specialty Income Gap: Why It Matters. Ann Intern Med. 2007;146:301-306. doi:10.7326/0003-4819-146-4-200702200-00011

Bodenheimer, T., & Pham, H. (2010). Primary Care: Current Problems And Proposed Solutions. *Health Affairs, 9*(5), 799-805. doi:10.1377

Clinic Physician, personal communication.

Dictionary.com. (n.d.). Retrieved September 21, 2015, from http://dictionary.reference.com/browse/throughput

Graig, L., & Perosino, K. (2011, July 1). Applying Lean to Improve the Patient Visit Process at Three Federally Qualified Health Centers. Retrieved September 12, 2015.

Clinic Data, October 2015

Jimmerson, C. (2010). *Value stream mapping for healthcare made easy*. Boca Raton: CRC Press.

Milestones in Medical Technology. (2012, October 9). Retrieved October 17, 2015, from http://www.nytimes.com/interactive/2012/10/05/health/digital-doctor.html?_r=0#/#time15_328

Onisuru T. Okotie, Neel Patel, and Chris M. Gonzalez, “The Effect of Patient Arrival Time on Overall Wait Time and Utilization of Physician and Examination Room Resources in the Outpatient Urology Clinic,”Advances in Urology, vol. 2008, Article ID 507436, 3 pages, 2008. doi:10.1155/2008/507436

Schulte, M. (2013). *Healthcare delivery in the U.S.A.: An introduction*(2nd ed.). Boca Raton, FL: Taylor & Francis/CRC Press.

Toussaint, J., & Berry, L. (n.d.). The Promise of Lean in Health Care. *Mayo Clinic Proceedings,* 74-82.

Appendix A

A3 Report in the clinic's preferred format

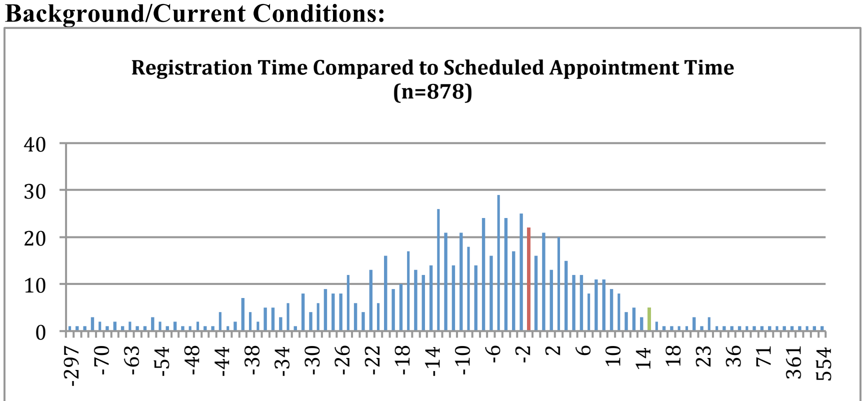
Title: Managing Lateness to Improve Process Flow in Primary Care	**Date:** 3/23/2016
Lead:	**Page:**

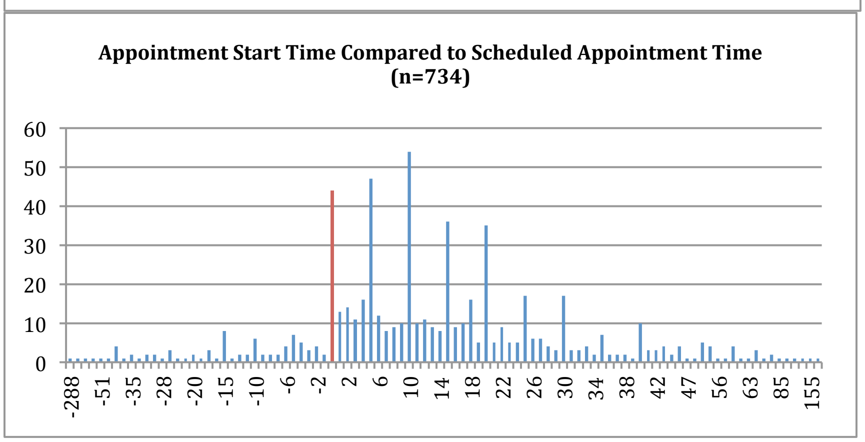
Team Members: Sarah Abell, Amy Stevens

Problem Statement: There is a high volume of patients served by the clinic and as a result, appointments are scheduled very tightly, so any lateness by the provider or patient can cause delays for the remainder of the day. When appointments begin late, it can result in either a shortened appointment, thus reducing the quality of the visit for the patient, or causes the provider to be behind schedule for the remainder of the day, often requiring them to stay late to catch up.

Scope (In/Out): We focused on patients who have appointments with attending physicians in the primary care clinic, from the time of their arrival to the time they finish meeting with their provider.

Background/Current Conditions:





Though data collected from the clinic's scheduling and registration system shows that the majority of patients are registering before their appointment time, data collected by physicians indicates that the majority of appointments are starting after their scheduled start times.

Root Causes/Areas of Focus:

- Patient Arrival Patterns
- PCA Staffing Schedules
- Difference between scheduled appointment lengths and time required by patient
- Policies regarding patients who arrive late.

Goals: Analyze the effects of delays on the clinic's functioning and use simulation to identify the areas in which changes are most likely to significantly improve the flow of patients through the clinic.

Plan · Do · Study · Act (PDSA)

Countermeasures (Plan):

The first step was to gain an understanding of the clinic's flow and performance, both qualitatively and qualitatively. This understanding allowed the project team to identify areas in which there was an opportunity for improvement.

Implementation (Do):

The team constructed a simulation model to represent the clinic and tested a series of alternate models against the base case to determine the effects of simulated changes. Models in the following areas were tested:

- Patient arrival patterns
- The staff schedules of the Patient Care Associates (PCAs)
- The duration of appointments relative to the scheduled duration
- Asking late patients to reschedule their appointments for another day.

Results/Conclusion (Study):

The results from the simulation suggested that patient arrivals and differences between scheduled appointment lengths and actual appointments had the most significant effects on the clinic's functioning. The model also suggested that a policy that asks extremely late patients to reschedule their appointments has the potential to have a positive effect on appointment start time and overall appointment time.

Follow-up Actions (Act):

Based on the results, the clinic should first focus their efforts on reducing the difference between scheduled and actual appointment durations, since the results from the simulation indicated that the positive effects from this group of scenarios were likely to be the most significant.

Appendix B
Data collection sheet handed out to patients in November 2015

Primary Care Clinic Time Study

We are collecting data about the appointment process in hopes of reducing wait times and improving patient experience. Your participation is completely voluntary and anonymous. Please complete this form as you move through each step of the process, and make sure to use the same device (watch, clock, cell phone) to keep time throughout your visit..

Your participation is greatly appreciated. After your visit ends, please return this sheet to the check-out desk.

Today's Date:	
Scheduled appointment time:	
Physician name:	

Please write what time the following things happened:	
Got in line at the registration desk	
Started speaking with registration staff	
Left the registration desk	
A medical assistant came to get you for vital signs (height, weight, etc)	
Left the area where vital signs were checked	
Entered the exam room	
Physician entered the exam room	
Physician completed the visit and exited the room	

Thank you very much for your participation. Enjoy the rest of your day!