Submitted to

Using Analytics to Improve Patient Flow in Outpatient Clinics

(Authors' names blinded for peer review)

Demands for increased capacity and reduced costs in outpatient settings creates the need for a coherent strategy regarding how to collect, analyze, and use data to facilitate and lead to process improvements. Specifically, this note focuses upon system performance related to patient flows in outpatient clinics in Academic Medical Centers that schedule patients by appointments. We describe ways to map these visits as we map processes, collect data to formally describe the system, create discrete event simulations of these systems, use the simulation as a virtual lab to explore possible system improvements, and identify proposals as candidates for implementation. We close with a discussion of several projects in which we have used our approach to understand and improve these complex systems.

1. Introduction

As of 2016 the Affordable Care Act (ACA) had extended access to health insurance coverage to roughly 30 million previously uninsured americans and that coverage expansion is linked to between 15 and 26 million additional primary care visits annually. (Glied and Ma (2015), Beronio et al. (2014)) In addition, the number of people 65 and older in the US is expected to grow from 43.1 million in 2012 to 83.7 million by 2050. (Ortman et al. (2014)) This jump in the number of insured Americans coupled with the anticipated growth in the size of the population above the age of 65 will correlate with rising demand for healthcare services.

At the same time, Medicare and other payers are moving away from the older ”fee for service” model towards ”bundled payment” schemes (Cutler and Ghosh (2012)). Under these arrangements providers are paid a lump sum to treat a patient or population of patients. This fixes patient related revenue and means that these payments can only be applied to fixed costs if variable costs are less than the payment. We expect the continued emergence of bundled payment schemes to accelerate the gradual move away from inpatient treatment to the delivery of care through outpatient settings that has been taking place for over 20 years. Consequently, a disproportionate share of the growth in demand will be processed through outpatient clinics, as opposed to hospital beds. This evolution is also seen as one of the key strategies needed to help get healthcare cost in the US down closer to the costs experienced in other developed countries. (Lorenzoni et al. (2014))

An additional complicating factor is that healthcare delivery in the US is often interspersed with teaching and training of the next generation of care providers. In 2007 roughly 40 million outpatient visits were made to teaching hospitals known as Academic Medical Centers (AMC). (Hing et al. (2010)) Inclusion of the teaching component within the care process dramatically increases the complexity of each patient visit. The classic model of an outpatient visit where a nurse leads the patient to an examination room, the patient is next seen by the physician, and then leaves the clinic is not a sufficient description of the process in the AMC. Adding a medical resident or fellow (Trainee) into the process introduces steps for the interaction between the Trainee and the patient as well as interactions between the Trainee and the Attending Physician (Attending). These added steps increase flow times, the number and levels of resources deployed, and system congestion (Boex et al. (2000), Franzini and Berry (1999), Hosek and Palmer (1983), Hwang et al. (2010)). The delays added are easy to understand when one considers the fact that the Trainee typically takes longer then the Attending to complete the same task and many teaching settings demand that both the Trainee and the Attending spend time with each patient on the clinic schedule. (Williams et al. (2007), Taylor et al. (1999), Sloan et al. (1983))

The addition of the teaching mission is not simply adding steps to a well managed process. The added complexity is akin to changing from a single server queueing system to a hybrid system. (Williams et al. (2012, 2015)) The Trainee may function as a parallel (but slower) server, or the Trainee and Attending may function as serial servers such that a one step activity becomes a two step process, or decisions on how the Trainee is intertwined in the process may be made dynamically, meaning that the Trainee's role may change depending on system status.

In short we are asking our current healthcare system to improve access to care, to a rapidly growing and aging population as demand is shifted from inpatient to outpatient services in teaching hospitals using delivery models that are not well understood. While the extent to which this is even possible is debatable (Moses et al. (2005)) it is quite clear that efforts to make this workable require thoughtful data analysis and extremely high quality operations management. (Sainfort et al. (2005))

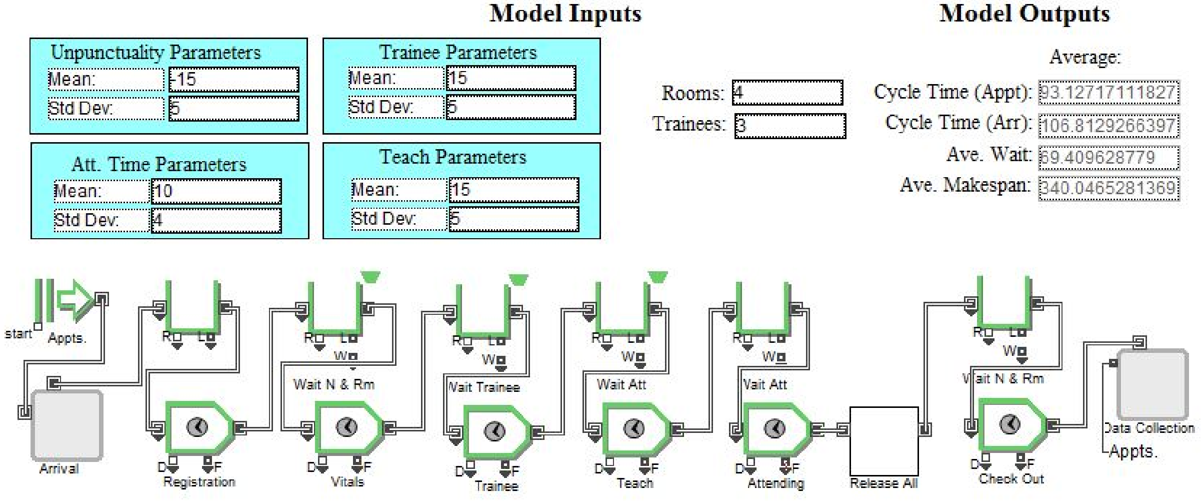
The primary objective of this chapter is to lay out a strategy toward gaining an understanding of these complex systems, identifying means to improve their performance, and predicting how proposed changes will effect system behavior. We present this in the form of a 6-step process and provide some details regarding each step. We close with a discussion of several projects in which our process has been applied.

1.1. A Representative Clinic

To make the remainder of our discussion more concrete let us introduce a representative unit of analysis. Data associated with this unit will be taken from a composite of clinics that we have studied, but is not meant to be a complete representation of any particular unit. Consider a patient with an appointment to see the Attending at a clinic within an AMC. We will work with a Discrete Event Simulation (DES) of this process. DES is the approach of creating a mathematical model of the flows and activities present in a system and using this model to perform virtual experiments seeking to find ways to improve measurable performance. (Benneyan (1997), Clymer (2009), Hamrock et al. (2013), Jun et al. (1999)) A screen-shot from such a DES is presented in Figure 1, and will double as a simplified process map. By simplified, we mean that several of the blocks shown in the figure actually envelope multiple blocks that handle details of the model. (Versions of this and similar models along with exercises focused on their analysis and use are included as an appendix.) Note that the figure also contains a sample of model inputs and outputs from the simulation itself. We will discuss several of these metrics shortly.

In this depiction a block "Creates" work units (patients) according to an appointment schedule. The block labeled "Arrival" combines these appointment times with a random variable reflecting

Figure 1: Process Map of Clinic Visit



Model Inputs

Unpunctuality Parameters
Mean: -15
Std Dev: 5

Trainee Parameters
Mean: 15
Std Dev: 5

Att. Time Parameters
Mean: 10
Std Dev: 4

Teach Parameters
Mean: 15
Std Dev: 5

Rooms: 4
Trainees: 3

Model Outputs

Average:
Cycle Time (Appt): 93.12717111827
Cycle Time (Arr): 106.8129266397
Ave. Wait: 69.409628779
Ave. Makespan: 340.0465281369

Figure 1 **Copied from DES of Representative Clinic**

patient unpunctuality to get actual arrival times. Once created, the patients move to Step 1. Just above Step 1 we show a block serving as a queue just in case the resources at Step 1 are busy. In Step 1 the patient interacts with staff at the front desk. We will label this step ”Registration” with the understanding that it may include data collection, and perhaps some patient education. In Step 2 a Nurse leads the patient into an examination room, collects data on vital signs, and asks a few questions about the patient’s condition. We will label this step ”Vitals”. In Step 3 a Trainee reviews the patient record and enters the examination room to interact with the patient. We label this step ”Trainee”. In Step 4, the Trainee leaves the exam room and interacts with the Attending. We label this step ”Teach”. During this time the Trainee may ”present” case information to the attending and the pair discuss next steps, possible issues, and needs for additional information. In Step 5 the Trainee and Attending both enter the exam room and interact with the patient. We label this step ”Attending”. Following this step, the Trainee, Attending, and Room are ”released” meaning they are free to be assigned to the next patient. Finally the patient returns to the front desk for ”Check Out.” This step may include collection of payment, and making an appointment for a future visit.

In order to manage this system we need an understanding of its behavior. This behavior will be reflected in quantifiable metrics such as, cycle times, waiting times and how long it will take to complete the appointment schedule (makespan). Note that Cycle Times may be calculated based on Appointment times or patient Arrival times. Both of these values are included among the Model Outputs shown here. While this model is fairly simple, some important questions may be addressed with its use. For example, we may make different assumptions regarding the Attending's processing time and note how this changes the selected output values. This is done by altering the parameters labeled ”Att. Time Parameters” among the model inputs. For this illustration we assume that these times are drawn from a log-normal distribution and the user is free to change the mean and standard deviation of that distribution. However, one benefit of simulation is that we may use a different distribution or sample directly from collected activity time data. We will discuss these issues later. This model can also be used as part of a more holistic approach to address more subtle questions including; how does the added educational mission affect output metrics, and what is the best appointment schedule for this system? In the next section we lay out a more complete approach to handle more complex questions such as these.

2. How to Fix Healthcare Processes

Much of the earliest development of research oriented universities in the US was driven by the needs for research related to healthcare (Chesney (1943)). Consequently, when working with physicians and other healthcare professionals in the AMC a convenient starting point for the discussion is already in place. Research in most parts of healthcare addresses questions using randomized trials or pilot implementations. These typically center on formal experiments which are carefully designed and conducted in clinical or laboratory settings. This experiment-based approach to research has proven to be highly effective and is assumed by many to be the best way to produce evidence based results on medical questions relating to issues including the efficacy of a new drug or the efficiency of a new technique. One way to get buy in from practitioners in the AMC is to take a very similar approach to issues related to patient flow.

At the same time Operations Research (*OR*) has a long history of using tools to improve service delivery processes. *OR* employs a predictive modeling investigative paradigm that uses mathematical equations, computer logic, and related tools to forecast the consequences of particular decision choices (Sainfort et al. (2005)). Typically, this is done in abstraction without a formal experiment. This approach permits the consideration of alternative choices to quickly be evaluated and compared to see which are most likely to produce preferred outcomes. Many traditional areas of OR are prevalent in clinic management. These topics include appointment scheduling (Cayirli et al. (2006)), nurse rostering problems (Burke et al. (2004)), resource allocation problems (Chao et al. (2003)), capacity planning (Bowers and Mould (2005)), and routing problems (Mandelbaum et al. (2012)).

Given this confluence of approaches and needs, it seems natural for those working to improve healthcare processes to employ OR techniques such as DES to conduct controlled, virtual experiments as part of the improvement process. However, when one looks more closely one finds that the history of implementations of results based on OR findings in AMC's is actually quite poor. For example, a review of over 200 papers that use DES in Healthcare settings identified only 4 that even claim that physician behavior was changed as a result. (Wilson (1981)) A more recent review found only 1 instance of a publication which included a documented change in clinic performance resulting from a simulation-motivated intervention. (van Lent et al. (2012))

This raises a major question; since there is clearly an active interest in using DES models to improve patient flow and there is ample talent working to make it happen, what can we do to make use of this technique in a way that results in real change in clinic performance. Virtually any Operations Management textbook will provide a list of factors needed to succeed in process improvement projects such as getting all stakeholders involved early, identifying a project champion, setting clear goals, dedicating necessary resources, etc. (Trusko et al. (2007)) However, we want to focus this discussion on two additional elements that are a bit more subtle and, in our experience, often spell the difference between success and failure when working in outpatient clinics in the AMC.

First finding an important problem is not sufficient. It is critically important to think in terms of finding the right question which also addresses the underlying problem. As outside agents or consultants, we are not in a position to pay faculty and staff extra money to implement changes to improve the system. We need a different form of payment to motivate their participation. One great advantage in the AMC model is that we can leverage the fact that physicians are also dedicated researchers. Thus, we can use the promise of publications in lieu of a cash payment to induce participation.

Second, we need to find the right combination of techniques. Experiments and data collection resonate with medical researchers. However, the translation from ”lab” to ”clinic” is fraught with confounding factors outside of the physician's control. On the other hand, OR techniques can isolate a single variable or factor but, modeling by itself does not improve a system, and mathematical presentations that feel completely abstract do not resonate with practitioners. The unique aspect of our approach is to combine OR tools with ”clinical” experiments. This allows clinicians to project themselves into the model in a way that is more salient then the underlying equations could ever be. The key idea is that value exists in finding a way to merge the tools of OR with the methodologies of medical research to generate useful findings that will actually be implemented to improve clinic flow.

3. The Process Improvement Process

Given this background, we need a systematic approach to describing, analyzing, and predicting improvements in performance based on changes that can be made to these systems. In order to do this we need to accomplish at least six things, and this list forms the statement of our method.

1. Describe processes that deliver care and/or service to patients in a relevant way
2. Collect data on activity times, work flows, and behavior of key agents
3. Create a DES of the system under study
4. Experiment with both real and virtual systems to identify and test possible changes
5. Develop performance metrics of interest to both patients and care providers
6. Predict changes in metrics which stem from changes in process

We now turn to providing a bit more detail about each of these steps.

3.1. Step 1: Process Description

Much has been written concerning process mapping in healthcare settings. (Trusko et al. (2007), Trebble et al. (2010)) In many instances the activity of process mapping itself suggests multiple changes that may improve process flow. However, some insights related to the healthcare-specific complications to this activity warrant discussion.

Perhaps the most obvious way to develop a process map is to ask the agents in the system to describe the work flow. We have found that this is absolutely necessary and serves as an excellent starting point but is never sufficient. Agents in the system often provide misleading descriptions of process flow. In many cases physicians are not fully aware of what support staff do to make the process work and Trainees and staff are often quite careful to not appear to contradict more senior physicians. To get high quality process descriptions we must gather unbiased insights from multiple levels of the organization. Ideally, this will include support staff, nursing, Trainees, and Attendings. In some cases other administrators are valuable as well, especially if there is a department manager or some other person who routinely collects and reports performance data. It is ideal to have all of these agents working on the development of a process map as a group. However, if this cannot be done it is even more vital to carefully gather information about process flows from as many different angles as possible.

Second, we have found that no matter how much information about the process has been gathered, direct observation by outside agents working on the process improvement process is ALWAYS required. We have yet to find a process description created by internal agents that completely agreed with our observations. Health care professionals (understandably) put patient care above all other considerations. Consequently, they make exceptions to normal process flows routinely and do not give them a second thought. As a result, their daily behavior will almost always include several subtleties that they do not recall when asked about process flow.

3.2. Step 2: Data Collection

In our experience, this is the most time consuming step in the improvement process. Given a process map, it will be populated with some number of activities undertaken by various agents. The main question that has to be asked at this stage is how long each agent spends to complete each step. This approach makes sense for several reasons. First, the dominant patient complaint in outpatient settings is waiting times. Thus, time is a crucial metric from the patient's perspective. Second, many costing systems have been developed which accumulate costs based on hourly or minute by minute charges for various resources. (Kaplan and Anderson (2003), Kaplan and Porter (2011), King et al. (1994)) Consequently, time is a crucial metric from the process manager's perspective as well. Therefore, how long each step takes becomes the central question of interest.

We have utilized four ways to uncover this information. Agents within the system can be asked how long a process step takes. This is useful as a starting point and can be sufficient in some rare instances. On the other hand, quizzing agents about activity times is problematic because most people think in terms of averages and find it difficult to measure variances. This can only be done after a sufficient number of observations are in hand.

We have also used approaches in which the care givers record times during patient visits. For example, in one clinic we attached a form to each patient record retrieved during each clinic session. In Step 1 staff at the front desk record the patient arrival time and appointment time. The Nurse then records the start and end of Step 2, and so on. This approach can be automated through the use of aids such as phone or ipad apps where applicable. However, this approach introduces several issues. Recording data interrupts normal flow and it is not possible to convince the participants that data recording compares in importance to patient care. As a consequence, we repeatedly see instances where the providers forget to record the data and then try to "fill it in" later in the day when things are less hectic. This produces data sets where mean times may be reasonable estimates but the estimates of variances are simply not reliable.

A third approach to data collection often used in AMCs is to use paid observers to record time stamps. This approach can generate highly reliable information as long as the process is not "too" complex and the observer can be physically positioned to have lines of sight that make this method practical. This approach is common in AMC's because they are almost always connected to a larger university and relatively high quality, low cost labor is available in the form of students or volunteers. While we have used this technique successfully on multiple occasions, it is not without its problems. First, the observers need to be unobtrusive. This is best done by having them assigned to specific spaces. If personnel travel widely, this becomes problematic. For example, a Radiation Oncology clinic that we studied had rooms and equipment on multiple floors so tracking becomes quite complex. Second, the parties serving patients know they are being observed. Many researchers have reported significant improvements to process flow using this approach, only to find that after the observers left, the system drifted back to its previous way of functioning and the documented improvement was lost.

We have also used a fourth approach to data collection. Many hospitals and clinics are equipped with Real-Time Location Systems (RTLS). Large AMC's are often designed to include this capability because tracking devices and equipment across hundreds of thousands of square feet of floor space is simply not practical without some technological assistance. Installations of these systems typically involve placing censors in the ceilings or floors of the relevant spaces. These sensors pick up signals from transmitters that can be imbedded within "tags" or "badges" worn by items or people being tracked. Each sensor records when a tag comes within range and again when it leaves that area. When unique tag numbers are given to each care giver, detailed reports can be generated at the end of each day showing when a person or piece of equipment moved from one location to another. This approach offers several dramatic advantages. It does not interfere with the care delivery process, the marginal cost of using it is virtually 0, and since these systems are always running, the observation periods can begin and end as needed.

In closing we should highlight three key factors in the data collection process; 1) data collection needs to be done in a way that does not interfere with care delivery; 2) audits of the data collection system are needed to insure accuracy; and 3) a sufficient time span must be covered to eliminate any effects of the "novelty" of the data collection and its subsequent impact on agent behaviors.

3.3. Step 3: Create a DES of the System

We have often found it useful to create DES models of the systems under study as early in the process as possible. This can be a costly process in that a great deal of data collection is required and model construction can be a non-trivial expense. Other tools such as process mapping and queueing theory can be applied with much less effort. (Kolker (2010)) However, we have repeatedly found that these tools are insufficient for the analysis that is needed for several reasons. Because the variances involved in activity times can be extremely high in healthcare, distributions of the metrics of interest are important findings. Consequently, basic process analysis is rarely sufficient and often misleading.

Queueing models do a much better job of conveying the significance of variability. However, many common assumptions of these models are routinely violated in clinic settings. These include: some processing times are not exponentially distributed; processing times are often not from the same distribution, and; if arrivals are based on appointments, inter-arrival times are not exponentially distributed.

However, none of these issues pose the largest challenge to applying simple process analysis or queuing models in outpatient clinics. Consider three additional issues. First, the basic results of process analysis or queueing models are only averages which appear in steady state. A clinic does not start the day in steady state - it begins in an empty state. It takes some time to reach steady state. However, if one plots average waiting times for a clinic over time, one quickly sees that it may take dozens or even hundreds of cases for the system to reach steady state. Clearly a clinic with one physician is not going to schedule hundreds of patients for that resource in a single session. Thus steady state results are often not informative.

Second, if activity times and/or the logic defining workflow change in response to job type or system status, then the results of simple process analysis or queueing models become invalid. We have documented such practices in multiple clinics that we have studied. (Chambers et al. (2016), Conley et al. (2016)) Consequently, what is needed is a tool that can account for all of these factors simultaneously, make predictions about what happens when some element of the system changes, and can give us information about the broader distribution of outcomes - not just means for systems in steady state. DES is a tool with the needed capabilities.

A brief comment on the inclusion of activity times in DES models is warranted here. We have used two distinct approaches. We can select an activity time at random from a collection of observations. Alternatively, we can fit a distribution to collected activity time data. We have found both approaches to work satisfactorily. However, if the data set is sufficiently large, we recommend sampling directly from that set. This generates results that are both easier to defend to statisticians and more credible to practitioners.

3.4. Step 4: Field and Virtual Experiments

It is at this point that the use of experiments comes into play and we merge the OR methodology of DES with the experimental methods of medical research. The underlying logic is that we propose an experiment involving some process change that we believe will alter one or more parameters defining system behavior. We can use the DES to predict outcomes if our proposal works. In other cases, if we have evidence that the proposed change works in some settings, we can use the DES to describe how that change will effect system metrics in other settings. The construction of these experiments is the ”art” of our approach. It is this creation that leads to publishable results, and creates novel insights.

We will provide examples of specific experiments in the next section. However, at this juncture we wish to raise two critical issues; confounding variables, and unintended consequences. Confounding variables refers to system or behavioral attributes that are not completely controlled when conducting an experiment but can alter study results. For example, consider looking at a system before an intervention, collecting data on its performance, changing something about the system, and then collecting data on the performance of the modified system. This is the ideal approach, but implicitly assumes that nothing changed in the system over the span of time of study other than what you intended to change. If data collection takes place over a period of months, it is quite possible that the appointment schedule changed over that span of time due to rising or falling demand. In this example, the change in demand would be a confounding variable. It is critically important to eliminate as many confounding variables as you can before concluding that your process change fully explains system improvement. DES offers many advantages in this regard because it allows you to fix some parameter levels in a model even if they may have changed in the field.

It is also critical to account for unintended consequences. For example, adding examination rooms is often touted as a way to cut waiting times. However, this also makes the relevant space larger, increasing travel times as well as the complexity of resource flows. This must be accounted for before declaring that the added rooms actually improved performance. It may improve performance along one dimension while degrading it in another.

DES modeling has repeatedly proven invaluable at this stage. Once a DES model is created, it is easy to simulate a large number of clinic sessions and collect data on a broad range of performance metrics. With a little more effort it can also be set up to collect data on use of overtime, or waiting times within examination rooms. In addition, DES models can be set up to have patients take different paths or have activity times drawn from different distributions depending on system status. Finally, we have found it useful to have DES models collect data on subgroups of patients based on system status because many changes to system parameters affect different groups differently.

3.5. Step 5: Metrics of Interest

As one famous adage asserts, ”if you can’t measure it you can’t manage it.” Focusing on measurements removes ambiguity and limits misunderstandings. If all parties agree on a metric then it is easier for them to share ideas about how to improve it. However, this begs an important question - what metrics do we want to focus on? In dealing with this question Steps 4 and 5 of our method really become intertwined and cannot be thought of in a purely sequential fashion. In some settings we need novel metrics to fit an experiment, while in other settings unanticipated outcomes from experiments suggest metrics that we had not considered earlier.

Both patients and providers are concerned with system performance, but their differing perspectives create complex tradeoffs. For example, researchers have often found that increases in face time with providers serves to enhance the patient experience, (Thomas et al. (1997), Seals et al. (2005), Lin et al. (2001)) but an increase in waiting times degrades that experience. (Meza (1998), McCarthy et al. (2000), Lee et al. (2005)) The patient may not fully understand what the care provider is doing but they can always understand that more attention is preferable and waiting for it is not productive. Given a fixed level of resources, increases in face time result in higher provider utilization, and higher utilization increases patient waiting times. Consequently, the patient's desire for increased face time and reduced waiting time creates a natural tension and suggests that the metrics of interest will almost always include both face time and waiting time.

Consider one patient that we observed recently. This patient arrived 30 minutes early for an appointment, and waited 20 minutes before being lead to the exam room. After being lead to the room the patient waited for 5 minutes before being seen by a Nurse for 5 minutes. The patient then waited 15 minutes before being seen by the Resident. The Trainee then spoke with the patient for 20 minutes before leaving the room to discuss the case with the Attending. The patient then waited 15 minutes before being seen by the Resident and the Attending working together. The Attending spoke with the patient for 5 minutes before being called away to deal with an issue for a different patient. This took 10 minutes. The Attending then returned to the exam room and spoke with the patient for another 5 minutes. After that the patient left. By summing these durations we see that the patient was in the clinic for roughly 100 minutes. The patient waited for 20 minutes in the waiting room. However, the patient also spent 45 minutes in the exam room waiting for service. Time in the examination room was roughly 80 minutes of which 35 minutes was spent in the presence of a service provider. Thus we may say that face time was 35 minutes. However, of this time only 10 minutes was with the attending physician. Consideration of this more complete description suggests a plethora of little-used metrics that may be of interest such as:

1. Patient Punctuality
2. Time spent in the waiting room before the appointment time
3. Time spent in the waiting room after the appointment time

4. Waiting time in the Examination Room
5. Proportion of Cycle Time spent with a care provider
6. Proportion of Cycle Time spent with the Attending

The key message here is that the metrics of interest may be specific to the problem that one seeks to address and must reflect the nuances of the process in place to deliver the services involved.

3.6. Step 6: Predict Impact of Process Changes

Even after conducting an experiment in one setting we have found that it is extremely difficult to predict how changes will affect a different system simply by looking at the process map. This is another area where DES proves quite valuable. For example, say that our experiment in Clinic A shows that by changing the process in some way, the time for the Attending step is cut by 10%. We can then model this change in a different clinic setting by using a DES of that setting to predict how implementing our suggested change will be reflected in performance metrics of that clinic in the future. This approach has proven vital to get the buy-in needed to facilitate a more formal experiment in the new setting or to motivate implementation in a unit where no formal experiment takes place.

4. Experiments, Simulations, and Results

Our work has included a collection of experiments that have led to system improvements for settings such as that depicted in Figure 1. We now turn to a discussion of a few of these efforts to provide context and illustrations of our approach. Figure 1 includes an arrival process under an appointment system. This is quickly followed by activities involving the Trainee and/or Nurse and/or Attending. Finally, the system hopes to account for all of these things when searching for an optimized schedule. We discuss a few of these issues in turn.

4.1. Arrival Process

We are focusing on clinics which set a definite appointment schedule. One obvious complication is that some patients are no-shows, meaning that they do not show up for the appointment. No-show rates of as much as 40% have been cited in prior works. (McCarthy et al. (2000), Huang (1994))

However, there is also a more subtle issue of patients arriving very early or very late and this is much harder to account for. Early work in this space referred to this as patient ”unpunctuality.” (Bandura (1969), Blanco White and Pike (1964), Alexopoulos et al. (2008), Fetter and Thompson (1966), Tai and Williams (2012), Perros and Frier (1996)) Our approach has been used to address two interrelated questions. Does patient unpunctuality affect clinic performance, and can we affect patient unpunctuality? To address these questions we conducted a simple experiment. Data on patient unpunctuality was collected over a 6 month period. We found that most patients arrive early but patient unpunctuality ranged from -80 to +20. In other words some patients arrived as much as 80 minutes early while others arrived 20 minutes late. An intervention was performed that consisted of 3 elements. In reminders mailed to each patient before their visit, it was stated that late patients would be asked to reschedule. All patients were called in the days before the visit and the same reminder was repeated over the phone. Finally, a sign explaining the new policy was posted near the registration desk. Unpunctuality was then tracked 1 month, 6 months, and 12 months later. Additional metrics of interest were waiting times, use of overtime, and the proportion of patients that were forced to wait to be seen. (Williams et al. (2014))

This lengthy follow up was deemed necessary because some patients only visit the clinic once per quarter, thus the full effect of the intervention could not be measured until after several quarters of implementation. To ensure that changes in clinic performance were related only to changes in unpunctuality we needed a way to control for changes in the appointment schedule that happened over that time span. Our response to this problem was to create a DES of the clinic, use actual activity times in the DES, and consider old versus new distributions of patient unpunctuality. This allowed us to isolate the impact of our intervention.

Before the intervention 7.7% of patients were tardy and average tardiness of those patients was 16.75 minutes. After 12 months, these figures dropped to 1.5% and 2 minutes respectively. The percentage of patients who arrived before their appointment time rose from 90.4% to 95.4%. The proportion who arrived at least 1 minute tardy dropped from 7.69% to 1.5%. The range of unpunctuality decreased from 100 minutes to 58 minutes. The average time to complete the session dropped from 250.61 minutes to 244.49 minutes. Thus, about 6 minutes of overtime operations was eliminated from each session. The likelihood of completing the session on time rose from 21.8% to 31.8%.

Our use of DES allowed us to create metrics of performance that had not been explored earlier. For example, we noticed that the benefits from the change were not the same for all patients. Patients that arrived late saw their average wait drop from 10.7 minutes to 0.9 minutes. Those that arrived slightly early saw their average waiting time increase by about 0.9 minutes. Finally, for those that arrived very early, their waiting time was unaffected. In short, we found that patient unpunctuality can be affected and it does alter clinic performance, but this has both intended and unintended consequences. The clinic session is more likely to finish on time and overtime costs are reduced. However, much of the benefit in terms of waiting times is actually realized by patients that still insist on arriving late.

4.2. Physician Processing Times

Historically, almost all research on outpatient clinics assumed that processing times were not related to the schedule or whether the clinic was running on time. Is this indeed the case? To address this question we analyzed data from three clinic settings. One was a low volume clinic that housed a single physician, one was a medium volume clinic in an AMC that had one Attending working on each shift along with 2 or 3 Trainees. The last was a high volume service that had multiple Attendings working simultaneously. (Chambers et al. (2016))

We categorized patients into three groups: Group A patients were those who arrived and were placed in the examination room before their scheduled appointment time; Group B patients were those who arrived before their appointment times, but were placed in the examination room after their appointment time, indicating that the clinic was congested; and, Group C patients were those who arrived after their appointment time. The primary question was whether the average processing time for patients in Group A was the same as for patients in Group B. We also had questions about how this affected clinic performance in terms of waiting times, and session completion times.

In the low volume clinic with a single physician average processing times and standard errors are 38.31 (3.21) for Group A and 26.23 (2.23) for Group B. In other words, the physician moves faster when the clinic is behind schedule. Similar results had been found in other industries, but this was the first time (to the best of our knowledge) that this had been demonstrated for outpatient clinics.

In the medium volume clinic the relevant values were 65.59 (2.24) and 53.53 (1.97). Again, the system works faster for Group B then it does for Group A. Note, the drop in average times is about 12 minutes in both settings. This suggests that the finding is robust, meaning that it occurs to a similar extent in similar (but not identical) settings. Additionally, remember that the medium volume clinic included Trainees in the process flow. This suggests that the way that the system gets this increase in speed may be different. In fact, our data show that the average amount of time the Attending spends with the patient was no more than 12 minutes to begin with. Thus, we know that it is not just the behavior of the Attending that makes this happen. The AMC must be using the Trainees differently when things fall behind schedule.

In the high volume clinic, the parallel values were 47.15 (0.81) and 17.59 (0.16). Here we see that the drop in processing times is much more dramatic than we saw before. Again, the message is that processing times change when the system is under stress and the magnitude of the change implies that multiple parties are involved it making this happen. In hindsight, this seems totally reasonable, but the extent of the difference is still quite startling.

As we saw in the previous section, there is an unintended consequence of this system behavior as it related to patient groups. Patients that show up early should help the clinic stay on schedule. This may not be so because these patients receive longer processing times. Thus their cycle times are longer. Patients that arrive late, have shorter waiting times and shorter processing times. Thus their cycle times are shorter. If shorter cycle times are perceived as a benefit, this seems like an unfair reward for patient tardiness, and may explain why it will never completely disappear.

4.3. Impact of the Teaching Mission

The result from the previous section suggests that the way that the Trainee is used and managed within the clinic makes a difference when considering system performance. To explore this point further we wanted to compare a clinic without Trainees with a similar clinic that included Trainees. This is difficult to do as an experiment, but we were lucky when looking at this question. An Attending from a clinic with no Trainees was hired as the director of a clinic in the AMC that included Trainees. Thus we could consider the same Attending, seeing the same patients in both settings. One confounding variable was that the two clinics used different appointment schedules. (Williams et al. (2012))

We collected data on activity times in both settings. Given these times we could seed DES models of both clinics and compare results. Within the DES we could look at both settings as though they had the same appointment schedule. If we consider the two settings using the schedule in place for the AMC we see that the average Cycle Time in the AMC was 76.2 minutes and this included an average waiting time of 30.0 minutes. The average time needed to complete a full schedule was 291.9 minutes. If the same schedule had been used in the private practice model, the average cycle time would be 129.1 minutes and the average waiting time would be 83.9 minutes.

The capacity of the AMC is clearly greater than it is in the private practice model. This is interesting because the flow times in the private practice setting using the schedule that was optimized for that setting were much lower. It turns out that the total processing time for each patient is greater in the AMC but the capacity is higher. This is explained by the use of parallel processing. In the AMC setting the Attending spends time with one patient while Trainees simultaneously work with other patients. We were able to conduct a virtual experiment by changing the number of Trainees in the DES model. We found that having 1 Trainee created a system with cycle times that were much greater than the private practice model. Using 2 Trainees produced cycle times that were about the same. Using 3 Trainees created the reduced cycle times that we noticed in practice. Using more than 3 Trainees produced no additional benefit because both clinics had only

3 available exam rooms. This enabled us to comment on the optimal number of Trainees for a given clinic.

The use of DES also highlighted a less obvious result. It turns out that the waiting time in this system is particularly sensitive to the time taken in the step we labeled ”Teach”. This is the time that the Trainee spends interacting with the Attending after interacting with the patient. In fact, we found that reducing this time by 1 minute served to reduce average waiting time by 3 minutes. To understand this phenomenon, recall that when the Trainee and the Attending are discussing the case while the patient waits in the examination room for one minute the three busiest resources in the system; the Trainee, the Attending, and the examination room are simultaneously occupied for that length of time. Thus it is not surprising that waiting times are sensitive to the duration of this activity, although the degree of this sensitivity is still eye opening.

4.4. Pre-processing

Given that waiting times are extremely sensitive to teaching times, we created an experiment designed to alter the distribution of these times. Instead of having the Trainee review the case after the patient is placed in the examination room and then having the first conversation about the case with the Attending after the Trainee interacts with the patient, we can notify both the Trainee and Attending in advance which patient each Trainee will see. That way the Trainee can review the file before the session starts and have a conversation with the Attending about what should happen upon patient arrival. We also created a template to guide the flow and content of this conversation. We refer to this approach as ”pre-processing”. (Williams et al. (2015))

We recorded activity times using the original system for 90 days. We then introduced the new approach and ran it for 30 days. During this time we continued collecting data on activity times. Before the intervention was made the average Teach time was 12.9 minutes for new patients and 8.8 minutes for return patients. The new approach reduced these times by 3.9 minutes for new patients and 2.9 minutes for return patients. Holding the schedule as a constant we find that average waiting times drop from 36.1 minutes to 21.4 minutes and the session completion time drops from 275.6 minutes to 247.4 minutes.

However, in this instance, it was the unintended consequences that proved to be more important. When the Trainees had a more clearly defined plan about how to handle each case, their interactions with the patients became more efficient. The Trainees also reported that they felt more confident when treating the patients than they had before. While it is difficult to measure this effect in terms of times, both the Trainees and the Attending felt that the patients received better care under the new protocol.

4.5. Cyclic Scheduling

Considering the works mentioned above, one finding that occurred repeatedly was that the way the Trainee was involved in the process had a large impact on system performance, and how that was done was often state dependent. Recall that we found that the system finds ways to move faster when the clinic is behind schedule. When a physician is working alone, this can be done simply by providing less face time to patients. When the system includes a Trainee, an additional response is available in that either the Attending or the Trainee can be dropped from the process for one or more patients. Our experience is that doctors strongly believe that the first approach produces a huge savings and they strongly oppose doing the second.

Our direct observation of multiple clinics produced some insights related to these issues. First, omitting the Attending does not save as much time as most Attendings think. The Trainee is slower than the Attending. In addition, the Attending gets involved in more of these cases than they seem to realize. Many Attendings feel compelled to "at least say hi" to the patient even when the patient is not really on their schedule, and these visits often turn out to be longer than expected. Regarding the second approach, we have noticed a huge variance in terms of how willing the Attending is to omit the Trainee from a case. Some almost never do it while others do it quite often. In one clinic we studied, we found that the Trainee was omitted from roughly 30% of the cases on the clinic schedule. If this is done, it might explain why a medium volume or high volume clinic within the AMC could reduce cycle times after falling behind schedule to a greater extent than the low volume clinic can achieve. This can be done by instructing the Trainee to handle one case while the Attending handles another and having the Attending exclude the Trainee from one or more cases in an effort to catch up to the clinic schedule.

Accounting for these issues when creating an appointment schedule led us to the notion of Cyclic Scheduling. This idea is that the appointment schedule can be split into multiple subsets which repeat. We label these subsets ”cycles”. In each cycle we include one New patient and one Return patient scheduled to arrive at the same time. A third patient is scheduled to arrive about the middle of the cycle. If both patients arrive at the start of the cycle we let the Trainee start work on the New patient and the Attending handles the Return patient without the Trainee being involved. This was deemed acceptable because it was argued that most of the learning comes from the visits of New patients. If only one of the two patients arrives the standard process is used.

Process analysis tools produce some results about average cycle times in this setting, but since waiting times are serially correlated, we want a much clearer depiction of how each patient's waiting time was related to that for the following patients. Considering the problem using a queuing model is extremely difficult because the relevant distribution of activity times is state dependent, and the number of cycles is small. Consequently, steady state results are misleading. Studying this approach within a DES revealed that average makespan, waiting time, and cycle times are significantly reduced using our Cyclic approach and the Trainee is involved in a greater proportion of the cases scheduled.

5. Conclusion

While a great deal of time, effort, and money has been spent to improve health care processes, the problems involved have proven to be very difficult to solve. In this work we focused on a small but important sector of the problem space - that of appointment based clinics in Academic Medical Centers. One source of difficulty is that the medical field favors an experimental design based approach while many OR tools are more mathematical and abstract. Consequently, one of our core messages is that those working to improve these systems need to find ways to bridge this gap by combining techniques. When this is done progress can be made and the insights generated can be spread more broadly. Our use of DES builds on tools of process mapping that most managers are familiar with and facilitates virtual experiments that are easier to control and use to generate quantitative metrics amenable to the kinds of statistical tests that research physicians routinely apply.

However, we would be remiss if we failed to emphasize the fact that data driven approaches are rarely sufficient to bring about the desired change. Hospitals in AMC's are often highly politicized environments with a hierarchical culture. This fact can generate multiple roadblocks that no amount of ”number crunching” will ever overcome. One, not so subtle aspect of our method is that it typically involves imbedding ourselves in the process over some periods of time and interacting repeatedly with the parties involved. We have initiated many projects not mentioned above because they did not result in real action. Each and every project that has been successful involved many hours of working with faculty, physicians, staff, and technicians of various types to collect information and get new perspectives. We have seen dozens of researchers perform much more impressive data analysis on huge data sets using tools that were much more powerful than those employed in these examples, only to end up with wonderful analysis not linked to any implementation. When dealing with healthcare professionals we are often reminded of the old adage, ”No one cares how much you know. They want to know how much you care.” While, we believe that the methodology outlined in this chapter is useful, our experience strongly suggests that the secret ingredient to making these projects work is the attention paid to the physicians, faculty, and especially staff involved who ultimately make the system work.

References

Alexopoulos, C, D Goldman, J Fontanesi, D Kopald, Wilson JR. 2008. Modeling patient arrivals in community clinics. *Omega* **36** 33–43.

Bandura, A. 1969. *Principles of behavior modification.*. Holt, Rinehart, & Winston.

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

Beronio, K, R Po, L Skopec, S Glied. 2014. Affordable care act will expand mental health and substance use disorder benefits and parity protections for 62 million americans. *Mental Health* **2**.

Blanco White, MJ, MC Pike. 1964. Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care* 133–145.

Boex, JR, AA Boll, L Franzini, A Hogan, D Irby, PM Meservey, RM Rubin, SD Seifer, JJ Veloski. 2000. Measuring the costs of primary care education in the ambulatory setting. *Academic Medicine* **75**(5) 419–425.

Bowers, J, G Mould. 2005. Ambulatory care and orthopaedic capacity planning. *Health Care Management Science* **8**(1) 41–47.

Burke, EK, P De Causmaecker, GV Berghe, H Van Landeghem. 2004. The state of the art of nurse rostering. *Journal of Scheduling* **7**(6) 441–499.

Cayirli, T, E Veral, H Rosen. 2006. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science* **9**(1) 47–58.

Chambers, CG, M Dada, SM Elnahal, SA Terezakis, TL DeWeese, JM Herman, KA Williams. 2016. Changes to physician processing times in response to clinic congestion and patient punctuality: a retrospective study. *BMJ open* **6**(10) e011730.

Chao, X, L Liu, S Zheng. 2003. Resource allocation in multisite service systems with intersite customer flows. *Management Science* **49**(12) 1739–1752.

Chesney, AM. 1943. *The Johns Hopkins Hospital and John Hopkins University School of Medicine: a chronicle*. Johns Hopkins University Press.

Clymer, John R. 2009. *Simulation-based engineering of complex systems*, vol. 65. John Wiley & Sons.

Conley, WK, CG Chambers, SM Elnahal, A Choflet, KA Williams, TL DeWeese, JM Herman, M Dada. 2016. Using real-time location system technology to facillitate process analysis and time-driven activity based costing in a radiation oncology outpatient clinic.

Cutler, DM, K Ghosh. 2012. The potential for cost savings through bundled episode payments. *New England Journal of Medicine* **366**(12) 1075–77.

Fetter, RB, JD Thompson. 1966. Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research* **1**(1) 66.

Franzini, L, JM Berry. 1999. A cost-construction model to assess the total cost of an anesthesiology residency program. *The Journal of the American Society of Anesthesiologists* **90**(1) 257–268.

Glied, S, S Ma. 2015. *How will the Affordable Care Act affect the use of health care services?*. Commonwealth Fund.

Hamrock, E, J Parks, J Scheulen, FJ Bradbury. 2013. Discrete event simulation for healthcare organizations: a tool for decision making. *Journal of Healthcare Management* **58**(2) 110.

Hing, E, MJ Hall, JJ Ashman, J Xu. 2010. National hospital ambulatory medical care survey: 2007 outpatient department summary. *Natl Health Stat Report* **28** 1–32.

Hosek, JR, AR Palmer. 1983. Teaching and hospital costs: the case of radiology. *Journal of Health Economics* **2**(1) 29–46.

Huang, XM. 1994. Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Research* **7**(1) 2–8.

Hwang, CS, KA Wichterman, EJ Alfrey. 2010. The cost of resident education. *Journal of Surgical Research* **163**(1) 18–23.

Jun, JB, Sheldon H Jacobson, JR Swisher. 1999. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society* **50**(2) 109–123.

Kaplan, RS, SR Anderson. 2003. Time-driven activity-based costing. *SSRN 485443* .

Kaplan, RS, ME Porter. 2011. How to solve the cost crisis in health care. *Harvard Business Review* **89**(9) 46–52.

King, M, I Lapsley, F Mitchell, J Moyes. 1994. Costing needs and practices in a changing environment: the potential for abc in the nhs. *Financial Accountability & Management* **10**(2) 143–160.

Kolker, A. 2010. Queuing theory and discrete events simulation for health care. *Health Information Systems: Concepts, Methodologies, Tools, and Applications* **4** 1874–1915.

Lee, VJ, A Earnest, MI Chen, B Krishnan. 2005. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Research* **5**(1) 1.

Lin, CT, GA Albertson, LM Schilling, EM Cyran, SN Anderson, L Ware, RJ Anderson. 2001. Is patients' perception of time spent with the physician a determinant of ambulatory patient satisfaction? *Archives of Internal Medicine* **161**(11) 1437–1442.

Lorenzoni, L, A Belloni, F Sassi. 2014. Health-care expenditure and health policy in the usa versus other high-spending oecd countries. *The Lancet* **384**(9937) 83–92.

Mandelbaum, A, P Momcilovic, Y Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

McCarthy, K, HM McGee, CA O'Boyle. 2000. Outpatient clinic waiting times and non-attendance as indicators of quality. *Psychology, Health & Medicine* **5**(3) 287–293.

Meza, JP. 1998. Patient waiting times in a physician's office. *The American Journal of Managed Care* **4**(5) 703–712.

Moses, H, SO Thier, DHM Matheson. 2005. Why have academic medical centers survived? *Journal of the Americal Medical Association* **293**(12) 1495–1500.

Ortman, JM, VA Velkoff, H Hogan. 2014. An aging nation: the older population in the united states. *Washington, DC: US Census Bureau* 25–1140.

Perros, P, BM Frier. 1996. An audit of waiting times in the diabetic outpatient clinic: role of patients' punctuality and level of medical staffing. *Diabetic Medicine* **13**(7) 669–673.

Sainfort, F, J Blake, D Gupta, RL Rardin. 2005. Operations research for health care delivery systems. *WTEC Panel Report* .

Seals, B, CA Feddock, CH Griffith, JF Wilson, ML Jessup, SR Kesavalu. 2005. Does more time spent with the physician lessen parent clinic dissatisfaction due to long waiting times? *Journal of Investigative Medicine* **53**(1) S324–S324.

Sloan, FA, RD Feldman, AB Steinwald. 1983. Effects of teaching on hospital costs. *Journal of Health Economics* **2**(1) 1–28.

Tai, G, P Williams. 2012. Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine* **108**(2) 467–476.

Taylor, DH, DJ Whellan, FA Sloan. 1999. Effects of admission to a teaching hospital on the cost and quality of care for medicare beneficiaries. *New England Journal of Medicine* **340**(4) 293–299.

Thomas, S, R Glynne-Jones, I Chait. 1997. Is it worth the wait? a survey of patients' satisfaction with an oncology outpatient clinic. *European Journal of Cancer Care* **6**(1) 50–58.

Trebble, TM, J Hansi, T Hides, MA Smith, M Baker. 2010. Process mapping the patient journey through health care: an introduction. *British Medical Journal* **341**(7769) 394–397.

Trusko, BE, C Pexton, HJ Harrington, P Gupta. 2007. *Improving Healthcare Quality and Cost with Six Sigma*. Financial Times Press.

van Lent, Wineke AM, P VanBerkel, WH van Harten. 2012. A review on the relation between simulation and improvement in hospitals. *BMC Medical Informatics and Decision Making* **12**(1) 1.

Williams, JR, MC Matthews, M Hassan. 2007. Cost differences between academic and nonacademic hospitals: a case study of surgical procedures. *Hospital topics* **85**(1) 3–10.

Williams, KA, CG Chambers, M Dada, PJ Christo, D Hough, R Aron, JA Ulatowski. 2015. Applying jit principles to resident education to reduce patient delays: A pilot study in an academic medical center pain clinic. *Pain Medicine* **16**(2) 312–318.

Williams, KA, CG Chambers, M Dada, D Hough, R Aron, JA Ulatowski. 2012. Using process analysis to assess the impact of medical education on the delivery of pain servicesa natural experiment. *The Journal of the American Society of Anesthesiologists* **116**(4) 931–939.

Williams, KA, CG Chambers, M Dada, JC McLeod, JA Ulatowski. 2014. Patient punctuality and clinic performance: observations from an academic-based private practice pain centre: a prospective quality improvement study. *BMJ open* **4**(5) e004679.

Wilson, JC Tunnicliffe. 1981. Implementation of computer simulation projects in health care. *Journal of the Operational Research Society* **32**(9) 825–832.