A DATA-DRIVEN APPROACH FOR OPERATIONAL IMPROVEMENT IN EMERGENCY DEPARTMENTS

Wanyi Chen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2018

Approved by:

Nilay Argon

Serhan Ziya

Chuanshu Ji

Vidyadhar Kulkarni

Debbie Travers

ABSTRACT

WANYI CHEN: A Data-Driven Approach for Operational Improvement in Emergency Departments
(Under the direction of Nilay Argon and Serhan Ziya)

Emergency departments (EDs) in the US are experiencing significant stress from crowding, of which one of the main contributors is the lengthy boarding process, which is the process of to-be-admit patients waiting in the ED for the hospital to ready beds for them. We explored ways to reduce crowding by initiating hospital bed request (BeRT) early on for likely to-be-admit patients. In Chapter 2, we modeled the ED patient flow as a Markov decision process. With the objective of balancing the tradeoff between waiting cost and the cost of false early BeRTs, we found the optimal early BeRT policy to be of threshold type, where the threshold is a function of census and patients probability of admission. Chapter 3 built a fluid model, where patients flow into the ED (a fluid tank) as continuous fluid flowing at a time-dependent deterministic rate. To control the number of false early BeRTs, we imposed a constraint on the length of time for the early BeRT option. The optimal policy that minimizes the fluid level (congestion level) in the ED dictates that when ED is under heavy traffic regime, one should BeRT early as early, and as long, as allowed. In chapter 4, we looked at several early BeRT heuristics that are inspired by the theoretical optimal policies found previously. We tested and compared their performances in terms of length-of-stay and waiting time using a simulation model built for the UNC ED based on 2012 patient data. We observed that as the admission probability distributions of the patient population became less variable, the heuristics that take more information into account performed better. Lastly, we offered a different perspective on ED crowding in Chapter 5, where we explored the association between ED cencus and providers' triage and admission decisions. We found that the more crowded the ED was, the more conservative providers were, in that nurses tend to triage more patients as critical, and physicians tend to admit more patients into the hospital.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Emergency departments (EDs) are gateways to hospitals and play a critical role in the US healthcare system. The majority of them are experiencing significant operational stress caused by overcrowding. Failure to serve on time can put patients at risk for suboptimal care and potential health harm. Researchers have been seeking to identify primary causes of ED overcrowding and ways to reduce its adverse effects to the extent possible (Olshaker, 2009), (Hoot and Aronsky, 2008), (Welch et al., 2011).

One of the main contributors to ED crowding is the lengthy process of transferring an admitted patient from the ED to an inpatient department. It has been suggested that if the hospital admissions of ED patients can be predicted early during triage and communicated to different departments of a hospital, then necessary steps can be taken earlier to reduce transfer delays (Peck et al., 2012a). Typically, a patient visiting the ED is first triaged, i.e., examined to determine the complexity of the condition, where some basic information is collected and the patient is assigned a priority level. Patients then wait to be seen by an ED physician, who decides on the treatment and whether the patient needs to be admitted to the hospital or be discharged. Normally, request for a hospital bed and preparations to receive the patient are delayed until the admission decision by the doctor is ascertained. If hospital admission decisions can be predicted in advance (i.e., upon triage or soon after), then this information can be passed on to the target inpatient ward where staff can begin preparations early on and thereby reduce patient transfer delays and boarding. Our research is motivated by the aforementioned idea and seeks to find efficient ways to take advantage of early prediction of hospitalizations and thus enable an earlier start for the boarding process to shorten total length of stays at the ED. From this point on, we term this early request for a hospital bed as early BeRT (bed request).

With this motivation, we formulated a queueing model in Chapter 2 that approximates the patient flow in the ED in a stylized fashion. Each job (patient) that arrives to the queueing system (i.e., the ED) belongs to one of two types. Type 1 jobs need only a primary service given by a single server while type 2 jobs need an additional secondary service. The primary service corresponds to the lump sum service patients receive at the ED and the secondary service refers to the inpatient admission procedure (the secondary service time corresponds to boarding). The type of a patient determines whether he/she will be admitted to the hospital. Secondary service is conducted by servers that are always available when it is initiated. However, primary server cannot serve a new job until secondary service of a job is over. Jobs incur waiting costs and there is an option of starting primary and secondary services at the same time with an extra cost. The decision is whether or not to use that option for each job given the probability that the job is of type 1. We formulate this problem as a Markov decision process and prove that the optimal policy that minimizes the long-run average cost is of threshold-type.

In Chapter 3, we take an alternative approach and build a deterministic and continuous fluid model aiming to capture the general behavior of patient flow in the ED. As mentioned earlier, when implementing early bed requests in the ED one needs to carefully manage the tradeoff between the cost of overcrowding associated with holding hospital admitted patients in the ED and the cost of wasting hospital resources by making too many early bed requests based on false admission prediction. Of course, knowing the relative magnitude of these two counteracting costs can aid in our decision making yet it might be unrealistic to estimate the cost of a false early bed request. Hence, in this alternative formulation, we impose a constraint on the length of time during which one can make early bed requests and thus speed up service. To be more specific, we treat the patients arriving to an ED as fluid flowing into a tank. The fluid is pumped out of the tank at some deterministic outflow rate as patients receive service at the ED. There is an option to speed up the outflow rate, which corresponds to the option of starting preparing the hospital bed for patients early on based on their predicted probability of admission to hospital. Using this fluid model we identify the optimal period of time during each day to use that option given the aforementioned operational constraint on the total amount of time early bed requests can be made.

In Chapter 4, we evaluate several different heuristic policies that are motivated by our mathematical models discussed in Chapter 2 and Chapter 3. With the accessibility of a dataset that

consists of 12 months of all patient encounters at the UNC ED in 2012, we built a simulation model of the ED. We utilize this simulation model to evaluate the heuristics considered in terms of the improvement they bring in reducing patients' length-of-stay, waiting, and boarding times.

While the primary focus on ED crowding has been its influence on patient outcomes (e.g., patient mortality) there has been emerging research that pays attention to the operational responses to a congested ED such as the changes in rates of admission. Many ED patients fall into a gray area as to their needs for admission, and coming up with an appropriate discharge plan for these patients may require significant cost of staff time and physical resources. Consequently, ED physicians may choose to admit these patients as a safe alternative (Miller, 1960). (Gorski et al., 2017) explored the association between crowding, which is measured by the occupancy level, and likelihood of admission in a US ED and found there to be a positive correlation. Similarly, since the amount of time and resources needed to be invested in patients with different acuity levels vary dramatically, it is natural to conjecture that ED crowding might have an effect on the operational strategies being adopted at the forefront of the entire patient flow, i.e., triage area, as well. Following this stream of research ideas, in Chapter 5 we examine the impact of census, i.e., the occupancy level in the ED, on nurses' triage decisions, i.e., categorization of patients into acuity levels, and physicians' admission decisions. More specifically, we performed a retrospective analysis on all 2012 patients encounter data in the UNC ED. We employed a cumulative logistic regression model to assess the association between census and triage levels. To evaluate the relationship between census and admission decisions we used a logistic regression model.

CHAPTER 2

Dynamic Decision Making in a Queueing System with Secondary Service

2.1 Introduction

Emergency departments (EDs) are complex service systems, most of them deal with operational problems that are caused by high congestion, and seek novel ways of reducing patient waiting times and length-of-stay. In a typical emergency department, an arriving patient first goes through triage, which determines the patient's criticality and priority level. Then, once the patient is admitted to the emergency department, the patient is seen by a physician who makes a diagnosis. In some cases, it might be necessary for the physician to order some tests before finalizing his/her decision. After diagnosis, the patient is either discharged from the emergency department or is admitted to the hospital. While most emergency departments typically work in this fashion, some have been experimenting with or considering making some changes. One idea, which has been implemented in a number of departments, is to predict whether or not a patient would need a diagnostic test at the time of triage and possibly order tests at that time. It is possible that the patient might end up not needing a test but if the test is needed, having the test results available sooner will reduce the time the patient will keep the emergency department bed occupied. Another proposed idea, which faces some implementation challenges, is to predict whether or not a patient will eventually be admitted to the hospital at the time of triage and request a bed from the hospital at that time. This has the potential to significantly reduce the time an admitted patient occupies a bed since the hospital bed the patient will transfer to might already be available by the time the "admit" decision for the patient is given or at least would be available soon after. However, if the patient for whom an "admit" prediction is made ends up being discharged from the emergency department, that would mean that hospital resources were unnecessarily used to make the bed available, which would also turn into a problem between the emergency department and the hospital.

Mainly motivated by these novel practices, we consider a queueing system in which each arriving customer is either of type-1 or type-2. Customers of either type require primary service, which is provided by a single server (we will refer to this server as *the server* throughout the paper) while type-2 customers additionally require the secondary service. This secondary service is provided by another collection of servers, which are assumed to be infinitely many. All customers queue in front of the server. When the server picks up the next customer to serve, it cannot observe the type of the customer but can observe the probability that the customer is of type-2, i.e., that the customer will need the secondary service. Only after completion of the primary service, the server knows with complete certainty whether the customer will need the secondary service. However, there is nothing that prevents the primary and the secondary services to proceed simultaneously and thus the system controller can order the secondary service to start at the same time as the server starts the primary service even though the secondary service for that particular customer could be unnecessary.

If the controller does not initiate the secondary service together with the primary service, the server proceeds with the primary service and by the end of the service, it determines whether the customer needs the secondary service. If s/he does not (meaning a type-1 customer), the customer leaves right away and the server picks up the next customer. If the customer is of type-2, and thus needs the secondary service, the customer starts receiving the secondary service right away from the pool of infinitely many servers. However, the server cannot serve a new customer. It remains *blocked* until the secondary service of the customer is over. If the controller initiates the secondary service together with the primary service, the server again proceeds with the primary service and it determines whether the customer needs the secondary service. If she does not or if she does but the secondary service, which started earlier together with the primary service, is already over, then the customer leaves right away and the server picks up the next customer. Each customer incurs a waiting cost, which is linearly increasing with the time s/he spends waiting. Without loss of generality, there is no cost associated with the primary service but the secondary service has a cost and this cost is smaller when it is started after the primary service is over.

By choosing to start the secondary service together with the primary service, the controller hopes to shorten the time the server is occupied with the customer and makes the server available for other waiting customers more quickly. However, by doing that, the controller not only pays

more for the secondary service but is also taking a risk because the secondary service may in fact be completely unnecessary for that customer. Thus, the goal of the controller is to carefully manage this trade-off. More specifically, the objective of the controller is to minimize the long-run average cost the system incurs by identifying whether the secondary service should be initiated together with the primary service given the number of customers waiting and the probability that the customer who is about to start the primary service is of type-2.

The model broadly described above is stylized and is not meant to capture the motivating applications at a highly detailed, realistic level. Our goal in this paper is to provide some general insights and possibly pave the way for the analysis of more advanced formulations in the future. However, it might be necessary to provide some explanations for the reasons behind some of our modeling choices. It is likely clear to the reader that the customers in the model correspond to the patients arriving at the emergency department and the type of a customer determines whether the patient needs a diagnostic test in the first setting described above and whether the patient will eventually be admitted to the hospital in the second setting. Somewhat more difficult to see is what exactly the server corresponds to and what exactly it means to have an infinite collection of servers for the secondary service.

In EDs, patients are mainly served by the attending physician and other medical personnel including the residents and nurses but particularly for crowded EDs, one can also view each ED bed as a server as well. Until a bed is physically vacated and cleaned, a new patient cannot be admitted. In our model, the server should be seen as a *physician-bed* pair. It is the physician who is performing the primary service but the bed cannot be vacated until the whole service, which possibly includes the secondary service (corresponding to a diagnostic test or the hospital bed preparation time), is over and the server remains blocked. While this is the case in reality as well, the limitation of our formulation is that it assumes that there is a single physician-bed pair serving the patients, which is almost never the case. However, particularly for highly crowded EDs, which experience long waiting times, our analysis would provide useful insights as one can view waiting patients as being preassigned to each bed at the time of their arrivals and each bed having its own queue.

2.2 Literature Review

There are three streams of research relevant to the study undertaken in this paper, one concerned with the dynamic control of queues, and in particular the control imposed on the service process, one concerned with Business Process Management (BPM) and the evaluation of the changes in service structure along the dimension of cost, and the final one concerned with service outsourcing.

With regard to dynamic control of queues via varying service rates, one commonality found in most literature, a similarity to our paper, is that the decision making is typically centered around the tradeoff between two kinds of costs, namely, the cost of holding customers in the queue, which is nondecreasing with respect to the queue length, and the cost of applying faster service rates, which is nondecreasing with respect to the service rates applied. Additionally, among the papers dealing with the characterization of optimal policies most of them show that they have certain monotonicity structure in terms of the queue lengths.

Crabill and Thomas B. (Crabill, 1972) derive the form of a minimal cost rate stationary operating policy for an M/M/1 queueing system with K possible service rates, where the optimal service rate is shown to be nondecreasing in the state of the system. The two costs of the system are a general cost rate dependent on the state of the system and a cost rate associated with each of the possible service rates. Weber et al. (Weber and Stidham, 1987) prove a monotonicity result for the problem of optimal service rate control in certain queueing networks. As an illustrative example, they show that for a number of G/M/1 queues arranged in a cycle with some number of customers moving around the cycle, the policy that minimizes the expected total discounted cost has a monotone structure: namely, that by moving one customer from the current queue to the following queue, the optimal service rate in the current queue is not increased and the optimal service rates elsewhere are not decreased. In their setting the cost is charged for holding customers in the queues and for each unit of time the service rate is in effect in the queues. Later, Stidham Jr. et al. (Stidham Jr and Weber, 1989) present a unified, simple method for proving that an optimal policy is monotonic in the number of customers in the system. Compared to (Crabill, 1972) they consider weaker and more general conditions and both exponential and nonexponential models. George et al. (George and Harrison, 2001) consider the problem of minimizing average cost per time unit over an infinite horizon for a single-server queue where the queue length evolves as a birth-and-death

process with constant arrival rate and state-dependent service rates that can be chosen from a fixed subset. There is a nondecreasing cost-of-effort function on the subset of values that service rates can be chosen from and holding costs are continuously incurred as a nondecreasing function of the queue length. They find that the optimal service rates are nondecreasing as a function of queue length. They also present a method for computing the minimum achievable average cost.

As to the redesign of the underlying mechanics of business processes, Buzacott and John A. (Buzacott, 1996) gives a comprehensive review of different kinds of system structure reengineering and explores conditions under which such changes are beneficial. In general, a high degree of variability in task times seems to be necessary. In his paper, the scenario where several tasks are combined into one is a similar version of our problem in the sense that for both the purpose is to reduce or eliminate subdivision of the overall processing requirements into individual tasks, each performed by a different facility, person or machine. To be more specific, our model assumes there is the option of starting stage-2 service together with stage-1 service with a resulting service time of the maximum of the two. This is in parallel with the use of case teams, suggested by Hammer and Champy and summarized in the paper, where in our case the team corresponds to the team formed by stage-1 and stage-2 servers. Similar to their setting, each job is not complete until both tasks are complete and the next job cannot begin until the previous job is complete. The difference lies in the fact that with no collaboration, which in our case corresponds to starting stage-2 service after stage-1 service is complete, the next job cannot enter stage-1 service until both tasks on the previous job are complete. To our knowledge, most BPM research is primarily concentrated upon identifying the objective and developing heuristic policies for business processes through a qualitative perspective. Interested readers can refer to (Van Der Aalst, 2013) for a comprehensive review of the state-of-the-art in BPM research. On the other hand, this paper starts with a redesign of the service flow, namely, to conduct second stage service together with first stage service at the same time, and focuses on finding the dynamically optimal way to apply this control.

Outsourcing refers to the act to procure (as some goods or services needed by a business or organization) under contract with an outside supplier. In this paper, due to the availability of outside servers that can perform stage-2 service one has the option of starting stage-2 service together with stage-1 service, and the motive for doing this is that arrivals might need a second stage of service on top of one, which will lengthen the total service time, add to the congestion, and thus

increase the holding cost. Hence the act of starting stage-2 service can be considered as outsourcing, and the decision-making is centered around balancing in-house holding cost and outsourcing cost (starting ahead is more expensive than starting on-demand). In relation to the literature on service outsourcing, our paper is most similar to (Koçağa et al., 2015). Most papers in this literature study contracting issues in the context of call center outsourcing, where a firm (which we will call the user) that sends some or all of its calls to an outside server (which we willl call the vendor) must determine appropriate terms for the contract to induce the vendor to make system-optimal decisions and the vendor must make decisions about staffing level and effort level. Unlike these papers, (Koçağa et al., 2015) focus on real-time routing decisions instead. They are faced with the issue of under/over-staffing in call centers when arrival rates are uncertain. To mitigate this issue, they find a joint policy for staffing and real-time call co-sourcing, i.e., by sometimes outsourcing calls, that minimizes long run average cost when there is staffing cost and costs associated with abandoments and outsourcing. They formulate a Markov decision process and propose a policy that uses a square-root safety staffing rule, and outsources calls in accordance with a threshold rule that is dependent on the queue length. They show that this policy is asymptotically optimal. Both the optimality of a threshold-type policy and the cost structure that influences dynamic decision making is very similar to our work presented here.

2.3 Model Description

Customers arrive at a queueing system according to a Poisson process with rate λ. Each customer is one of two types. Type-1 customers need only *primary* service while type-2 customers need both *primary* and *secondary* services. Arriving customers line up for primary service, which is provided by a single server. There is no queue for the secondary service. It is provided by one of infinitely many secondary servers and can be performed either immediately after primary service or simultaneously with primary service depending on the system controller's decision. For any given customer, if secondary service follows the completion of primary service or if they are started simultaneously but primary service finishes first and it is revealed that the customer does indeed need the secondary service (i.e., a type-2 customer) the single server, which performs primary services remains blocked and cannot serve a new customer until the secondary service of the customer is

complete. (In the rest of the paper, unless otherwise specified, "the server" will always refer to the server performing the primary service.) However, if primary service and secondary service are started simultaneously and primary service finishes first but it is revealed that secondary service is in fact not needed, the customer leaves right away and the server becomes immediately available for the next customer.

The system controller cannot observe the type of the customers before they go through primary service but it can observe the probability of any given customer being of type-2. The type of the customer is revealed with certainty only after the completion of primary service. Let Z_k denote the random variable representing the probability that the kth customer to arrive to the system is of type-2. We assume that $\{Z_k\}_{k=1}^{\infty}$ is a sequence of independent and identically distributed (iid) random variables with the common discrete probability distribution specified as $P\{Z_k = \alpha_i\} = q_i\}$ for $\alpha_i \in \Omega$ and $k \in \{1, 2, \dots\}$, where $\Omega = \{\alpha_1, \alpha_2, \dots, \}$ is the set of possible values Z_k can take. Without loss of generality, we assume that α_i is increasing in i. We also let $\alpha = \sum_{i=1}^{\infty} q_i \alpha_i$ so that α represents the probability that a randomly chosen customer is of type-2.

Let X_{1k} denote the primary service time for the kth customer and X_{2k} denote the secondary service time for the kth customer to receive this service. We assume that $\{X_{1k}\}_{k=1}^{\infty}$ is a sequence of iid random variables with exponential distribution with rate γ_1 and $\{X_{2k}\}_{k=1}^{\infty}$ is a sequence of iid random variables with exponential distribution with rate γ_2. To clearly describe the service time an arriving customer experiences, let X_1 and X_2 denote generic random variables respectively representing the time it takes for primary service and secondary service. Consider a customer whose probability of being type-2 is z. The system controller can either choose to serve this customer by performing primary and secondary services together starting them simultaneously (called "parallel service") or can choose to have the server perform primary service first, and only then initiate secondary service if primary service reveals that secondary service is needed (called "service in sequence"). Note that the customer may eventually turn out to be type-1 and not need secondary service but the system controller might still choose the parallel service option in hopes of making server-1 available for other customers more quickly considering the possibility of the customer being type-2. Let S_p^z denote the total time the customer keeps the server busy under parallel service and

let S_s^z denote the same time if the services are performed in stage. Then,

$$S_p^z = \begin{cases} \max(X_1, X_2) & \text{w.p. } z \\ X_1 & \text{w.p.} 1 - z \end{cases},$$

and

$$S_s^z = \begin{cases} X_1 + X_2 & \text{w.p. } z \\ X_1 & \text{w.p. } 1 - z \end{cases}.$$

Thus, the benefit of choosing the parallel service option is that with probability z, the service time of the customer shortens to $\max(X_1, X_2)$ from $X_1 + X_2$. As we explain, next, however, there are costs associated with taking different actions and therefore choosing this option for all the customers may not be desirable.

Specifically, we assume that the system incurs a holding cost of C_w for each waiting customer per unit of time. The cost of performing secondary service after the completion of primary service is denoted by C_s and the cost of performing secondary service in parallel with primary service is denoted by C_p. Note that any cost of primary service is irrelevant and is thus ignored because all customers have to go through primary service. We assume throughout the paper that $\alpha_i C_s \leq C_p$ for all i, which implies that for any single customer in isolation the cost of performing parallel service is larger than the expected cost of performing service in sequence. An obvious sufficient condition for this assumption to hold is that $C_s \leq C_p$, i.e., service in sequence does not cost more than service in parallel, which is likely to hold in our motivating applications where the cost under either service option would likely be about the same. The objective of the service controller is to minimize the long-run average cost for this system by determining when to choose parallel service and when to choose service in sequence depending on the system state.

We model this problem as a Markov decision process (MDP). The state space $\mathbb{X}$ can be described as $\mathbb{X} = \{0\} \cup \{(m, n) \mid m \in \{\alpha_i\}_{i=1}^{\infty} \cup \{2, 3\},\ n \in \mathbb{Z}^+\}$ where state (0) is the state where the system is empty, states (m, n) are the states in which there are n customers in the system including the customer with the server with $m = \alpha_i$ corresponding to the state in which the server is performing a primary service on a customer with probability α_i of being type-2, $m = 2$ corresponding to the state in which the server is performing a primary service and secondary service has been completed,

and $m = 3$ corresponding to the state in which the server has already completed primary service and the customer is now going through secondary service. We restrict ourselves to the policy set Π, where any $\pi \in \Pi$ is a stationary, non-idling, state-dependent policy, and is a mapping from the system state $\mathbb{X}$ to the action space $\mathcal{A} = \{0, 1\}$ where 0 corresponds to the decision of starting a "service in sequence", i.e., not initiating a secondary service together with primary service and 1 corresponds to the decision of starting a parallel service with the restriction that no action is available in state (0) and action 1 is only available in states x where $x = (\alpha_i, n \geq 1)$, for some i, i.e. when there is at least one customer and either a primary or a secondary service has not already been completed for the customer because of a parallel service decision made earlier. Note that the policies we consider here can be seen as preemptive in the sense that the system controller can switch from "parallel service" to "service in sequence" at a decision epoch, which can correspond to either an arrival time or a service completion time, as long as neither primary nor secondary service is complete for the customer or from "service in sequence" to "parallel service" as long as the primary service of the customer is still in progress.

Using uniformization, the continuous-time MDP formulation can equivalently be written as a discrete-time MDP. Let $\beta = \lambda + \gamma_1 + \gamma_2$ denote the uniformization constant. We set $\beta = 1$ without loss of generality. For any $x \in \mathbb{X}$, $h(x)$ denotes the relative value or bias for state x. For expositional convenience below, we further define $h(\alpha_j, 0) = h(0)$ for $j = 1, 2, \ldots$ as the relative value function although $(\alpha_j, 0)$ is not an element of the state space $\mathbb{X}$. Finally, let g denote the long-run average cost under an optimal policy. Then, the optimality equations can be written as follows:

$$h(0) = \lambda \sum_j q_j h(\alpha_j, 1) + (\gamma_1 + \gamma_2) h(0). \tag{2.1}$$

For all $n \geq 1$, and $\alpha_i \in \Omega$,

$$\begin{aligned} h(\alpha_i, n) = nC_w + \lambda h(\alpha_i, n+1) + (1 - \alpha_i)\gamma_1 \sum_j q_j h(\alpha_j, n-1) \\ + \alpha_i \gamma_1 h(3, n) + \gamma_2 \min\{h(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, h(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\}, \end{aligned} \tag{2.2}$$

where $h(\alpha_j, 0) = h(0) = \sum_j q_j h(\alpha_j, 0)$, for all j. For all $n \geq 1$,

$$h(2,n) = nC_w + \lambda h(2, n+1) + \gamma_1 \sum_j q_j h(\alpha_j, n-1) + \gamma_2 h(2,n), \tag{2.3}$$

$$h(3,n) = nC_w + \lambda h(3, n+1) + \gamma_2 \sum_j q_j h(\alpha_j, n-1) + \gamma_1 h(3,n). \tag{2.4}$$

We know that if there is a solution to the optimality equations above, then there exists a stationary, deterministic policy, $\pi^* \in \Pi$ under which the long-run average cost is $g = g^*$ and the policy is described by the action that minimizes the right hand side of the optimality equation for each state $x \in \mathbb{X}$.

2.4 Existence of a Stationary Optimal Policy

In this section, we show that under a particular condition on the arrival and service rates, the solution to the optimality equations exist and thus there exists an optimal stationary policy. The condition we need is that $\lambda\left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2}\right) < 1$. Recall that α is the probability that a randomly chosen customer is of type-2 and thus the term in the parentheses, $\left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2}\right)$ is the total expected time the server will be occupied with a random customer if a decision is made to perform the two stages of service in sequence and the condition is basically the stability condition for the queueing system if all customers are served in a service-in-sequence fashion. It is important to note that this a sufficient condition and that there could be solutions to the optimality equations if it does not hold.

Theorem 1. *Suppose $\lambda\left(\frac{1}{\gamma_1} + \frac{\mathrm{E}[\alpha]}{\gamma_2}\right) < 1$, then there exists a finite constant J and a finite function h that satisfy the ACOE (average cost optimality equalities):*

$$J + h(i) = \min_a \left\{ C(i,a) + \sum_j P_{ij}(a) h(j) \right\}, \; i \in \mathcal{S}.$$

Let f be a stationary policy realizing the equality in the ACOE. Then f is average cost optimal with average cost J.

Proof. According to Theorem 7.2.3 in Sennott (Sennott, 2009) the (SEN) Assumptions ensures the existence of a finite constant J and a finite function h that satisfy the ACOI (average cost optimal

inequalities)

$$J + h(i) \geq \min_a \left\{ C(i,a) + \sum_j P_{ij}(a)h(j) \right\}, \; i \in \mathcal{S}.$$

Also, there exists an average cost optimal policy f that achieves the minimum in the ACOI. According to Theorem 7.5.6 (Sennott, 2009), the (BOR) Assumptions ensure that the (SEN)s Assumptions hold and that the ACOE is valid. Hence we only need to show that (BOR) Assumptions hold under the condition that $\lambda\left(\frac{1}{\gamma_1} + \frac{\mathrm{E}[\alpha]}{\gamma_2}\right) < 1$.

We consider the stationary policy d that always chooses the parallel service option. Then under the assumption that $\lambda\left(\frac{1}{\gamma_1} + \frac{\mathrm{E}[\alpha]}{\gamma_2}\right) < 1$, the Markov chain induced by d is a M/G/1 with service time distribution:

$$S = \begin{cases} X_1 & \text{w.p. } 1 - \mathrm{E}[\alpha] \\ \max(X_1, X_2) & \text{w.p. } \mathrm{E}[\alpha] \end{cases}.$$

Thus

$$\begin{aligned} \mathrm{E}[S] &= (1 - \mathrm{E}[\alpha]) \cdot \mathrm{E}[X_1] + \mathrm{E}[\alpha] \cdot \mathrm{E}[\max(X_1, X_2)] \\ &\leq (1 - \mathrm{E}[\alpha]) \cdot \mathrm{E}[X_1] + \mathrm{E}[\alpha] \cdot \mathrm{E}[X_1 + X_2] \\ &= (1 - \mathrm{E}[\alpha]) \cdot \frac{1}{\gamma_1} + \mathrm{E}[\alpha] \cdot \left(\frac{1}{\gamma_1} + \frac{1}{\gamma_2}\right) \\ &= \frac{1}{\gamma_1} + \frac{\mathrm{E}[\alpha]}{\gamma_2}. \end{aligned}$$

The utilization of the system is

$$\rho = \lambda \mathrm{E}[S] \leq \lambda \left(\frac{1}{\gamma_1} + \frac{\mathrm{E}[\alpha]}{\gamma_2}\right) < 1.$$

Hence the Markov chain induces by d is stable and induces a positive recurrent class $\mathcal{R}_d$. It is easy to see that $\mathcal{R}_d = \mathcal{S} = \{(0)\} \cup \{(\alpha_i, n), 1 \leq i \leq K, n \geq 1\} \cup \{(s, n), s = 2 \text{ or } 3\}$. Suppose we choose a distinguished state $z = 0$. By Definition 7.5.1 (Sennott, 2009) and Definition C.2.5 (Sennott, 2009), d is a z standard policy because the MC under d is positive recurrent, and hence the expected first passage time and associated total expected cost from one state to another are both finite. Hence (BOR1) holds. In fact, d is a z standard policy for any $z \in \mathcal{R}_d = \mathcal{S}$.

Next, since the MC under d is positive recurrent, the long run average cost under d, denoted by J_d, is finite. Choose $\varepsilon = 1$. Define $D = \{s \mid C(s;a) \leq J_d + 1 \text{ for some } a\}$ as in (BOR2), where

$$C(0) = 0;\ C(2,n) = C(3,n) = nC_w,\ \forall n \geq 1,$$

and for $1 \leq i \leq K$ and $n \geq 1$,

$$C(\alpha_i, n; 0) = nC_w + \alpha_i \gamma_1 C_s,\ C(\alpha_i, n; 1) = nC_w + \gamma_2 C_p.$$

then $D = \{0\} \cup A \cup B$, where

$$A = \{(\alpha_i, n) \mid 1 \leq i \leq K \text{ and } 1 \leq n \leq \left\lfloor \frac{1}{C_w}\left(J_d + 1 - \min\{\alpha_i \gamma_1 C_s, \gamma_2 C_p\}\right) \right\rfloor.$$

$$B = \{(2,n) \text{ and } (3,n) \mid 1 \leq n \leq \left\lfloor \frac{J_d + 1}{C_w} \right\rfloor\}.$$

It is easy to see that D is a finite set since J_d is finite, meaning that (BOR2) holds. Finally, (BOR3) holds because $D - \mathcal{R}_d = \emptyset$. □

2.5 Structure of the Optimal Policy

This section is devoted to proving that if $\lambda\left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2}\right) < 1$, i.e., under the condition with which we can ensure the existence of an optimal policy, the optimal policy has a threshold structure. More specifically, the optimal policy is such that for any given value of α_i, the probability for the customer to be of type-2, the parallel service option is chosen if and only if the number of customers in the system is above a particular threshold value. We start with the statement of the theorem.

Theorem 2. *Suppose that $\lambda\left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2}\right) < 1$. Then, the optimal policy, which minimizes the long-run average cost, is of threshold type. More specifically, there exists an integer $N(\alpha_i)$ such that if the system is in state (α_i, n), i.e., there are n customers in the system and the customer who is already receiving primary service or is about to start receiving service has a probability α_i of being type-2, then the optimal action is to perform parallel service if and only if $n \geq N(\alpha_i)$. Furthermore,*

$$N(\alpha_i) := \inf\{n : h(\alpha_i, n) - h(2,n) > \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p - \frac{\alpha_i \gamma_1}{\gamma_2} C_s\}.$$

From (2.2), one can see that the optimal action in state (α_i, n) is to perform parallel service if and only if $h(\alpha_i, n) - h(2, n) > \frac{\gamma_1+\gamma_2}{\gamma_2} C_p - \frac{\alpha_i \gamma_1}{\gamma_2} C_s$. Therefore, if the right hand side of this inequality, $h(\alpha_i, n) - h(2, n)$, is non-decreasing in n, Theorem 2 immediately follows. In the rest of this section, we prove that is indeed the case.

First, we introduce the finite-horizon version of the uniformized, discrete-time version of our problem described in Section 2.3. Let $V_m^\pi(x)$ denote the total expected cost under policy π over a period of m stages starting from state x. The optimal expected m-stage cost then can be expressed as

$$V_m(x) = \inf_{\pi \in \Pi} V_m^\pi(x),$$

and satisfies the following finite horizon optimality equations: For $m \geq 1$,

$$V_m(0) = \lambda \sum_j q_j V_{m-1}(\alpha_j, 1) + (\gamma_1 + \gamma_2) V_{m-1}(0). \tag{2.5}$$

For all $m \geq 1$, $n \geq 1$, and $\alpha_i \in \Omega$,

$$\begin{aligned} V_m(\alpha_i, n) = {} & nC_w + \lambda V_{m-1}(\alpha_i, n+1) + (1-\alpha_i)\gamma_1 \sum_j q_j V_{m-1}(\alpha_j, n-1) \\ & + \alpha_i \gamma_1 V_{m-1}(3, n) + \gamma_2 \min\{V_{m-1}(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\}, \end{aligned} \tag{2.6}$$

where $V_m(\alpha_j, 0) = V_m(0) = \sum_j q_j V_m(\alpha_j, 0)$, for all j. For all $m \geq 1$ and $n \geq 1$,

$$V_m(2, n) = nC_w + \lambda V_{m-1}(2, n+1) + \gamma_1 \sum_j q_j V_{m-1}(\alpha_j, n-1) + \gamma_2 V_{m-1}(2, n), \tag{2.7}$$

$$V_m(3, n) = nC_w + \lambda V_{m-1}(3, n+1) + \gamma_2 \sum_j q_j V_{m-1}(\alpha_j, n-1) + \gamma_1 V_{m-1}(3, n). \tag{2.8}$$

Next, we show that the optimality operator preserves certain conditions as stated in the following lemma. It is important to note that while only some of the conditions stated in the lemma will be key to establishing the threshold result, the proof of those essential conditions requires showing all of them together.

Lemma 1. *Suppose for any $m \geq 1$ we have that*

1) $V_m(2,n) - \sum_j q_j V_m(\alpha_j, n-1)$ *is a non-negative non-decreasing function of n for all $n \geq 1$.*

2) $V_m(\alpha_i, n) - \sum_j q_j V_m(\alpha_j, n-1)$ *is a non-decreasing function of n for all i and $n \geq 1$ and* $V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0) \geq \alpha_i C_s$.

3) $\min\{V_m(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p\} - V_m(2,n)$ *is a non-decreasing function of n for all i and $n \geq 1$.*

Then we have

Condition 1. $\min\{V_m(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p\} - \sum_j q_j V_m(\alpha_i, n-1)$ *is a non-decreasing function of n for all i and $n \geq 1$, and* $\min\{V_m(\alpha_i, 1) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_m(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p\} - \sum_j q_j V_m(\alpha_i, 0) \geq \frac{\alpha_i(\gamma_1+\gamma_2)}{\gamma_2} C_s$ *for all i.*

The proof of this lemma is provided in the appendix.

Lemma 2. *Suppose for any $m \geq 1$ we have that $V_m(\alpha_i, n) - V_m(2,n)$ is a non-decreasing function of n for all i and $n \geq 1$. Then we have*

Condition 2. $V_m(\alpha_i, n) - \min\{V_m(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_r, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_e\}$ *is a non-decreasing function of n for all i and $n \geq 1$, and* $V_m(\alpha_i, 1) - \min\{V_m(\alpha_i, 1) + \frac{\alpha_i \gamma_1}{\gamma_2} C_r, V_m(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_e\} \geq -\frac{\alpha_i \gamma_1}{\gamma_2} C_r$ *for all i.*

Condition 3. $\min\{V_m(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_r, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_e\} - V_m(2,n)$ *is a non-decreasing function of n for all i and $n \geq 1$.*

The proof of this lemma is provided in the appendix.

Lemma 3. *Let $\alpha_i C_s \leq C_p$, for all $\alpha_i \in \Omega$ and suppose that the following six conditions all hold for $0 \leq k \leq m-1$ where $m \geq 1$:*

Condition 4. $V_k(\alpha_i, n) - V_k(2,n)$ *is a non-decreasing function of n for all i and $n \geq 1$.*

Condition 5. $V_k(3,n) - \sum_j q_j V_k(\alpha_j, n-1)$ *is a non-negative non-decreasing function of n for all $n \geq 1$.*

Condition 6. $V_k(\alpha_i, n) - (1-\alpha_i) \sum_j q_j V_k(\alpha_j, n-1) - \alpha_i V_k(3,n)$ *is a non-decreasing function of n for all i and $n \geq 1$, and* $V_k(\alpha_i, 1) - (1-\alpha_i) \sum_j q_j V_k(\alpha_j, 0) - \alpha_i V_k(3,1) \geq \alpha_i C_s$.

Condition 7. $V_k(2,n) - \sum_j q_j V_k(\alpha_j, n-1)$ *is a non-negative non-decreasing function of n for all $n \geq 1$.*

Condition 8. $V_k(\alpha_i, n) - \sum_j q_j V_k(\alpha_j, n-1)$ *is a non-decreasing function of* n *for all* i *and* $n \geq 1$ *and* $V_k(\alpha_i, 1) - \sum_j q_j V_k(\alpha_j, 0) \geq \alpha_i C_s$.

Condition 9. $V_k(\alpha_i, n)$ *is a non-decreasing function of* i *for all* $n \geq 1$.

Then Condition 4 through 9 also hold for $k = m$*, i.e., Condition 4 through 9 are preserverd under the optimality equations.*

The proof of this lemma is provided in the appendix. Now, we choose the terminating costs so that $V_0(\alpha_i, n) = nC_s$ for $n \geq 0$ and $\alpha_i \in \Omega$, $V_0(2, n) = V_0(3, n) = (n-1)C_s$ for $n \geq 1$. One can then easily check that all the conditions of Lemma 3 hold for $m = 1$. Then, repeated use of Lemma 3 implies that all the conditions of the lemma hold for any integer $m \geq 1$. We also know from Theorem 1 that there exists an optimal policy for the long-run average cost problem with bias function $h(\cdot)$ satisfying the ACOEs (2.1) through (2.4). Thus, we must have

$$h(\alpha_i, n) - h(2, n) = \lim_{m \to \infty} [V_m(\alpha_i, n) - V_m(2, n)]$$

for $\alpha_i \in \Omega$ and $n \geq 1$. Then, because we know that all the conditions of Lemma 3 holds for any m and in particular Condition 1, i.e., $V_m(\alpha_i, n) - V_m(2, n)$ is a non-decreasing function of n, we can conclude that $h(\alpha_i, n) - h(2, n)$ is also non-decreasing in n for $n \geq 1$ and $\alpha_i \in \Omega$. This completes the proof of Theorem 2.

2.6 Monotonicity of the Optimal Threshold

In Section 2.5, we showed that the optimal policy is such that for a customer who is about to receive service or already receiving a preliminary service the parallel service option is chosen if and only if the number of customers in the system n at the decision time is greater than some threshold value $N(\alpha_i)$ where α_i is the probability that the customer is of type-2. Recall that type-2 customers are those who must have the secondary service and thus the incentive for choosing the parallel service option is stronger when this probability is larger. In other words, it would be reasonable to expect the minimum number of customers in the system that would justify parallel service to be smaller when this probability is larger. In this section, we prove that that is indeed the case.

Theorem 3. *The optimal threshold $N(\alpha_i)$ is a non-increasing function of α_i.*

Proof. The proof follows along the lines of the proof of Theorem 2. First, we choose the terminating costs so that $V_0(\alpha_i, n) = nC_s$ for $n \geq 0$ and $\alpha_i \in \Omega$, $V_0(2, n) = V_0(3, n) = (n-1)C_s$ for $n \geq 1$. One can then easily check that the conditions of Lemma 3 hold for $m = 1$. Then, repeated use of Lemma 3 implies that all the conditions of the lemma hold for any integer $m \geq 1$. We also know from Theorem 1 that there exists an optimal policy for the long-run average cost problem with bias function $h(\cdot)$ satisfying the ACOEs (2.1) through (2.4). Thus, we must have

$$h(\alpha_i, n) - h(2, n) = \lim_{m \to \infty} [V_m(\alpha_i, n) - V_m(2, n)]$$

for $\alpha_i \in \Omega$ and $n \geq 1$. Then, because we know that the conditions of Lemma 3 hold for any m and in particular Condition 4, i.e., $V_m(\alpha_i, n) - V_m(2, n)$ is a non-decreasing function of α_i, we can conclude that $h(\alpha_i, n) - h(2, n)$ is also non-decreasing in α_i. This completes the proof of the theorem. $\square$

CHAPTER 3

Optimal Timing for Early Bed Request for Admitted Patients in an Emergency Department

3.1 Introduction

In this chapter, we are interested in implementing as well as evaluating the efficacy of early bed requests in EDs to reduce patient sojourn times. The idea is to predict whether or not a patient will eventually be admitted to the hospital at his/her time of arrival, rather than later when the ED service is completed for the patient, and request a bed from the hospital at that time. From now on, we term this operational strategy as early bed request, or BeRT. BeRT could possibly reduce the time an admitted patient occupies a bed in the ED, because by the time the service at the ED is completed for the patient, the bed at the hospital might have already been prepared for him, or at least will be soon after, since it was called ahead earlier on at the time of arrival as opposed to at the end of the ED service according to the usual practice. However, if the patient for whom a BeRT is made turns out to be a discharged patient, i.e., the prediction is a false positive, that would mean that the hospital resources were unnecessarily employed to make the BeRT, which could turn into a problem between the hospital and the emergency department. To incorporate this fact into our model, we assume in this chapter that there is a limit on the maximum number of BeRTs per day, and this limit is derived based on discussions with the ED management about the hospital management's tolerance for the number of false BeRTs per day, and the sensitivity of the admission prediction.

Most emergency departments lack a valid and well-performing admission prediction tool. As part of a UNC Healthcare Innovations project, we have developed a logistic regression model that can predict each individual patient's probability of being admitted based on his/her demographic characteristics and clinical information that are available during triage (Travers et al., 2017). We term this logistic regression model as APT (Admission Prediction Tool). Although APT is not

at the center of our discussion in this chapter, it is important to know the performance metrics derived from it, and we will discuss it later to aid in our primary discussion.

To implement BeRT, we first need to come up with a tool that guides us with when to initiate the bed preparation process at the hospital each day and for which patients in the ED. Because of the cost of potentially wasting hospital resources due to incorrect prediction of admission, there has to be a limit on the number of times one can use the option of BeRT on ED patients. To find a decision rule that dictates when to implement BeRT during the day, we propose a mathematical fluid model to approximate the behavior of the patient flow and service process in an ED. In this model, we regard patients arriving to an ED as fluid flowing into a tank with unlimited capacity according to a deterministic inflow (arrival) rate. Upon arrival, the tank (ED) will immediately start emptying the fluid (serving the patient) at a deterministic rate. There are two options for outflow (service) rate. The minimum outflow rate corresponds to the overall service rate, which is the inverse of mean sojourn time at the ED, under normal operating conditions, i.e., with no BeRT applied to any patient. The maximum outflow rate corresponds to the overall service rate at the ED when BeRT is applied to all patients for whom the predicted admission probabilities are above a certain threshold. Our problem is to determine the optimal time at which one should start applying the maximum service rate to the system and the length during which one should keep applying the maximum service rate so as to minimize the time averaged fluid level in the system subject to a constraint on the maximum length during which the maximum service rate can be applied. Although the model broadly described here is stylized and is not meant to capture the ED operations at a highly detailed and realistic level, the main purpose here is to shed light on how to apply the BeRT strategy at an actual emergency department optimally depending on its operating conditions such as daily arrival volumes and its service capacities, as well as to evaluate the benefit that this novel strategy can bring to the ED in reducing patient sojourn times.

To evaluate the optimized decision rule suggested by the mathematical fluid model, we will test the optimal policy on a simulation model developed for the ED at the UNC Medical Center in Chapel Hill. The preliminary version of the simulation model is developed by previous graduate students based on a smaller dataset. As part of my dissertation, this simulation model is currently being refined and validated based on the 2012 hospital data. We are not going to explain details about how different components of the simulation model was built and the rationale behind it.

Instead, our focus is on testing the optimal policy found by the mathematical fluid model on the simulation model, evaluating the system performance in terms of the sojourn times, and getting insights into how to implement this policy in the real ED.

3.2 Literature Review

Our research relies on a logistic regression model that we developed as part of an UNC Healthcare Innovations Grant to predict the probability of admissions for ED patients (Travers et al., 2017). The recent published literature offers a handful of classification tools to predict admissions of ED patients. Peck et al. (Peck et al., 2012b) evaluated three models (expert opinion, naive Bayes and a generalized linear regression model) that predict the number of ED patients that will be admitted and introduced a methodology for implementing these models in a hospital setting. Barack-Corren et al. (Barak-Corren et al., 2017) developed a logistic regression model to predict patient disposition (hospitalization vs. discharge) at three progressive time points throughout the ED visit using clinical, operational and demographic data retrospectively collected in an Israeli hospital. LaMantia et al. (LaMantia et al., 2010) focuses on elderly patients, and derived and validated a triage-based model that predicts hospital admission of elderly patients and probabilities of them returning to the ED. Despite different modeling techniques, whether it is statistics based or it relies on solely the judgement of experienced nurse and/or physicians, the similarity we found is that most of the aforementioned work favor simple probabilistic models using only a minimal number of predictors available at triage and renders reasonable accuracy.

More recent studies seek means to actually employ the predicted information to reduce the boarding time of ED patients. Peck et al. (Peck et al., 2012b) recommends starting bed coordination early on while patients are still receiving the ED treatment to reduce the boarding delays. While they recognize the potential benefit of introducing admission prediction into the ED setting on reducing boarding time they do not have a model that is used to optimize this decision process as their focus is on developing a good model for prediction. Qiu et al. (Qiu et al., 2015) proposed a cost sensitive bed reservation policy that recommends optimal bed reservation times for patients. Their policy is cost sensitive in that it accounts for costs associated with admission prediction misclassification as well as costs associated with incorrectly selecting the reservation time. However,

unlike our work, which considers the ED as a queueing system and evaluates the total "cost" incurred for the system through all customers (patients) waiting, they do not have a queueing model and only assess costs at an individual patient level.

Our work is novel in the sense that we study the ED as a queueing system and use a fluid model to approximate the patient flow in the ED in a continuous and deterministic manner to optimize the early bed request decision. And not only are we able to draw conclusions about the optimal timing and length of time to take advantage of admission prediction, which speeds up the patient flow, we will also use a simulation model tailored to the operating conditions at the UNC ED to validate as well as evaluate the optimal policy found by the mathematical fluid model.

3.3 The Fluid Model

We consider the ED as a fluid system, where the patient flow coming into the ED is characterized by a deterministic function of time, $\lambda(t)$, which is the inflow rate at time $t \geq 0$, i.e., the number of patients arriving per unit of time. The system has a fixed, s, number of servers. Let $\mu(t)$ be the per server service rate at time t, i.e., the number of patients served per unit of time. $\mu(t)$ can take two values at any time point: $\overline{\gamma}(t)$, which denotes the maximum service rate per server at time t, and $\underline{\gamma}(t)$, which denotes the minimum service rate per server at time t. The decision is about how to switch $\mu(t)$ between these two values at any given point in time.

We study the problem over a finite horizon $t \in [0, T]$. One needs to decide when to use the maximum service rate in order to minimize the time average fluid level under a constraint on the length of time during which we can apply the maximum service rate. We let $x(t)$ denote the fluid level at time t. The time average fluid level, i.e., our objective function, can be expressed as

$$A = \frac{1}{T} \int_0^T x(t) dt.$$

We let $\overline{\delta}$ denote the upper bound on the length of time during which we can apply the maximum service rate. For this problem to be interesting and realistic, we assume that $0 < \overline{\delta} < T$. Let $I(t)$ be the indicator function representing whether the maximum service rate is applied at time t, i.e.,

for $t \in [0, T]$, let

$$I(t) = \begin{cases} 1 & \text{if } \mu(t) = \overline{\gamma}(t), \\ 0 & \text{if } \mu(t) = \underline{\gamma}(t). \end{cases}$$

Note that $I(t)$ is our decision variable. Based on the definition of $I(t)$, the constraint can be expressed as

$$\int_0^T I(t)dt \leq \overline{\delta}.$$

Also, we can express $\mu(t)$ in terms of $I(t)$ as

$$\mu(t) = \overline{\gamma}(t)I(t) + \underline{\gamma}(t)\left[1 - I(t)\right].$$

We can also express $x(t)$ using $\mu(t)$ as (Harrison, 1985)

$$x(t) = \sup_{0 \leq t' \leq t} \max\{x_0 + \int_0^t [\lambda(u) - s\mu(u)]du, \int_{t'}^t [\lambda(u) - s\mu(u)]du\},$$

where x_0 denotes the initial fluid level at time zero, i.e., $x_0 = x(0)$. Then our problem can be formulated as

$$\begin{aligned} \min_{I(t):\, t \in [0,T]} & \int_0^T x(t)dt \\ \text{s.t.} \quad & \begin{array}{l} x(t) = \sup_{0 \leq t' \leq t} \max\{x_0 + \int_0^t [\lambda(u) - s\mu(u)]du, \\ \qquad\qquad \int_{t'}^t [\lambda(u) - s\mu(u)]du\} \end{array}, \forall t \in [0, T], \\ & \mu(t) = \overline{\gamma}(t)I(t) + \underline{\gamma}(t)\left[1 - I(t)\right], \forall t \in [0, T], \\ & \int_0^T I(u)du \leq \overline{\delta}, \\ & I(t) = 0 \,\text{or}\, 1, \forall t \in [0, T]. \end{aligned}$$

The optimal solution to our problem should be dependent on the actual form of the rate functions $\lambda(t)$, $\overline{\gamma}(t)$ and $\underline{\gamma}(t)$. Using the 2012 patient data from UNC ED we can make some reasonable assumptions for the rate functions. Figures 3.1 and 3.2 show how the arrival and service rates (under normal operating conditions, i.e., without any BeRT) change over the course of a day at the UNC ED during 2012.

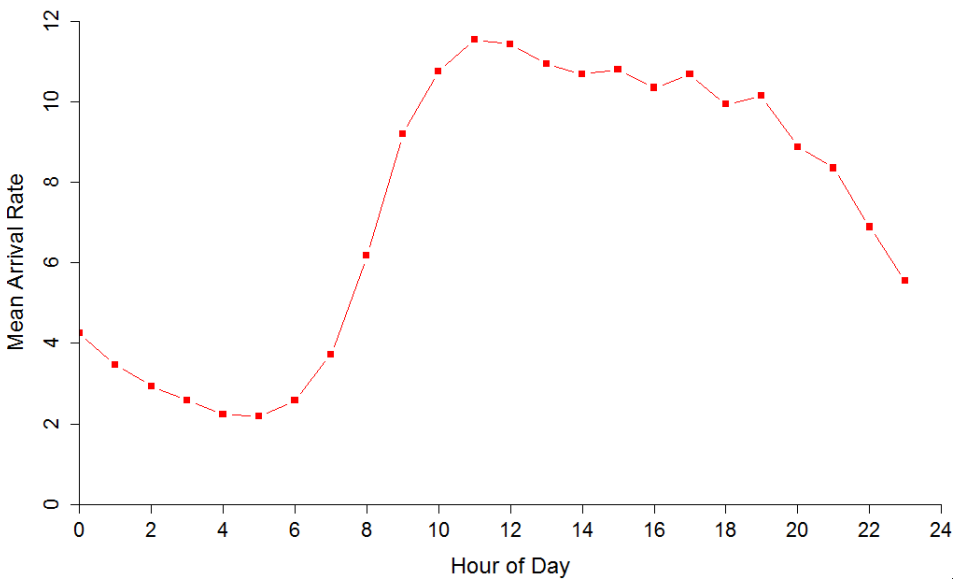

Figure 3.1: Arrival rate (number of arrivals per hour) vs. hour-of-day based on UNC ED 2012 data

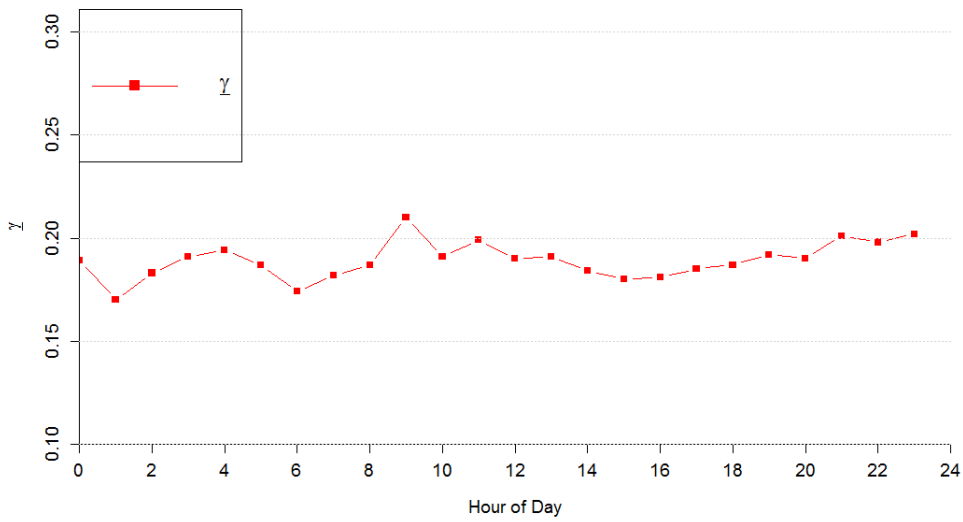

Figure 3.2: Service rate (number of patients served per hour) per server under normal operating conditions vs. hour-of-day based on UNC ED 2012 data

The arrival rate for each hour of day is taken to be the number of patients who arrived to the UNC ED during that hour divided by the total number of that hour in the year of 2012, which was 366. The service rate per server under normal operating conditions for each hour of day is taken to be the inverse of the mean sojourn time of all patients who start being served during that hour. It is easy to see that service rate is approximately constant over time. Based on this observation, we assume that both $\overline{\gamma}(t)$ and $\underline{\gamma}(t)$ aare constant over time, and we denote $\overline{\gamma}(t) = \overline{\gamma}$ and $\underline{\gamma}(t) = \underline{\gamma}$. Further, we observe that the arrival rate, $\lambda(t)$, although is not constant over time, approximately

grows linearly from early morning to noon, and then stay constant at its peak for a couple hours during the daytime, and then linearly declines. We call the period during which the arrival rate stays constantly high the peak hours (9am to 5pm). Upon discussing with the ED staff we reached the agreement that at current stage it is only necessary to BeRT during the peak hours because the ED is usually not crowded in other time of the day. Since the arrival rate is approximately constant over time during the peak hours, we also assume that $\lambda(t) = \lambda$ for our problem. With the assumptions that all rates are constant over time we can re-express the previous formulation of our problem as below

$$
\begin{aligned}
\min_{I(t):\, t\in[0,T]} & \int_0^T x(t)dt \\
\text{s.t.} \quad & x(t) = \begin{array}{l} \sup_{0\le t'\le t} \max\{x_0 + \int_0^t [\lambda - s\mu(u)]du, \\ \qquad\qquad \int_{t'}^t [\lambda - s\mu(u)]du\} \end{array}, \forall t \in [0,T], \\
& \int_0^T I(t)dt \le \overline{\delta}, \\
& I(t) = 0 \,\text{or}\, 1, \forall t \in [0,T].
\end{aligned}
$$

3.4 The Optimal Policy

Even after the assumptions that $\lambda(t)$, $\overline{\gamma}(t)$ and $\underline{\gamma}(t)$ stay constant during $[0,T]$, the optimization problem is still challenging as there can be infinitely many solutions where $I(t)$ takes the value 1/0 during infinite number of intervals that reside within $[0,T]$. To simplify the problem further, we limit the BeRT option to a single interval within $[0,T]$. Another reason for doing so is that implementing BeRT in the ED would entail a big change for the ED management team. We thus want to make it as easy to implement as possible. This way, the ED management will only need to implement BeRT during a single interval throughout the day. Under this case, we can fully characterize a BeRT policy using two values. Let $t_0 = \min_{0\le t\le T}\{I(t) = 1\}$ be the first time one starts to BeRT

and $t_s = \max_{0 \le t \le T}\{I(t) = 1\}$ be the time BeRT ends. Then $\mu(t)$ can be re-expressed as

$$\mu(t) = \begin{cases} \overline{\gamma} & \text{if } t_0 \le t \le t_s, \\ \underline{\gamma} & \text{if } t \in [0, T] \backslash [t_0, t_s]. \end{cases}$$

Also, the constraint on the total amount of time during which BeRT can be applied can be re-written as

$$t_s - t_0 \le \overline{\delta}.$$

Hence the optimization problem becomes

$$\begin{aligned} &\min_{t_0, t_s} \int_0^T x(t)dt \\ &\text{s.t.} \quad x(t) = \begin{array}{c} \sup_{0 \le t' \le t} \max\{x_0 + \int_0^t [\lambda - s\mu(u)]du, \\ \int_{t'}^t [\lambda - s\mu(u)]du\} \end{array}, \forall t \in [0, T], \\ &\quad \mu(t) = \begin{cases} \overline{\gamma} & \text{if } t_0 \le t \le t_s, \\ \underline{\gamma} & \text{if } t \in [0, T] \backslash [t_0, t_s], \end{cases} \\ &\quad t_s - t_0 \le \overline{\delta}, \\ &\quad 0 \le t_0 \le t_s \le T. \end{aligned} \tag{3.1}$$

The next lemma further reduces the number of decision variables to only one.

Lemma 4. *It is suboptimal to let $t_0 > T - \overline{\delta}$ and $t_s - t_0 < \overline{\delta}$.*

Although Lemma 4 is intuitive, it still requires a rigorous proof, which is provided in the Appendix. Based on Lemma 4, the problem reduces to finding only the optimal starting point for the BeRT interval, denoted by t_0^*, which should be searched in $[0, T - \overline{\delta}]$.

Theorem 4. *The optimal policy π^* that solves constrained problem (3.1) is to let $t_0^* = 0$ except when $\overline{\delta} > t_1$ and $t_3 \ge 0$, we have $t_0^* = \min\{t_2, t_3, T - \overline{\delta}\}$, where*

$$t_1 = \frac{x_0}{s\overline{\gamma} - \lambda},$$

$$t_2 = \frac{\overline{\delta}(s\overline{\gamma} - \lambda) - x_0}{\lambda - s\underline{\gamma}},$$

and

$$t_3 = \frac{T - \overline{\delta} - \frac{s(\overline{\gamma}-\underline{\gamma})}{(s\overline{\gamma}-\lambda)(\lambda-s\underline{\gamma})}x_0}{2 + \frac{\lambda-s\underline{\gamma}}{s\overline{\gamma}-\lambda}}.$$

The proof of Theorem 4 is provided in the Appendix.

CHAPTER 4
Numerical Study

4.1 Introduction

As mentioned briefly in previous chapters, one proposed idea to reduce ED congestion is to predict whether or not a patient will eventually be admitted to the hospital upon or shortly after arrival and request a bed from the hospital at that time. We term this strategy early BeRT (bed request) or call-ahead. Calling ahead has the potential to significantly reduce the time an admitted patient occupies a bed since the hospital bed the patient will transfer to might already be available by the time the "admit" decision for the patient is given or at least would be available soon after. However, if the patient for whom an "admit" prediction is made ends up being discharged from the emergency department, that would mean that hospital resources were unnecessarily used to make the bed available, which would also turn into a problem between the emergency department and the hospital.

In this chapter, we will conduct numerical studies to evaluate the performances of several early BeRT heuristic policies using a discrete-event simulation model built for the UNC ED. The measurements we use to compare the efficacy of different heuristics are the long-run average length-of-stay, waiting time, and daily number of false early BeRTs. A good heuristic policy will balance the trade-off between the length-of-stay and waiting time, and daily counts of false early BeRTs. The primary purpose of performing the numerical studies is that we could identify heuristics that are easy to implement in a real ED setting and perform reasonably well.

The results in Chapter 2 show that the optimal policies are of threshold type, where the threshold is on the admission probability and occupancy level. In light of this fact, we consider heuristic policies that decide whether to call ahead based on certain thresholds, which are either functions of the admission probability, or the queue length, or a combination of both. Note that in Chapter 2, we used the terminology "type-2 probability", "sequential/parallel service". In the

context described above, here type-2 probability corresponds to an individual patient's admission probability, parallel service option corresponds to requesting a hospital bed early on and starting the bed preparation process in parallel with serving the patient in the ED, and sequential service option corresponds to waiting until the patient finish being served at the ED to request a hospital bed, if the patient turns out to be an admit. Chapter 3, on the other hand, employs a fluid model approximation and constrains the option of early BeRT to a single time interval, for which the start and end times are determined by system parameters. We are also going to consider a heuristic policy that is motivated by this fluid model solution.

This chapter is organized as follows: First, we will discuss heuristic policies considered in this chapter. Second, we will describe the simulation model that we used to represent the UNC ED. Third, we will discuss our experimental setting and then present the results that we found by implementing the aforementioned heuristics on the simulation model.

4.2 Heuristic Policies

4.2.1 The Current System

The simulation model we use to approximate the UNC ED system was built using the patient data for calendar year 2012. From this point on, we refer to the simulation model under the 2012 operating condition the *current system*. In the current system, no early BeRT is implemented for any patient that visits the ED. We refer the policy that does not call ahead for anyone *Hcurrent*. In later sections, we will compare the performance of other heuristic policies to *Hcurrent* and focus on how much improvement we can get for length-of-stay and waiting time, given certain tolerance for daily false call-aheads.

4.2.2 Fixed Threshold Policy (FT)

FT simply uses a constant threshold on the admission probability, denoted as T, to make a decision for calling ahead. When a patient enters service, one checks the patient's probability of admission α, and calls ahead if and only if

$$\alpha \geq T,$$

where T is a constant between 0 and 1.

4.2.3 Time-Dependent Threshold Policy (TT)

TT is a myopic policy that only takes into account the cost associated with one single patient, despite of the system's state of crowding, where we use the cost structure described in Chapter 2. For a single patient, if one does not call ahead for a bed, the expected total cost incurred from the time when the patient enters service to the time the patient leaves the system is

$$C_w \left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2} \right) + \alpha C_s,$$

while if one chooses to make an early BeRT for the patient, the expected total cost is

$$C_w \left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2} - \frac{\alpha}{\gamma_1 + \gamma_2} \right) + C_p.$$

TT says that to make an early BeRT for each individual patient if and only if

$$C_w \left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2} - \frac{\alpha}{\gamma_1 + \gamma_2} \right) + C_p \leq C_w \left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2} \right) + \alpha C_s,$$

or equivalently,

$$\alpha \geq \frac{C_p}{C_s + \frac{C_w}{\gamma_1 + \gamma_2}}. \tag{4.1}$$

The policy can be time-dependent because both γ_1 and γ_2 can be dependent on time of day, and the patient type.

4.2.4 Census and Time-dependent Threhold Policy (CTT)

CTT chooses actions differently depending on whether the system has a queue or not. When the system has no queue, then CTT follows what TT does, i.e., to call ahead if and only if (4.1) holds.

The action that CTT takes when the system has patient(s) waiting is motivated by the optimal policy we found for the clearing model as described in the Appendix of this chapter. Simply put, all assumptions for the clearing model are the same as for the queueing model described in Chapter

2, except that the system does not have any arrivals from the beginning of time, but starts with a fixed number of customers to be served until empty. All the terminology and notations used in Chapter 2 carry over therein. For the clearing model, our objective is to minimize the total cost until the system is emptied by choosing which customers we apply the early BeRT option. The optimal policy we found there is to start early BeRT for the customer if and only if

$$nC_w \frac{\alpha}{\gamma_1 + \gamma_2} \geq C_p - \alpha C_s,$$

where n denotes the number of patients present in the system at the decision epoch and α denotes the admission probability of the patient for whom we are making an admit decision. To apply the intuition of this formula in a real ED setting where there are multiple servers and random arrivals to the system, we replace n in the formula, which represents the number of customers left in the clearing model system at the decision epoch, by the expected number of patients that will be served until the queue is first emptied, denoted by n_e, given the current number of patients in the queue, denoted by m. Additionally, assuming the number of servers is denoted by K, then the service rate is $K(\gamma_1 + \gamma_2)$ instead of $\gamma_1 + \gamma_2$.

It is straightforward to see that n_e is a function of m, and is also dependent on how one serves the patients because the total service time distribution for a patient is different if one chooses to early BeRT. That being said, we approximate the system behavior through a way which leads to CTT, which assumes that given m, the current number of patients in the queue, n_e is approximately equal to the number of patients served until the queue reaches empty state for the first time with the system behaving as a M/G/1 queue, where the service time distribution is the same if early BeRT is applied to everyone being served during that cycle.

Using the well-known formula in classic queueing theory for the first-passage time from any state m to state 0 for a M/G/1 system with arrival rate λ and mean service time τ we have that, one chooses to early BeRT for a patient with admission probability α when there are m patients in the queue if and only if

$$n_e(m) C_w \frac{\alpha}{K(\gamma_1 + \gamma_2)} \geq C_p - \alpha C_s,$$

where $n_e(m) = \frac{m}{1-\lambda\tau}$, and $\tau = \left(\frac{1}{\gamma_1} + \frac{\alpha}{\gamma_2}\right)^{-1} K^{-1}$. Let N denote the number of patients in the system, then $m = N - K$. By rearranging the terms and substituting it into the above formula we have

$$\alpha \geq \frac{C_p}{C_s + \left(\frac{C_w}{\gamma_1+\gamma_2}\right)\left(\frac{m}{K(1-\lambda\tau)}\right)}.$$

Now we are one step away from formulating CTT. Notice that at this point, our policy is to call ahead if and only if

$$\alpha \geq \frac{C_p}{C_s + \frac{C_w}{\gamma_1+\gamma_2}}.$$

when there is no queue, or $N \leq K$. And to call ahead if and only if

$$\alpha \geq \frac{C_p}{C_s + \left(\frac{C_w}{\gamma_1+\gamma_2}\right)\left(\frac{N-K}{K(1-\lambda\tau)}\right)}.$$

when there is a queue, or $N > K$.

However, we want to make CTT continuous in the sense that substituting $N = K$ into the formula for the case when $N > K$ gives a threshold the same as that for the case when $N \leq K$. To achieve this, we simply add a constant to the right hand side of the above formula, and thus CTT dictates that when $N > K$,

$$\alpha \geq \frac{C_p}{C_s + \left(\frac{C_w}{\gamma_1+\gamma_2}\right)\left(\frac{N-K}{K(1-\lambda\tau)}\right)} + \frac{C_p}{C_s + \frac{C_w}{\gamma_1+\gamma_2}} - 1.$$

4.2.5 Constrained Fixed Threshold Policy (CFT)

CFT is inspired by the optimal policy found for the fluid model described in Chapter 3. Unlike other heuristic policies, CFT has two control parameters, a fixed threshold for admission, and a length-of-time during the peak hours (9am-5pm), as defined in Chapter 3, for the early BeRT option. Let $0 < T < 1$ denote the cutoff for admission, and $1 \leq \overline{\delta} \leq 8$ denote the length-of-time during the peak hours when call-ahead is allowed, then CFT dictates that one will call ahead for a patient if and only if s/he arrives during 9am to $(9 + \overline{\delta})$, and her/his admission probability is no less than T.

Keep in mind that the higher service rate as assumed in the fluid model is achieved, in a real ED setting, by calling ahead for a group of patients that are identified as to-be-admits. This is done by setting a threshold for admission, and categorize a patient to the admit group if and only if the patient's admission probability, which is generated upon arrival in the simulation model according to certain distributions, exceeds the preset threshold. As mentioned briefly in Chapter 3, we are only going to implement early BeRT during the peak hours, which is taken to be 9am to 5pm based on Figure 3.1 where it is a period of time when the arrival rate seems to be constant over time. When implementing Theorem 3.4.1 shows that to the structure of the optimal policy depends on the values for x_0, λ, s, $\underline{\gamma}$, $\overline{\gamma}$, and $\overline{\delta}$. We assume that $x_0 = 0$, i.e., the number of patients in the queue at 9am, because before the peak hours the arrival rate is constantly small, which implies that there will not be much accumulation in the queue before 9am. We discussed the method of estimating λ and $\underline{\gamma}$ in Section 3.3. While here, since we have a simulation model that is validated using the 2012 UNC ED data, we use the simulation to re-estimate all the parameters in a similar vein. To be more specific, for the peak hours during each day of a week, λ is taken to be the mean number of hourly total patient arrivals during the peak hours. $\underline{\gamma}$ is the inverse of the mean sojourn time (service + boarding) of all patients that entered service during the peak hours. Additionally, here we take $s\underline{\gamma}$ as the mean number of hourly total patient departures during the peak hours, and thus $\hat{s} = \frac{s\hat{\underline{\gamma}}}{\hat{\underline{\gamma}}}$.

Table 4.1 below shows the estimated λ, $s\underline{\gamma}$, $\underline{\gamma}$, ans s using the simulation model under normal operating conditions (without call-aheads).

	Sun	Mon	Tue	Wed	Thur	Fri	Sat
$\hat{\lambda}$	10.1	12.3	10.9	10.9	10.9	10.8	10.1
$s\hat{\underline{\gamma}}$	6.8	7.1	7.2	6.9	6.9	6.8	7.1
$\hat{\underline{\gamma}}$	0.211	0.198	0.200	0.197	0.197	0.197	0.211
$\hat{s} = \frac{s\hat{\underline{\gamma}}}{\hat{\underline{\gamma}}}$	32.2	35.9	36.0	35.0	35.0	34.5	33.6

Table 4.1: Estimated for 9am to 5pm

The estimation of $\overline{\gamma}$ should depend on the threshold we use for admission. This is because, the higher the threshold, the less patients we are going to categorize as admit patients, and thus the less early BeRTs one will make, hence the less improvement one can obtain for the service rate $\overline{\gamma}$.

Once we fix the threshold for early BeRT, then one can estimate $\overline{\gamma}$ the same way as one estimates $\underline{\gamma}$.

Table 4.2 below shows the estimated $\overline{\gamma}$ under different thresholds for admission.

	Sun	Mon	Tue	Wed	Thur	Fri	Sat
$\hat{\overline{\gamma}}$ (0.9)	0.212	0.200	0.200	0.200	0.200	0.200	0.211
$\hat{\overline{\gamma}}$ (0.8)	0.212	0.201	0.202	0.201	0.200	0.201	0.212
$\hat{\overline{\gamma}}$ (0.7)	0.214	0.204	0.204	0.204	0.203	0.204	0.214
$\hat{\overline{\gamma}}$ (0.6)	0.216	0.206	0.206	0.206	0.205	0.207	0.216
$\hat{\overline{\gamma}}$ (0.5)	0.218	0.209	0.209	0.208	0.208	0.208	0.218
$\hat{\underline{\gamma}}$	0.211	0.198	0.200	0.197	0.197	0.197	0.211

Table 4.2: Estimated $\overline{\gamma}$ under different admission thresholds for 9am-5pm

The estimation of $\overline{\delta}$ takes a bit more effort. First, keep in mind that $\overline{\delta}$ in our fluid model formulation represents the maximum amount of time one can use the maximum service rate. In a real ED setting, managers care about the number of false positive early BeRT per day, which we will use to determine $\overline{\delta}$. Additionally, $\overline{\delta}$ is also dependent upon the admission threshold because the higher the threshold, the less patients we will categorize as admits, and the less false positives we will cause daily.

Let Λ denote the number of incorrect call-aheads the ED managers can tolerate per day. λ' being the peak-hour number of incorrect call-aheads per hour, which is dependent on the cutoff for admission, then

$$\begin{aligned}\lambda' &= \text{peak-hour number of incorrect call-aheads}\\ &= \text{peak-hour number of arrivals per hour} \times \text{impact} \times \text{fp}\\ &= \lambda \times \text{impact} \times \text{fp}.\end{aligned}$$

Where impact is the percentage of patients who have predicted admission probabilities no less than the BeRT cutoff. And fp is the percentage of patients who are incorrectly called ahead out of

those who have predicted admission probabilities no less than the BeRT cutoff. Consequently, we have

$$\overline{\delta} = \frac{\Lambda}{\lambda'}.$$

Given our APT and a threshold for admission, one can find the corresponding impact and fp. Tables 4.3 through 4.6 below present the estimated $\overline{\delta}$ under different given Λ and thresholds for admission. If the estimated $\overline{\delta}$ is greater than the length of the peak hour, i.e., 8 hours, then we automatically let $\overline{\delta} = 8$, because we do not call ahead outside of the peak hours. Also, we round up $\overline{\delta}$ to the nearest integer for simplicity of implementation.

cutoff	Su(9-17)	M(9-17)	Tu(9-17)	W(9-17)	Th(9-17)	F(9-17)	Sa(9-17)
0.5	1	1	1	1	1	1	1
0.6	2	1	2	2	2	2	2
0.7	3	3	3	3	3	3	3
0.8	7	6	7	7	7	7	7
0.9	8	8	8	8	8	8	8

Table 4.3: Estimated $\overline{\delta}$ under different admission thresholds for $\Lambda = 0.5$

cutoff	Su(9-17)	M(9-17)	Tu(9-17)	W(9-17)	Th(9-17)	F(9-17)	Sa(9-17)
0.5	1	1	1	1	1	1	1
0.6	3	3	3	3	3	3	3
0.7	6	5	6	6	6	6	6
0.8	8	8	8	8	8	8	8
0.9	8	8	8	8	8	8	8

Table 4.4: Estimated $\overline{\delta}$ under different admission thresholds for $\Lambda = 1$

cutoff	Su(9-17)	M(9-17)	Tu(9-17)	W(9-17)	Th(9-17)	F(9-17)	Sa(9-17)
0.5	2	2	2	2	2	2	2
0.6	5	4	5	5	5	5	5
0.7	8	8	8	8	8	8	8
0.8	8	8	8	8	8	8	8
0.9	8	8	8	8	8	8	8

Table 4.5: Estimated $\overline{\delta}$ under different admission thresholds for $\Lambda = 1.5$

cutoff	Su(9-17)	M(9-17)	Tu(9-17)	W(9-17)	Th(9-17)	F(9-17)	Sa(9-17)
0.5	3	2	2	2	2	2	3
0.6	7	6	6	6	6	6	7
0.7	8	8	8	8	8	8	8
0.8	8	8	8	8	8	8	8
0.9	8	8	8	8	8	8	8

Table 4.6: Estimated $\overline{\delta}$ under different admission thresholds for $\Lambda = 2$

4.3 Simulation Model

This section discusses the simulation model that we employ to evaluate the efficacy of different heuristic policies in terms of their impact on reducing ED crowding. The simulation model is an extension of an early version, which was built and consistently refined by previous graduate students working on other projects related to UNC ED. Interested readers can refer to (Ahalt et al., 2016) for the first version of the simulation model and the project where it was used. Since many of our assumptions for the mathematical models are drawn from the 2012 UNC ED patient data, the earlier version of the simulation model needs to be updated so that the input parameters reflects the operating conditions at the ED during that time. The content of this section is organized as follows: First, input parameter analysis using the 2012 UNC ED data; Second, validation of the simulation model based on the 2012 UNC ED data.

The simulation model captures the patient flow going through the UNC ED at a highly detailed level, which can be viewed as a queueing process that consists of five components: arrival, triage, service, boarding, and departure. Figure 4.1 gives a general overview of this queueing process:

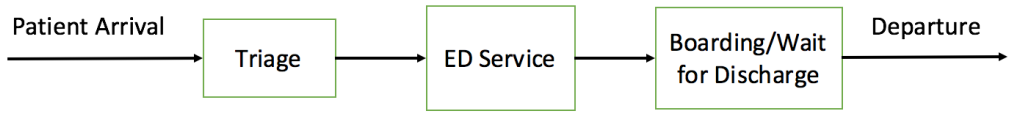

Figure 4.1: Patient flow at The ED

In the simulation model, patient arrivals are generated based on nonhomogeneous Poisson processes for which the arrival rates are dependent on hour of the day, day of the week, and the patient type. Patients are divided into groups based on their age, ESI (Emergency Severity Index) levels, and disposition categories. A patient is classified as a pediatric patient if he/she is younger than 18, otherwise he/she is an adult patient. Adult patients can be admitted to one of three wards: A, B, or D where A and B are open 24/7 and D is open from 9am to 2am. Pediatric patients are generally served in a pediatrics ward which is open from 9am to 2am. Pediatric patients can,

however, be admitted to ward A or B during after hours. Table 4.7 summarizes the hours of four wards' and bed capacities.

Ward	Hours	Bed Capacity
A	24/7	19
B	24/7	16
D	9am-2am	15
Peds	9am-2am	9

Table 4.7: Ward hours and bed capacity

ESI measures the severity of a patient's medical condition. There are five ESI levels ranging from 1 to 5 with lower numbers indicating higher criticality. Finally, there are two disposition categories: admitted and discharged. Admitted patients will be hospitalized after their ED visits while discharged patients leave the hospital system immediately after their ED visit.

Arriving patients join the queue for triage where they get assigned an ESI level. There are typically two triage nurses at triage and their service times are assumed to follow i.i.d. triangular distribution. We make this assumption on the distribution and its parameters based on experience of ED managers that we collaborate with since our data does not have triage times.

After triage, patients join a queue to wait for ED bed assignment. In the simulation model, we do not explicitly model the attending physicians or any other medical personnel. Instead, we regard each ED bed as a server. The first part of service a patient will receive at the ED starts when the patient is assigned to an ED bed and ends when a disposition decision is made for him/her.

The second part of ED service starts when a disposition decision is made for the patient and ends when the patient leaves the ED, during which time the patient will remain in his/her ED bed, keeping it blocked from being used by other patients. If it is decided that the patient does not need hospitalization, then the patient is of a discharge type, and he/she leaves the ED (and thus free the ED bed) shortly after the disposition decision is made. Otherwise, the patient is an admit type patient and the patient will wait a longer time (boarding) for the hospital to prepare a bed for him/her. For both parts of the service process, we estimate the service time distributions based on the data, where the distribution is dependent upon time and patient type (i.e., age, acuity level, and disposition categories).

4.3.1 Input Analysis

I performed input parameter estimations using the 2012 UNC ED patient data. The information available for each visit includes the patient's age, gender, ESI, disposition, arrival time, ED bed assignment time, disposition decision time and departure time. Entries with missing data or out-of-order time stamps are deleted. The cleaned data has approximately 56,000 entries (corresponding to 56,000 patient visits).

As mentioned earlier, the parameters that needed to be estimated are the arrival rates, service time, and boarding time distributions. Arrival rates are dependent on hour of the day, day of the week, and patient type broken down by age (adult vs. pediatric), ESI and disposition (admitted vs. discharged). Figure 3.1 displays the hourly average arrival rate for all patient types combined. Note that this is just a demonstration of the time varying nature of the arrival rate. In the actual simulation model the patients arrive according to time-varying arrival rates based on their types. The service times are dependent on hour of the day and patient type broken down by age (adult vs. pediatric) and ESI. Lastly, boarding times are dependent on hour of the day and patient type broken down by age (adult vs. pediatric), ESI and disposition (admitted vs. discharged).

Tables 4.8 through 4.11 below present the fitted service time and boarding time distributions. When fitting the distribution, we use Kolmogorov-Smirnov test (KS test) to determine the best fit. The p-values associated with all tests are provided in the tables. A large p-value means that the fitted distribution does not deviates from the empirical distribution significantly. Most fittings are good with a p-value greater than 0.05. For those that are not so good we also report the MSEs (Mean Squared Error) associated with the tests. In all cases where the p-value is not indicative of a good fit the MSE is in the order of 0.001, which means that the fitting is acceptable. In tables 4.8 through 4.11 the following notation is used to distinguish between patient types. The first letter can take two values A and P which correspond to Adult and Pediatric, respectively. The second letter can take two values A and D which correspond to Admitted and Discharged, respectively. And finally the numbers 1 through 5 represent ESI level. For example, a pediatric admitted patient with ESI level being 3 is abbreviated by PA3. And the following notation is used to represent random distributions. GAMM, ERLA and WEIB each stands for the Gamma, Erlang and Weibull distribution respectively, where the first and second parameters are the scale

and shape parameters. EXP stands for the Exponential distribution where the parameter is the mean. N stands for the Normal distribution where the first and second parameters represent the mean and standard variance. And finally, we use the notation $\mathrm{EXP}(p\mathrm{N}(\mu_1, \sigma_1) + (1-p)N(\mu_2, \sigma_2))$ to represent a random variable Y, such that $\log Y$ follows a mixed Normal distribution. To be more specific, with probability p, $\log Y$ follows $\mathrm{N}(\mu_1, \sigma_1)$, and with probability $1-p$, $\log Y$ follows $\mathrm{N}(\mu_2, \sigma_2)$. Also in the tables, p-value is based on KS test that examines the closeness between the fitted and empirical distribution.

Patient type	Hour	Service time distribution	P-value	MSE
A1	0-23	1+GAMM(101,0.952)	>0.15	
A2	2-9	1+GAMM(185,1.37)	0.022	0.00443
A2	9-14	4+ERLA(125,2)	0.027	0.00183
A2	14-20	EXP(0.68N(5.45,0.54)+0.32N(5.14,0.24))	>0.15	
A2	20-2	EXP(0.97N(5.41,0.82)+0.03N(2.63,0.90))	0.060	
A3	2-9	EXP(0.9N(5.43,0.60)+0.1N(4.31,0.96))	0.059	
A3	9-20	1+ERLA(83.9,3)	¡0.01	0.00026
A3	20-2	EXP(0.92N(5.38,0.58)+0.08N(4.23,1.02))	>0.15	
A4	2-9	1+GAMM(95.5,1.49)	0.090	
A4	9-14	1+GAMM(90.4,1.47)	0.070	
A4	14-20	1+GAMM(80.6,1.54)	0.134	
A4	20-2	1+ERLA(71.3,2)	0.028	0.00082
A5	0-5	1+EXP(105)	>0.15	
A5	5-23	EXP(0.8N(4.00,0.84)+0.2N(2.96,1.06))	>0.15	

Table 4.8: Service time distributions for adult patients

Patient type	Hour	Serivce time distribution	P-value	MSE
P1	0-23	2+EXP(78.7)	>0.15	
P2	2-9	7+WEIB(240,1.06)	>0.15	
P2	9-14	6+WEIB(305,1.38)	>0.15	
P2	14-2	EXP(0.95N(5.25,0.86)+0.05N(2.82,0.86))	0.116	
P3	0-23	2+GAMM(79.7,2.33)	0.050	0.00026
P4	0-23	1+GAMM(50.9,2.37)	0.095	
P5	0-23	1+GAMM(36.8,2.27)	>0.15	

Table 4.9: Service time distributions for pediatric patients

Patient type	Hour	Boarding time distribution	P-value	MSE
AA1	2-9	17+WEIB(122,0.935)	>0.15	
AA1	9-12	WEIB(180,1.05)	>0.15	
AA1	20-2	WEIB(153,0.955)	0.0423	0.00706
AA2	5-16	EXP(0.95N(5.47,0.634)+0.05N(3.77,1.31))	>0.15	
AA2	16-5	EXP(0.61N(5.03,0.440)+0.39N(5.20,1.22))	>0.15	
AA3	5-16	EXP(0.93N(5.50,0.563)+0.07N(4.43,0.998))	>0.15	
AA3	16-5	EXP(0.63N(5.05,0.419)+0.37N(5.37,1.00))	>0.15	
AA45	0-5	2+GAMM(188,1.18)	0.069	
AA45	5-0	11+ERLA(99.2,2)	>0.15	
AD12	3-8	1+943BETA(0.183,2.11)	0.001	<0.001
AD12	8-3	WEIB(34,0.616)	<0.01	0.00328
AD3	5-11	WEIB(33.6,0.663)	<0.01	0.00693
AD3	11-5	WEIB(25.1,0.657)	<0.01	0.00256
AD4	2-9	EXP(25.6)	<0.01	0.00583
AD4	9-2	EXP(19.9)	<0.01	0.00061
AD5	2-9	EXP(23.1)	0.046	0.00997
AD5	9-2	EXP(15.9)	<0.01	0.00435

Table 4.10: Boarding time distributions for adult patients

Patient type	Hour	Boarding time distribution	P-value	MSE
PA12	2-14	EXP(0.87N(4.81,0.923)+0.13N(5.17,0.233))	>0.15	
PA12	14-2	EXP(0.78N(4.79,0.543)+0.22N(4.25,1.45))	>0.15	
PA3	2-14	EXP(0.97N(5.06,0.666)+0.03N(2.04,0.829))	>0.15	
PA3	14-2	EXP(0.81N(4.87,0.496)+0.19N(4.72,1.24))	>0.15	
PA45	0-23	GAMM(92,1.79)	>0.15	
PD12	9-14	EXP(30.8)	<0.01	0.00432
PD12	14-20	EXP(26)	0.081	
PD12	20-9	EXP(28.3)	0.020	0.00507
PD3	4-10	EXP(25.4)	0.019	0.01015
PD3	10-4	EXP(20.6)	<0.01	0.00158
PD45	5-12	EXP(20.8)	>0.15	
PD45	12-5	EXP(18)	<0.01	0.00217

Table 4.11: Boarding time distributions for pediatric patients

4.3.2 Calibration and Validation

To validate that the simulation model we developed reflects the operating condition of the UNC ED in 2012, we compare the output of the simulation model with that estimated from the UNC ED 2012 data directly. The three metrics we consider are mean service time over the course of a day, mean boarding time by hour of day, and mean total length-of-stay by hour of day. Note that service time and boarding time distributions are estimated from the data and are direct input parameters in the simulation model, hence one should expect the output of them to be closely aligned with those from the data. Total length-of-stay is the sum of triage time, waiting time, service time, and boarding time. For triage time, since we do not have any data we impose an artificial triangular distribution with mean being based on ED nurses' suggestion. Waiting time is an organic product of running the simulation. This is because once the simulation is running, there will be limited number of servers (ED beds), with capacities changing by time of the day, and infinite arrivals to the system. Thus patients who arrive to a busy system where all servers are occupied will experience certain amount of waiting. Consequently, by comparing total length-of-stay from two

the simulation model and that of the original data, we will be able to see whether our assumptions on the bed capacities, arrivals, work together perfectly to produce a simulation model that mimics the behavior of the original system in 2012.

One important thing to note here is that in the simulation model, the notion of servers is modeled as ED bed resources. However, in the data we have available, the beginning of service time is defined as the first time that a patient was attended by an ED provider, which is not necessarily the first time that the patient gets assigned an ED bed. Because of the discrepancy, we had to calibrate bed capacity to achieve a match between the output of the simulation model and that estimated from the data for the three time measurements we consider. The resulting bed capacity is different from that of Table 4.7, and is summarized in Table 4.12.

Ward	Hours	Bed Capacity
A	24/7	12
B	24/7	16
D	11am-11pm	17
Peds	9am-2am	9

Table 4.12: Ward hours and bed capacity

After the aforementioned calibration, we arrive at a simulation model that accurately represent the operating condition of the UNC ED in 2012, as measured by service time, boarding time, and sojourn time. Figure 4.2 shows the result.

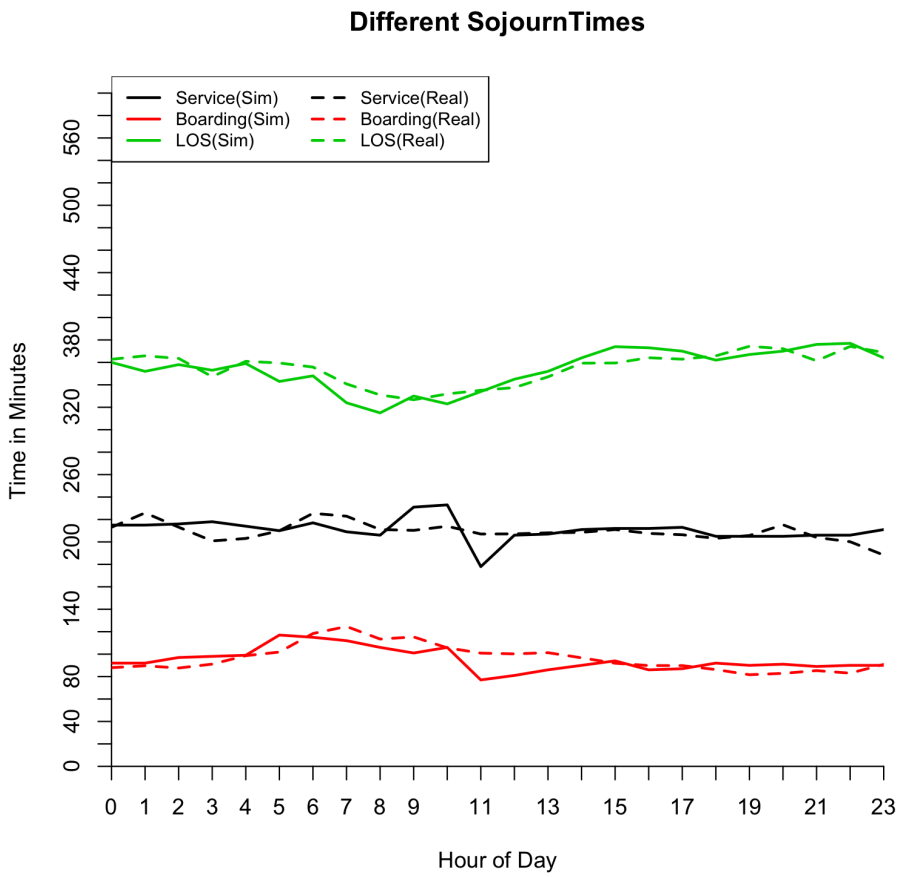

Figure 4.2: Sojourn Times Validation

4.4 Numerical Study

In this section, we will present results of a numerical study on the aforementioned heuristic policies, and compare their performances in terms of long-run average length-of-stay (LOS), where the average is taken over all patients. Keep in mind that all the heuristic policies we considered compare individual patient's probability of admission with a certain threshold, which can be a fixed constant (as in FT), or dependent on time (as in TT), or both the time and the system state (as in CTT). Consequently, depending on our assumptions on the admission probability distributions for all patients, each heuristic policy shall perform differently.

A natural assumption on the admission probability distribution would be to use the empirical distributions, as estimated by our APT. To be more specific, we applied APT on the 2012 UNC ED data, obtained a predicted admission probability for each individual patient, and fit an empirical distribution for each individual patient group broken down by their acuity, age (adult vs. pediatric), and disposition category (admit vs. discharge). According to the histograms of patients' predicted admission probabilities, we observe only a few distinct bars in all the histograms, meaning that for each individual patient group, the predicted admission probabilities tend to occur at a few distinct values most frequently. Based on this observation, for the empirical distributions we fit, we assumed uniform distributions between those most frequent values.

We examined the performance of all four aforementioned heuristic policies under the setting of empirical distribution. Figure 4.3 through Figure 4.5 present the result of our numerical study, and each compares the LOS under CTT with one other heuristic policy. Each point on the lines under the FT policy is obtained by varying the control parameter T, our fixed threshold for admission. For TT and CTT policies, the points are obtained by varying the cost parameter C_w, i.e., the waiting cost rate. Under CFT, we vary two control parameters, which are T, the fixed threshold for calling ahead, and $\overline{\delta}$, the length of time we call ahead during the peak hour. Each pair of parameters gives us one resulting length-of-stay and daily false positives. For each fixed daily false positives, we handpicked the threshold that renders the shortest length-of-stay, which corresponds

to each point in the plots. Notice that in the figures we also provide the 95% confidence interval (CI) bands around the mean values.

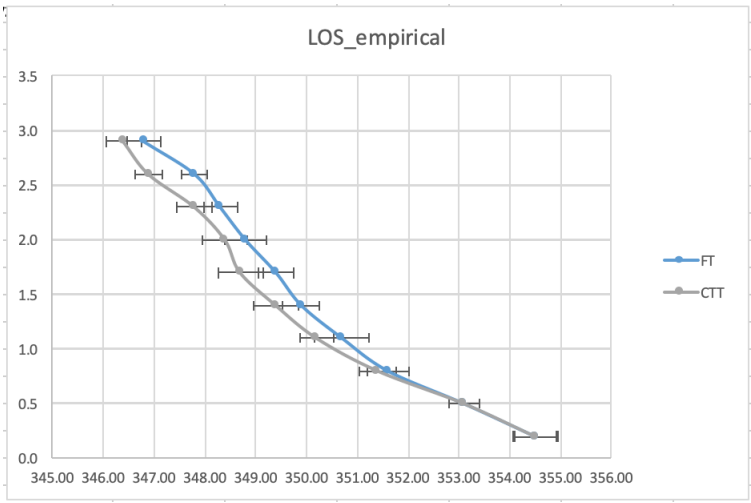

Figure 4.3: Length-of-stay (LOS) under CTT and FT

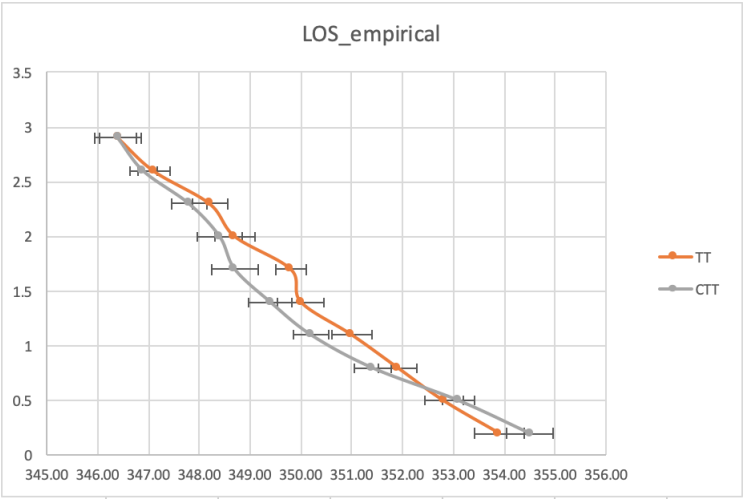

Figure 4.4: Length-of-stay (LOS) under CTT and TT

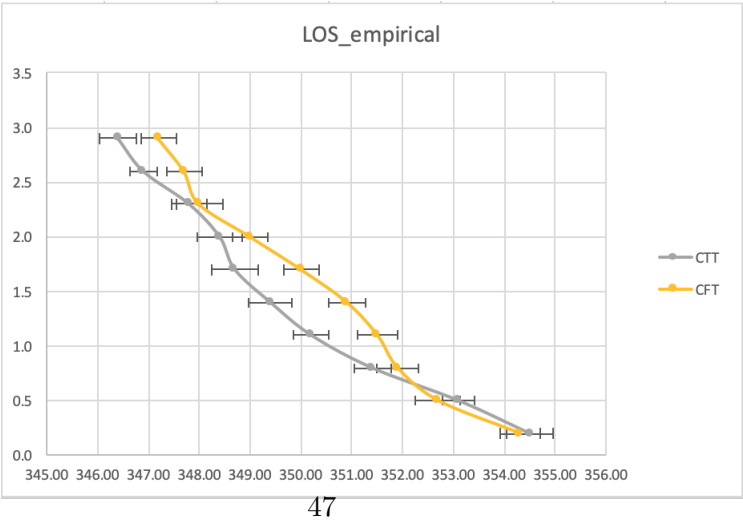

Figure 4.5: Length-of-stay (LOS) under CTT and CFT

4.5 Discussion

As is shown in Figure 4.3 through Figure 4.5, we looked at cases where the daily counts of false early BeRTs are between the value of 0 and 3. The general pattern is that the higher the daily counts of false positives, the larger the improvement on LOS and waiting time one achieves. Under the *current system*, the average LOS is 358min. It is evident that as the variances of the distributions decreases, larger improvement on the LOS can be expected given certain level of daily false positives. When one allows for 3 daily false positives, the heuristics result in a LOS in the range of 346min to 347min, corresponding to 11 to 12min reduction.

In addition, the three figures all show that the curve under CTT is always at or below the curve under other policies. This means first, given a fixed LOS, CTT results in less daily false positives. Note that in the figures we omit the vertical confidence intervals because they are almost neglegible compared to the horizontal ones. Second, given a fixed daily false positive, CTT results in a LOS that is no greater than that under other heuristic policies, in a statistically significant sense. This pattern is even stronger for a daily false positive that is greater than 1.5. For these two reasons, we conclude that CTT dominates the other heuristics in a statistically significant sense. And we project that the dominance could even be more significant if one allows for even more daily false positives.

CHAPTER 5

Impact of Census on Emergency Department Providers' Triage and Admission Decisions

5.1 Introduction

Emergency Departments (EDs) are busy places. In 2015 there were 136.9 million ED visits in the United States. This high volume often leads to ED crowding that has been associated with numerous negative patient outcomes including delays in lifesaving care that result in increased mortality and low patient satisfaction (George and Evridiki, 2015), (McCarthy et al., 2009), (Richardson et al., 2006), (McCusker et al., 2014).

It has been suggested that crowding of the emergency department can lead to difficulties with clinician decision-making and potentially impact equity in care (Hwang et al., 2011). Two such vital decision points that are tied to care quality and equity are the triage level assignment decision made by nursing staff and the disposition decision made by providers.

Nationally, emergency departments represent a significant source of hospital admissions accounting for nearly all the growth of hospital admissions in recent years (Morganti et al., 2013). The decision to admit a patient is made by emergency providers based upon available individual patient data, however recent research suggests that this decision may also be influenced by crowding of the ED itself (Gorski et al., 2017). This recently published study at a single academic medical center finds a statistical association between the likelihood of hospital admission and increased ED census. It was suspected that as EDs become busier there is a cognitive offloading that occurs for the physician by admitting patients rather than spending time and mental energy arranging safe discharges for patients who may be in a "gray area".

Making a disposition decision sooner during an individual patient's visit rather than waiting to see if a patient improves during the ED stay allows physicians to move on to see the next patient or complete the next task. There is some evidence from literature that as load increases in a system,

workers speed up their service rate (Kc and Terwiesch, 2009) and this effect may be what is being observed during times of high ED volume. Physicians may be, in effect, speeding up their services and increasing their "productivity" by choosing admission over discharge for patients who are in the gray area and for whom the right decision is not clear. Another study found that as the ED becomes more crowded the number of patients who are admitted to the hospital and have less than a 24-hour hospital stay increases; suggesting that some of these admissions that occur during times of high census may be avoidable (Freeman et al., 2017).

In other areas of healthcare, this relationship between decision making and crowding has also been found. One study found a correlation between ICU occupancy level and the rate of ICU discharges (Kc and Terwiesch, 2012). Another study found a similar relation in obstetrics, where midwives were more likely to refer high complexity patients to obstetricians at times of increased congestion as opposed to when census levels are much lower (Freeman et al., 2016).

This change in decision-making seems to occur even though it further contributes to system congestion. Ironically, boarding of admitted patients is thought to be a sizable contributor to crowding itself resulting in throughput delays of both admitted and discharged patients at an ED (Fogarty et al., 2014), (Kang et al., 2014). Understanding the relationship between ED census and individual provider and nurse decision-making may provide opportunity for operational changes in workflow to prevent decision fatigue at times of high census. Previous work has demonstrated the existence of a safety tipping point (Kuntz et al., 2014). Knowing that such a point exists and where it lays can aid in operational planning.

In addition to the admission decision, another critical decision that is made during a patient's ED visit is the triage classification. This is often the first important decision made during a patient's ED visit affecting how quickly the patient is evaluated by a provider. Only one other study has investigated the relationship between ED crowding and triage decisions and they concluded that there was no association (Richardson, 1998). Note that this study used the Australasian National Triage scale at a single tertiary care hospital in Australia. Furthermore, it treated patient census as a binomial categorical factor of "busy" or "non-busy" utilizing a single value to separate the two. A "busy" weekday in this study was defined as > 140 visits whereas ≤ 139 visits would constitute a "non-busy" weekday.

The aim of our study was to use statistical methods to test the hypotheses that ED census was associated with changes in triage and disposition decisions at an academic hospital in Southeastern US. To the best of our knowledge, our study is the first to look at ED census and triage assignment decisions by using the census level directly in the analysis rather than introducing arbitrary binary classifications (e.g., busy vs. non-busy) for the census level. Therefore, our modeling framework supports the exploration of how census count is associated with triage or admission decisions along the complete range of observed census levels.

5.2 Methods

5.2.1 Study Design and Setting

Following approval from the institutional review board, we performed a retrospective study using a data set of patient visits collected at the ED of an academic hospital in the Southeastern US. During the study period, which covered the year 2012, this ED received approximately 184 patient arrivals per day (67,203 patient visits per year). The triage system in place was the 5-level Emergency Severity Index (ESI) triage system, with levels from ESI 1 (patient dying) to ESI 5 (no ED resources needed) (Gilboy et al., 2012). At the time of the study the ED had 59 beds spread across five adult pods: A, B, C, D, and a behavioral health ED (BHED), as well as a pediatric pod. Pods A and B operated 24 hours a day seeing acute adult patients while pod D operated during peak hours and cared for primarily lower acuity patients. Pod C and BHED were dedicated to behavioral health patients although occasionally other patients were housed in these areas. Due to the non-homogeneity and inconsistent nature of their visits to the ED and hospital, behavioral health patients were excluded from our statistical analysis.

5.2.2 Data Analysis

The data available for each patient included demographic information (age, gender, and race), clinical information (triage acuity/ESI and chief complaint), disposition category (admit or discharge), and place of treatment (pod). Our goal was two-fold, to investigate the association between census and nurses' triage decision, and similarly the association between census and physicians' admission decision. We also considered other available variables as potential control variables in

the model (e.g., a patient's age may impact either the triage nurses' assessment or the admission decision by the provider) with reference to the relevant literature.

The data were cleaned before use in the statistical models. We deleted questionable data elements including but not limited to obviouserroneousentries, patient walkouts, behavioral health visits, or timeelements thatoccurred innon-chronologicorder. Additionally, we excluded patients with invalid or missing acuity scores. Duplicate records and those with missing or insufficient entries for the variables of interest were also excluded from the study. Whereas the original data had approximately 67,203 entries, after cleaning the data set contained 65,065 validated patient encounters eligible for statistical modeling.

Patient age was categorized into 8 clinically meaningful groups: <3month(m) old, 3m to 3, 3 to 8, 8 to 18, 18 to 40, 40 to 55, 55 to 70, and ≥70. These age groups were included as the levels of a categorical variable in subsequent statistical modeling. All other variables were also treated as categorical with the exception of census level, which was included in all models as a continuous variable, enabling us to associate any observed census count with the likelihood of admission or triage decisions. For race and pod, we combined categories that have less than 10 outcomes of each type of response (Agresti, 2003) to a single category named "Other".

Exploratory analysis confirmed that a patient's chief complaint could be highly predictive of admission and hence was a desirable component to include in the model. To control the complexity of the model, we selected the 45 most common chief complaints (out of 8,000), which had sufficient numbers of occurrences as to be informative. These 45 chief complaints were included explicitly in the model as levels of the "chief complaint" factor. (For a list of these 45 chief complaints, see Table A.1. All other chief complaints were included in the "Other" category. This way, we retained much of the information contained in the chief complaint data while limiting the complexity of the model.

Census, which was our primary control variable of interest, refers to the total number of patients in the ED, i.e., the number of patients in the waiting room and those occupying a bed. For our analysis of triage decisions, the census level used for each triage decision was the census level at the time of the corresponding patient's arrival, whereas for the analysis of disposition decisions, the census level was computed at the disposition decision time of the corresponding patient. Table 5.1 illustrates the breakdown of characteristics of all the patients in the cleaned data set with

the exception of chief complaints (due to its large number of categories) and census (because it is treated as a continuous variable). Prior to model fitting, we performed an exploratory data analysis to assess the univariate association between the control variables and the outcomes, i.e., triage level/ESI and disposition (admit and discharge). Also, we have not found any significant multicollinearity among control variables as we explain in more detail in Appendix. All data and statistical analysis in this work was performed in R.

Table 5.1: Breakdown of patient characteristics for variables of interest.

Characteristics	Percent in data set
Disposition	
Admit	29.6
Discharge	70.4
ESI	
1	0.9
2	13
3	57
4	24.9
5	4.2
Gender	
Female	54.6
Male	45.4
Race	
African American	30.0
Asian	1.1
Caucasian	53.8
Native American	0.4
Other	12.3
Unknown	2.4

Continued on next page

Table 5.1 – continued from previous page

Characteristics	Percent in data set
Age group	
$<$ 3m	0.8
3m to 3	5.2
3 to 8	4.7
8 to 18	7.5
18 to 40	34.3
40 to 55	21.6
55 to 70	15.3
$\geq$ 70	10.6
Pod	
A	27.8
B	23.4
C	2.8
D	27.2
Pediatrics	15.7
BHED	3.1

5.2.3 Statistical Modeling

Association between census and triage decision: To investigate how census might impact triage nurses' assignment of acuity levels, we fit a cumulative logit model (Agresti, 2003). We collapsed the five level ESI scale into three acuity groups: low (ESI 4/5), medium (ESI 3) and high (ESI 1/2). This reduced the complexity of the response variable in the model (acuity assignment) without losing much information as relatively few patients in the data set were assigned an ESI 1 or ESI 5 score. This resulted in a three-level cumulative logit model with low, medium or high acuity group as the response variable, which depended on census and the other relevant independent

variables discussed previously. Specifically, the cumulative logit modeling approach enabled us to understand how an independent variable (such as census) may be associated with the likelihood of a patient being placed into each of the categories of interest (such as low, medium or high acuity).

After the exploratory analysis, we conducted likelihood ratio tests between several candidate models (with different sets of independent variables) to identify a final model sufficient for testing the following hypothesis: ED census count has an impact on the likelihood of a patient being triaged in the low, medium or high category by the triage nurse. Table 5.2 provides the control variables of the resulting cumulative logit model for acuity group (low, medium, high) as the dependent variable and the p-value results of the likelihood ratio tests for each control variable. Note that all independent variables included in this cumulative logit model are significantly associated with the dependent variable at a 0.01 level of confidence.

Association between census and admission decision: In this part of the study, we fit a multivariate logistic regression model to assess the association between the disposition decision and census, which is calculated at the time a disposition decision is made for the corresponding patient. The logistic regression model is similar to the cumulative logit model, but only has two categories (admit or discharge) for the dependent variable. We considered multiple models and conducted likelihood ratio tests to identify which control variables to include in the final model. The control variables in the final model and the corresponding p-value results of the likelihood ratio tests for model selection are provided in Table 5.2. Note that all independent variables included in the final

logistic regression model are significantly associated with the dependent variable at a 0.05 level of confidence.

Cumulative Logit Model for Triage Decision	
Control variable	P-Value
Race	<0.01
Gender	<0.01
Age group	<0.01
Chief complaints	<0.01
Census	<0.01
Multivariate Logistic Regression Model for Disposition Decision	
Control variable	P-Value
Race	<0.01
Gender	<0.01
Age group	< 0.01
Acuity	<0.01
Pod	<0.01
Census	0.014
Chief complaint	<0.01
Interaction between age and acuity	<0.01

Table 5.2: P-values from likelihood ratio tests for all independent variables included in the selected cumulative logit model for triage decisions and multivariate logistic regression model for disposition.

5.3 Results

To estimate the impact of census on triage acuity assignment and disposition decision, we calculated odds ratios (OR) (Agresti, 2003) for both statistical models discussed in the statistical modeling section above. Specifically, in this case, the OR indicates how changes in a control variable (such as census) may increase or decrease the likelihood (odds) being assigned to a higher acuity level or being admitted. We next discuss our findings from each model separately.

Association between census and triage decision: We found by fitting the cumulative logit model with partial proportional odds that the relationship between nurses' triage decision and census (at time of arrival) was statistically significant. The OR for a patient being triaged as high acuity versus low or medium is 1.011 times greater when census is increased by one unit (95% CI = [1.009, 1.012]). We also found that for triaging a patient as medium or high versus low acuity is 1.009 times higher when census is increased by one unit (95% CI = [1.008, 1.010]). Results on odds ratios for all variables are reported in Table 5.3 except for chief complaints, which are provided in Table A.1 in Appendix. Using the cumulative logit model, we also calculated the marginal probabilities of being assigned each acuity level (low, medium, and high) at different census levels for a common group of patients (Caucasian females aged between 18 to 40 who had abdominal pain as their chief complaints); see Figure 5.1. Such a framework is useful for interpreting results for key patient subpopulations.

Association between census and admission decision: In the multivariate logistic regression model fitting, we found that there was a statistically significant association between providers' admission decision and census at the time when disposition decisions are made. The OR for admission per patient increase in census was 1.007 (95% CI = [1.006, 1.008]). ORs from the multivariate logistic regression analysis are reported in Table 5.4 except for chief complaints and interaction terms, which are provided in Table A.2 and A.3, respectively, in Appendix. For an example of the logistic regression model, we computed the probability of admission for a common group of patients: Caucasian females who are aged between 18 and 40, categorized as ESI3, with a chief complaint of abdominal pain and treated in Pod A, at different levels of census. The result is shown in Figure 5.2. The slope of the line is the same for all patients in the model however the probability of admission is higher or lower based on individual patient characteristics.

Table 5.3: Odds ratios of Prob(high acuity) versus Prob(low or medium acuity) and Prob(medium or high acuity) versus Prob(low acuity), and corresponding 95% confidence intervals for intercept, census, race, gender, and age group.

	P(high)/P(low or medium)	P(medium or high)/P(low)
Intercept		
		Continued on next page

Table 5.3 – continued from previous page

	P(high)/P(low or medium)	P(medium or high)/P(low)
	.057 [.052, .063]	1.403 [1.315, 1.496]
Census		
	1.011 [1.009, 1.012]	1.009 [1.008, 1.010]
Race (contrast: Caucasian)		
African American	.699 [.661, .739]	.693 [.665, .722]
Asian	.792 [.628, 1.002]	.898 [.759, 1.062]
Native American	1.219 [.847, 1.752]	1.387 [.998, 1.928]
Other	.540 [.493, .592]	.778 [.735, .822]
Unknown	.994 [.852, 1.160]	.898 [.800, 1.007]
Gender (contrast: Female)		
Male	1.345 [1.282, 1.410]	.901 [.869, .935]
Age Group (contrast: 18 to 40)		
$<$ 3m	2.143 [1.683, 2.729]	.970 [.799, 1.178]
3m to 3	.644 [.554, .749]	.422 [.390, .457]
3 to 8	.794 [.687, .918]	.462 [.426, .500]
8 to 18	1.741 [1.591, 1.905]	.812 [.760, .868]
40 to 55	1.165 [1.088, 1.247]	1.401 [1.334, 1.470]
55 to 70	1.551 [1.445, 1.664]	2.494 [2.343, 2.655]
$\geq$ 70	1.705 [1.577, 1.844]	5.601 [5.076, 6.181]

Table 5.4: Odds ratios of Prob(admit) versus Prob(discharge) and corresponding 95% confidence intervals for intercept, census, race, gender, acuity, age group, and pod.

	Prob(admit)/Prob(discharge)
Intercept	
	3.188 [2.763, 3.679]
	Continued on next page

Table 5.4 – continued from previous page

	Prob(admit)/Prob(discharge)
Census	
	1.007 [1.006, 1.008]
Race (contrast: Caucasian)	
African American	1.033 [.985, 1.084]
Asian	.892 [.729, 1.093]
Native American	2.138 [1.556, 2.938]
Other	.823 [.764, .887]
Unknown	.807 [.695, .938]
Gender (contrast: Female)	
Male	1.218 [1.167, 1.271]
Acuity (contrast: ESI3)	
ESI1	20.891 [12.519, 34.861]
ESI2	3.687 [3.313, 4.104]
ESI3	.115 [.095, .139]
ESI4	.018 [.006, .055]
Age Group (contrast: 18 to 40)	
$<$ 3m	3.179 [2.358, 4.285]
3m to 3	1.279 [1.072, 1.525]
3 to 8	1.199 [.999, 1.439]
8 to 18	1.252 [1.077, 1.456]
40 to 55	1.697 [1.587, 1.816]
55 to 70	2.913 [2.714, 3.125]
$\geq$ 70	4.325 [4.002, 4.676]
Pod (contrast: BHED)	
A	.661 [.587, .744]
B	.561 [.498, .631]

Continued on next page

Table 5.4 – continued from previous page

	Prob(admit)/Prob(discharge)
C	4.381 [3.680, 5.217]
D	.216 [.190, .247]
Pediatrics	.397 [.339, .465]

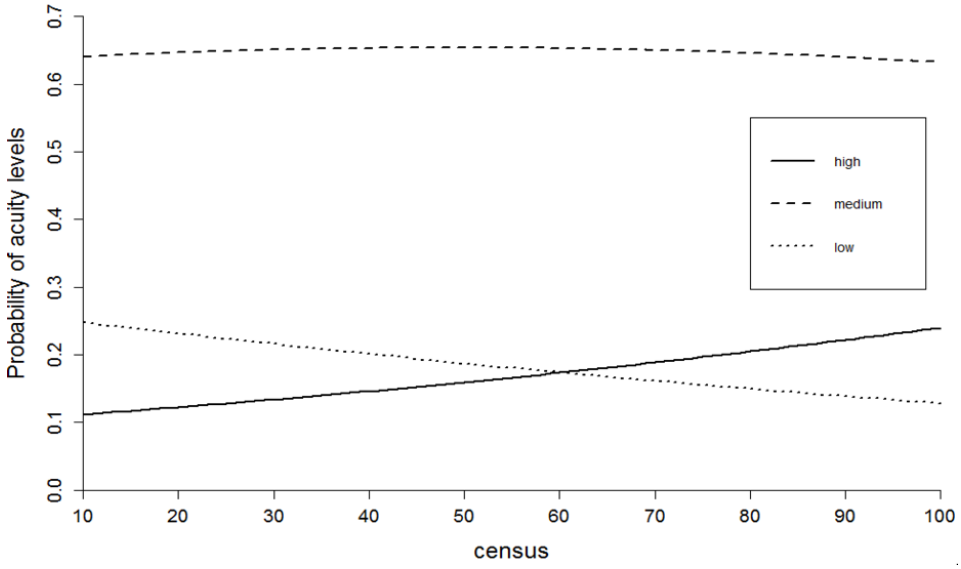

Figure 5.1: Marginal probabilities of different acuity levels versus census for a patient subgroup: Caucasian female, aged between 18 to 40, with abdominal pain.

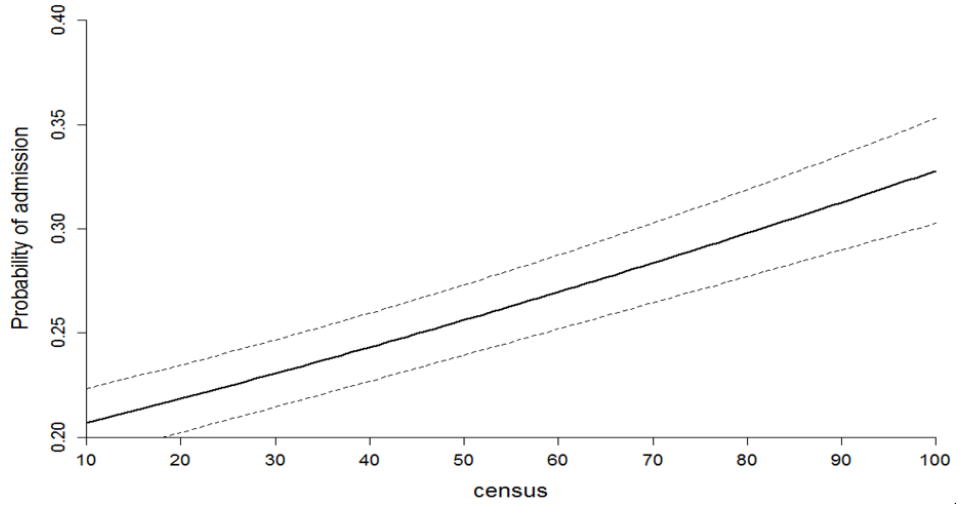

Figure 5.2: Probability of admission versus census (with 95% CI) for Caucasian female patients aged between 18 and 40, categorized as ESI3, presented with abdominal pain, and treated in Pod A.

5.4 Discusssion

To the best of our knowledge, there is only one other study that investigated the relationship between nurses' triage decision and ED census at the decision time and we are the first to consider census as a continuous variable (as opposed to a binary variable as in the prior work) and to use

a cumulative logit modeling to do so. In contrast to that previous study from (Richardson, 1998), we found a statistically significant association between ED census and nurses' triage decisions. Specifically, as can be seen from Figure 1, as census increases from 25 to 70 patients in the ED (representing, respectively, 10% and 90% quantiles of census from the data set), the probability of a patient being triaged as high acuity increases by about 50%, while the probability of a patient being triaged as low acuity decreases by approximately 25%. On the other hand, the probability of a patient being triaged as medium acuity (ESI 3) seems to change only slightly with census.

The relationship between physicians' admission decision and ED census at the decision time was observed in a prior work: (Gorski et al., 2017) performs a retrospective analysis using 18 months of all adult patient encounters seen in the main ED of an academic tertiary care center, and finds that there is a positive association between the likelihood that a patient would be admitted and the waiting room census and physician load census. Our results firmly support this earlier study in that we found a similar odds ratio for admission that increases as census does. From Figure 5.2, we can see that as census increases from 25 to 75 patients in the ED, the probability of a patient being categorized as admit increases by around 25%. Note that our study includes pediatric patients in addition to adults unlike (Gorski et al., 2017) that only considered adults and yet we still observed similar results.

Establishing an association does not prove cause and effect. Nevertheless, the correlations we found support what ED providers, nurses, and managers have suspected all along: As the ED becomes more crowded, there may be a tendency among providers and nurses to change their behavior in decision making towards being more risk averse. It may be that as the executive and cognitive function is taxed by the load, the clinicians of care make the decision that appears to be the safest choice for the individual patient. In the case of providers, they may opt for admission over a discharge in cases where the best disposition is in doubt. The same may hold true for triage nurses. As decisions become more pressured triage nurses may err on the side of caution and triage the patient a higher acuity than they otherwise would have. Work outside of health care has found similar decision fatigue in parole hearings (Danziger et al., 2011). Parole decisions made late in the day or long after a meal are more likely to result in the parolee staying in prison, the decision that is viewed as more cautious. As more and more decisions are made a decision maker tends to

pick what is considered the less risky of two choices even though this may not always be the best decision for the directly affected individual or others in the system.

5.5 Conclusion

This study includes data from a single academic center. The findings on relation between census and disposition are similar to a previous study at an academic center but it may be that academic centers have unique patient populations or organizational structures differing from community settings. Processing of admitted patients does tend to provide a greater challenge in academic centers (Horwitz et al., 2010). Also, our findings on relation between census and triage decisions should not be generalized to EDs that use a triage system other than ESI. Finally, a prospective case-control study would allow better identification of factors that affect nurses' triage and providers' admission decisions in the ED.

In this study, we found a correlation between overall ED census and likelihood of admission as well as changes in triage decisions that result in more patients being triaged to higher acuity levels. This supports a growing body of evidence that situational stressors such as high census may influence decisions made by nurses and physicians in the ED.

CHAPTER 6
Conclusion

This dissertation was centered around emergency department operations. Being an essential part of the US healthcare system, the majority of them are experiencing severe congestion issues. Motivated by some novel practices of streaming patient flow in the ED via re-engineering the service process, we studied the congestion problem through different angles. Chapter 2 and 3 targeted reducing crowding by seeking operational strategies that stream the patient flow in the ED, where the ED was modeled in stylized fashions. Chapter 4 employed a simulation model built for the UNC ED to evaluate the performances of several heuristic policies, which have structural forms inspired by the theoretical optimal policies found in previous two chapters. Chapter 5, on the other hand, offered a different perspective on the congestion problem. We were mainly interested in how ED providers' decision making would change according to fluctuations in congestion levels in the ED. Using two statistical models, we were able to get a sense of the potential impact of ED congestion level, as measured by census, on physicians' admission decision and nurses' triage decision making.

To be more specific, in Chapter 2, we modeled the ED as a single-server queueing system where there are two types of Poisson arrivals. Type-1 jobs only need a single stage of service (primary service). Type-2 jobs need an extra stage of service. One has the option to initiate secondary service, given the probability of a job being type-2 and the system state, at the time of arrival. For a job that is truly a type-2, doing so will shorten its sojourn time because by the time primary service is completed, secondary service for the job might have already been finished as well, or will be soon after. However, if we made a false judgement, then we waste the resources needed for starting secondary service in advance. With the purpose of balancing the tradeoff between the waiting cost and cost of false positives, we found the optimal policy to be of threshold type, where the threshold is a monotone decreasing function of both the system state, and the type-2

probability of the job for which we are making a decision. Although our model can fit into a few different settings in the ED, we have been focusing on its application to the boarding process. In particular, in our model, primary service corresponds to the lump sum service patients receive at the ED, while the secondary service corresponds to the boarding process. Additionally, type-2 probability corresponds to a patient's probability of being admitted to the hospital, and the server in our model represents the provider-bed pair in an ED. Given that EDs are mostly crowded, our single server assumption can be well justified under the heavy-traffic regime, despite the fact that they generally have multiple beds (servers).

Bearing in mind that the lengthy process of boarding serves as one of the main contributors to ED crowding, we continued to explore ways of shortening boarding via mathematical modeling and optimization. In Chapter 3, we modeled the patient flow in the ED as a deterministic fluid model. Instead of viewing the patients that arrive to an ED as discrete random arrivals, we treated them as continuous fluid flowing at a time-dependent rate. The ED, is a tank that receives the fluid that flows in and pumps it out at certain rate that is time-dependent, and controllable by the decision maker. Faster service rate is achieved, in a real ED context, by initiating hospital bed preparation for patients that we predict to be admit patients (early BeRT). As emphasized repeatedly, when implementing early BeRT in the ED one needs to carefully manage the tradeoff between the cost of overcrowding associated with holding patients in the ED and the cost of wasting hospital resources by making too many early BeRTs based on false admission prediction. In our fluid model, we managed to balance the tradeoff by imposing a constraint on the length of time during which one can make early BeRTs and thus speed up service. With our objective being to minimize the time averaged fluid level, which corresponds to the ED census, our fluid model then took the form of a constraint optimization problem. Based on this formulation, we identified the optimal period of time during each day to use the option of early BeRTs given the aforementioned operational constraint on the total amount of time that early BeRTs can be made. In cases when the arrival rate is larger than the maximum service rate that one can achieve, which is exactly the case during the ED's peak hours, we found that the optimal solution to the fluid model is to start early BeRT as early as possible, and for as long as possible, while keeping the operational constraint satisfied.

Following the lines of research in Chapter 2 and 3, we evaluated the performances of several heuristics that make early BeRTs based on different measurements of the system state, which

have structural forms inspired by the theoretical optimal policies found in previous chapters. We employed a simulation model built for the UNC ED and tested the performances of our heuristic policies on the simulation model in terms of the average length-of-stay and waiting time. We considered four sets of admission probability distributions, which vary in variances, as to evaluate the heuristics under different settings of prediction accuracy. Our results show that as the variance of distribution decreases, which corresponds to scenarios where admission prediction are more accurate, CFT (constraint fixed threshold) policy and CTT (census and time-dependent) policy began to dominate the other two heuristics that either uses a constant threshold (FT), or only time-dependent threshold (TT). This is to be expected, because the more information we take into account, the better we will do in shortening the sojourn times while keeping the daily counts of false positives unchanged.

Instead of focusing on resolving ED crowding, Chapter 5 offered another perspective and followed the stream of research that pays attention to operational responses to a congested ED such as the changes in rates of admission. As as far as we know, few have looked into the relationship between nurses' triage decision and ED census at the decision time. In Chapter 5, we are the first to apply a cumulative logit model to examine such an association. Our study shows that nurses tend to triage more patients into more critical categories when the ED is more congested. Further, as census increases from 0 to 100, the probability of a patient being triaged as high acuity doubles, while the probability of a patient being triaged as low acuity decreases by half. Some have already examined the relationship between physicians' admission decision and ED census at the decision time. By fitting a logistic regression model to the 2012 UNC ED data, we found result that firmly supports an earlier work (Gorski et al., 2017) that the odd ratio (OR) for admission is very similar in terms of magnitude and sign. Additionally, there is significant statistical evidence which suggests that as ED census increases, physicians tend to categorize more patients as admit patients. Further, our study is the first to include pediatric in addition to adult patients.

APPENDIX: PROOF OF THEOREMS, LEMMA, AND SUPPLEMENTARY TABLES A

A.1 Proof of Lemma 1

For the first part of Condition 1, we first write

$$\begin{aligned}
&\min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - \sum_j q_j V_m(\alpha_j, n-1) \\
&\qquad = \left[\min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2,n)\right] \\
&\qquad\qquad + \left[V_m(2,n) - \sum_j q_j V_m(\alpha_j, n-1)\right].
\end{aligned}$$

Then, the first part of the condition immediately follows by noting that the right hand side is a non-decreasing function of n for all $\alpha_i \in \Omega$ and $n \geq 1$ using 3) and 1) for V_m.

To show the second part of Condition 1, let $N_m(\alpha_i) = \inf\{n : V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s > V_m(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}$. First suppose that $N_m(\alpha_i) = 1$. Then

$$\begin{aligned}
&\min\{V_m(\alpha_i, 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - \sum_j q_j V_m(\alpha_j, 0) \\
&\qquad = \left[V_m(2,1) - \sum_j q_j V_m(\alpha_j, 0)\right] + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p \\
&\qquad\qquad \geq \frac{\gamma_1+\gamma_2}{\gamma_2}C_p \geq \frac{\alpha_i(\gamma_1+\gamma_2)}{\gamma_2}C_s,
\end{aligned}$$

where for the first inequality, we used 1) for V_m. Now, suppose that $N_m(\alpha_i) \geq 2$. Then

$$\begin{aligned}
&\min\{V_m(\alpha_i, 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - \sum_j q_j V_m(\alpha_j, 0) \\
&\qquad = \left[V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0)\right] + \frac{\alpha_i\gamma_1}{\gamma_2}C_s \\
&\qquad\qquad \geq \alpha_i C_s + \frac{\alpha_i\gamma_1}{\gamma_2}C_s = \frac{\alpha_i(\gamma_1+\gamma_2)}{\gamma_2}C_s,
\end{aligned}$$

where the inequality follows from 2) for V_m. This completes the proof for the condition.

A.2 Proof of Lemma 2

Proof of Condition 2. The second part of the condition is immediate by noting that

$$V_m(\alpha_i, n) - \min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2}C_p\} \geq$$
$$V_m(\alpha_i, n) - \left[V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s\right] = -\frac{\alpha_i\gamma_1}{\gamma_2}C_s.$$

To prove the first part of the condition, we need to show that for all i and $n \geq 1$

$$V_m(\alpha_i, n+1) - \min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1 + \gamma_2}{\gamma_2}C_p\}$$
$$\geq V_m(\alpha_i, n) - \min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2}C_p\}.$$

Now, let $N_m(\alpha_i) = \inf\{n : V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s > V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}$. Then, since $V_m(\alpha_i, n) - V_m(2, n)$ is a non-decreasing function of n for all i and $n \geq 1$ we have that $V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s > V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p$ if and only if $n \geq N_m(\alpha_i)$. Consequently, for $1 \leq n \leq N_m(\alpha_i) - 2$ we have

$$\left[V_m(\alpha_i, n+1) - \min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1 + \gamma_2}{\gamma_2}C_p\}\right]$$
$$- \left[V_m(\alpha_i, n) - \min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2}C_p\}\right] =$$
$$\left[V_m(\alpha_i, n+1) - V_m(\alpha_i, n+1) - \frac{\alpha_i\gamma_1}{\gamma_2}C_s\right]$$
$$- \left[V_m(\alpha_i, n) - V_m(\alpha_i, n) - \frac{\alpha_i\gamma_1}{\gamma_2}C_s\right] = 0 \geq 0.$$

For $n \geq N_m(\alpha_i)$ we have

$$\begin{aligned}
&\left[V_m(\alpha_i, n+1) - \min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\right] \\
&\qquad - \left[V_m(\alpha_i, n) - \min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\right] = \\
&\qquad\qquad \left[V_m(\alpha_i, n+1) - V_m(2, n+1) - \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\right] \\
&\qquad\qquad - \left[V_m(\alpha_i, n) - V_m(2, n) - \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\right] \\
&\qquad\qquad\qquad = [V_m(\alpha_i, n+1) - V_m(2, n+1)] - [V_m(\alpha_i, n) - V_m(2, n)] \geq 0,
\end{aligned}$$

because $V_m(\alpha_i, n) - V_m(2, n)$ is a non-decreasing function of n by assumption. For $n = N_m(\alpha_i) - 1$ we have

$$\begin{aligned}
&\left[V_m(\alpha_i, N_m(\alpha_i)) - \min\{V_m(\alpha_i, N_m(\alpha_i)) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, N_m(\alpha_i)) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\right] \\
&\qquad\qquad - [V_m(\alpha_i, N_m(\alpha_i) - 1) - \\
&\qquad \min\{V_m(\alpha_i, N_m(\alpha_i) - 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, N_m(\alpha_i) - 1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\Big] \\
&\qquad\qquad = \left[V_m(\alpha_i, N_m(\alpha_i)) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s\right] - \left[V_m(2, N_m(\alpha_i)) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\right] \geq 0,
\end{aligned}$$

by definition of $N_m(\alpha_i)$. Thus we have proved Condition 2.

Proof of Condition 3. We need to show that for all $n \geq 1$

$$\begin{aligned}
&\min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n+1) \geq \\
&\qquad\qquad \min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n).
\end{aligned}$$

Again, let $N_m(\alpha_i)$ be as defined previously in the proof of Condition 2. For $1 \leq n \leq N_m(\alpha_i) - 2$ we have

$$\begin{aligned}
&\left[\min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n+1)\right] \\
&\qquad - \left[\min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n)\right] \\
&\quad = \left[V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s - V_m(2, n+1)\right] - \left[V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s - V_m(2, n)\right] \\
&\qquad = \left[V_m(\alpha_i, n+1) - V_m(2, n+1)\right] - \left[V_m(\alpha_i, n) - V_m(2, n)\right],
\end{aligned}$$

because by assumption, $V_m(\alpha_i, n) - V_m(2, n)$ is non-decreasing with respect to n. For $n \geq N_m(\alpha_i)$ we have

$$\begin{aligned}
&\left[\min\{V_m(\alpha_i, n+1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n+1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n+1)\right] \\
&\qquad - \left[\min\{V_m(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, n)\right] \\
&\qquad = \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\gamma_1+\gamma_2}{\gamma_2}C_p = 0 \geq 0.
\end{aligned}$$

When $n = N_m(\alpha_i) - 1$ we have

$$\begin{aligned}
&\left[\min\{V_m(\alpha_i, N_m(\alpha_i)) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, N_m(\alpha_i)) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - V_m(2, N_m(\alpha_i))\right] \\
&\qquad - \Big[\min\{V_m(\alpha_i, N_m(\alpha_i) - 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_m(2, N_m(\alpha_i) - 1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} \\
&\qquad\qquad - V_m(2, N_m(\alpha_i) - 1)\Big] \\
&\qquad = \left[\frac{\gamma_1+\gamma_2}{\gamma_2}C_p + V_m(2, N_m(\alpha_i) - 1)\right] - \left[V_m(\alpha_i, N_m(\alpha_i) - 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s\right] \geq 0,
\end{aligned}$$

by definition of $N_m(\alpha_i)$. Thus we've proved Condition 3.

A.3 Proof of Lemma 3

To show each of the properties we use induction in a similar manner. Let $m \geq 1$ and suppose that condition 4 through 9 as given in the statement of the lemma all hold for $0 \leq k \leq m-1$. We will show that the same conditions then also hold for $k = m$.

Proof of Condition 4: For $n \geq 1$, we have

$$
\begin{aligned}
&V_m(\alpha_i, n) - V_m(2, n) = \\
&\quad \lambda \left[V_{m-1}(\alpha_i, n+1) - V_{m-1}(2, n+1)\right] + \alpha_i \gamma_1 \left[V_{m-1}(3, n) - \sum_j q_j V_{m-1}(\alpha_j, n-1)\right] \\
&\qquad + \gamma_2 \left[\min\{V_{m-1}(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\} - V_{m-1}(2, n)\right].
\end{aligned}
$$

From Conditions 4, 5 and 3 (with $k = m-1$), we know that the right hand side of the above equation is non-decreasing in n for $n \geq 1$ and thus we can conclude that Condition 4 also holds for $k = m$.

Proof of Condition 5: For $n \geq 2$, we have

$$
\begin{aligned}
V_m(3, n) - \sum_j q_j V_m(\alpha_j, n-1) = C_w + \lambda &\left[V_{m-1}(3, n+1) - \sum_j q_j V_{m-1}(\alpha_j, n)\right] \\
&+ \gamma_1 \left[V_{m-1}(3, n) - \sum_j q_j V_{m-1}(\alpha_j, n-1)\right] \\
+ \gamma_1 \sum_j q_j &\left[V_{m-1}(\alpha_j, n-1) - (1-\alpha_j) \sum_k q_k V_{m-1}(\alpha_k, n-2) - \alpha_j V_{m-1}(3, n-1)\right] \\
&+ \gamma_2 \sum_j q_j \left[V_{m-1}(\alpha_j, n-1) - \right. \\
&\left. \min\{V_{m-1}(\alpha_j, n-1) + \frac{\alpha_j \gamma_1}{\gamma_2} C_s, V_{m-1}(2, n-1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\}\right].
\end{aligned}
$$

From Conditions 5, 6 and 2 (with $k = m-1$), we know that the right hand side of the above equation is non-decreasing in n for $n \geq 2$ and thus we can conclude that Condition 5 also holds for $k = m$ but when $n \geq 2$. To establish the conditions for the case of $n = 1$, we need to show that $V_m(3, 1) - \sum_j q_j V_m(\alpha_j, 0) \geq 0$ and $V_m(3, 2) - \sum_j q_j V_m(\alpha_j, 0)$. To establish the first inequality, we

can write

$$
\begin{aligned}
V_m(3,1) - \sum_j q_j V_m(\alpha_j, 0) \\
&= C_w + \lambda \left[V_{m-1}(3,2) - \sum_j q_j V_{m-1}(\alpha_j, 1) \right] + \gamma_1 \left[V_{m-1}(3,1) - \sum_j q_j V_{m-1}(\alpha_j, 0) \right],
\end{aligned}
$$

which is non-negative by Condition 5 (with $k = m-1$). To establish the second inequality, we can write

$$
\begin{aligned}
&\left[V_m(3,2) - \sum_j q_j V_m(\alpha_j, 1) \right] - \left[V_m(3,1) - \sum_j q_j V_m(\alpha_j, 0) \right] = \\
&\qquad \lambda \left\{ \left[V_{m-1}(3,3) - \sum_j q_j V(\alpha_j, 2) \right] - \left[V_{m-1}(3,2) - \sum_j q_j V(\alpha_j, 1) \right] \right\} \\
&\quad + \gamma_1 \left\{ \left[V_{m-1}(3,2) - \sum_j q_j V_{m-1}(\alpha_j, 1) \right] - \left[V_{m-1}(3,1) - \sum_j q_j V_{m-1}(\alpha_j, 0) \right] \right\} \\
&\qquad + \gamma_1 \sum_j q_j \left[V_{m-1}(\alpha_j, 1) - (1-\alpha_j) \sum_k q_k V_{m-1}(\alpha_k, 0) - \alpha_j V_{m-1}(3,1) \right] \\
&\quad + \gamma_2 \sum_j q_j \left[V_{m-1}(\alpha_j, 1) - \min\{ V_{m-1}(\alpha_j, 1) + \frac{\alpha_j \gamma_1}{\gamma_2} C_s, V_{m-1}(2,1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p \} \right] \geq \\
&\qquad \gamma_1 \sum_j q_j \left[V_{m-1}(\alpha_j, 1) - (1-\alpha_j) \sum_k q_k V_{m-1}(\alpha_k, 0) - \alpha_j V_{m-1}(3,1) \right] \\
&\quad + \gamma_2 \sum_j q_j \left[V_{m-1}(\alpha_j, 1) - \min\{ V_{m-1}(\alpha_j, 1) + \frac{\alpha_j \gamma_1}{\gamma_2} C_s, V_{m-1}(2,1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p \} \right] \geq \\
&\qquad \gamma_1 \sum_j q_j \alpha_j C_s - \gamma_2 \sum_j q_j \frac{\alpha_j \gamma_1}{\gamma_2} C_s = 0,
\end{aligned}
$$

where we used the fact that Conditions 5, 6 and 2 hold for $k = m-1$. Hence, Condition 5 also holds for $k = m$.

Proof of Condition 6: For $n \geq 2$, we have

$$
\begin{aligned}
V_m(\alpha_i, n) - (1-\alpha_i)\sum_j q_j V_m(\alpha_j, n-1) - \alpha_i V_m(3,n) &= (1-\alpha_i)C_w \\
&+ \lambda\left[V_{m-1}(\alpha_i, n+1) - (1-\alpha_i)\sum_j q_j V_{m-1}(\alpha_j, n) - \alpha_i V_{m-1}(3, n+1)\right] \\
&+ \gamma_2\left[\min\{V_{m-1}(\alpha_i, n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_{m-1}(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\} - \sum_j q_j V_{m-1}(\alpha_j, n-1)\right] \\
&+ (1-\alpha_i)\gamma_1\sum_j q_j\,[V_{m-1}(\alpha_j, n-1) \\
&-(1-\alpha_j)\sum_k q_k V_{m-1}(\alpha_k, n-2) - \alpha_j V_{m-1}(3, n-1)\Big] \\
&+ (1-\alpha_i)\gamma_2\sum_j q_j\,[V_{m-1}(\alpha_j, n-1) - \\
&\min\{V_{m-1}(\alpha_j, n-1) + \frac{\alpha_j\gamma_1}{\gamma_2}C_s, V_{m-1}(2, n-1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\Big].
\end{aligned}
$$

From Conditions 6, 1 and 2 (with $k = m-1$), we know that the right hand side of the above equation is non-decreasing in n for $n \geq 2$ and thus we can conclude that the first part of Condition 6 also holds for $k = m$ but when $n \geq 2$. To complete the proof for Condition 6, it then remains to show that $V_m(\alpha_i, 1) - (1-\alpha_i)\sum_j q_j V_m(\alpha_j, 0) - \alpha_i V_m(3,1) \geq \alpha_i C_s$ and $V_m(\alpha_i, 2) - (1-\alpha_i)\sum_j q_j V_m(\alpha_j, 1) - \alpha_i V_m(3,2) \geq V_m(\alpha_i, 1) - (1-\alpha_i)\sum_j q_j V_m(\alpha_j, 0) - \alpha_i V_m(3,1)$.

To establish the first inequality, first, from (2.6) (with $n = 1$), we have

$$
\begin{aligned}
V_m(\alpha_i, 1) = C_w + \lambda V_{m-1}(\alpha_i, 2) + (1-\alpha_i)\gamma_1\sum_j q_j V_{m-1}(\alpha_j, 0) \\
+ \alpha_i\gamma_1 V_{m-1}(3,1) + \gamma_2\min\{V_{m-1}(\alpha_i, 1) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_{m-1}(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2}C_p\},
\end{aligned}
$$

and from (2.5), we have

$$
\begin{aligned}
(1-\alpha_i)\sum_j q_j V_m(\alpha_j, 0) &= (1-\alpha_i)V_m(0) \\
&= \lambda(1-\alpha_i)\sum_j q_j V_{m-1}(\alpha_j, 1) + (\gamma_1+\gamma_2)(1-\alpha_i)\sum_j q_j V_{m-1}(\alpha_j, 0).
\end{aligned}
$$

Using (2.8) (with $n = 1$) we can write

$$\alpha_i V_m(3,1) = \alpha_i C_w + \lambda \alpha_i V_{m-1}(3,2) + \gamma_2 \alpha_i \sum_j q_j V_{m-1}(\alpha_j, 0) + \gamma_1 \alpha_i V_{m-1}(3,1),$$

where we used the fact that $\sum_j q_j V_m(\alpha_j, 0) = V_m(0)$ for all $m \geq 0$. It then follows that

$$\begin{aligned} V_m(\alpha_i, 1) - (1-\alpha_i) \sum_j q_j V_m(\alpha_j, 0) - \alpha_i V_m(3,1) &= (1-\alpha_i) C_w \\ &+ \lambda \left[V_{m-1}(\alpha_i, 2) - (1-\alpha_i) \sum_j q_j V_{m-1}(\alpha_j, 1) - \alpha_i V_{m-1}(3,2) \right] \\ &+ \gamma_2 \left[\min\{V_{m-1}(\alpha_i, 1) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2,1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\} - \sum_j q_j V_{m-1}(\alpha_j, 0) \right]. \end{aligned}$$

Then, using Condition 6 and 1 (with $k = m - 1$), we have

$$V_m(\alpha_i, 1) - (1-\alpha_i) \sum_j q_j V_m(\alpha_j, 0) - \alpha_i V_m(3,1) \geq \lambda \alpha_i C_s + \gamma_2 \cdot \frac{\gamma_1 + \gamma_2}{\gamma_2} \cdot \alpha_i C_s = \alpha_i C_s.$$

For the second inequality, we can write

$$
\begin{aligned}
&\left[V_m(\alpha_i,2)-(1-\alpha_i)\sum_j q_j V_m(\alpha_j,1)-\alpha_i V_m(3,2)\right] \\
&\qquad -\left[V_m(\alpha_i,1)-(1-\alpha_i)\sum_j q_j V_m(\alpha_j,0)-\alpha_i V_m(3,1)\right]= \\
&\qquad \lambda\left\{\left[V_{m-1}(\alpha_i,3)-(1-\alpha_i)\sum_j q_j V_{m-1}(\alpha_j,2)-\alpha_i V_{m-1}(3,3)\right]\right. \\
&\qquad \left.-\left[V_{m-1}(\alpha_i,2)-(1-\alpha_i)\sum_j q_j V_{m-1}(\alpha_j,1)-\alpha_i V_{m-1}(3,2)\right]\right\} \\
&\quad +\gamma_2\left\{\left[\min\{V_{m-1}(\alpha_i,2)+\frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_{m-1}(2,2)+\frac{\gamma_1+\gamma_2}{\gamma_1}C_p\}-\sum_j q_j V_{m-1}(\alpha_j,1)\right]\right. \\
&\quad \left.-\left[\min\{V_{m-1}(\alpha_i,1)+\frac{\alpha_i\gamma_1}{\gamma_2}C_s, V_{m-1}(2,1)+\frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}-\sum_j q_j V_{m-1}(\alpha_j,0)\right]\right\} \\
&\quad +(1-\alpha_i)\gamma_1\sum_j q_j\left[V_{m-1}(\alpha_j,1)-(1-\alpha_j)\sum_k q_k V_{m-1}(\alpha_k,0)-\alpha_j V_{m-1}(3,1)\right]+ \\
&(1-\alpha_i)\gamma_2\sum_j q_j\left[V_{m-1}(\alpha_j,1)-\min\{V_{m-1}(\alpha_j,1)+\frac{\alpha_j\gamma_1}{\gamma_2}C_s, V_{m-1}(2,1)+\frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\right] \\
&\quad \geq (1-\alpha_i)\gamma_1\sum_j q_j\left[V_{m-1}(\alpha_j,1)-(1-\alpha_j)\sum_k q_k V_{m-1}(\alpha_k,0)-\alpha_j V_{m-1}(3,1)\right]+ \\
&(1-\alpha_i)\gamma_2\sum_j q_j\left[V_{m-1}(\alpha_j,1)-\min\{V_{m-1}(\alpha_j,1)+\frac{\alpha_j\gamma_1}{\gamma_2}C_s, V_{m-1}(2,1)+\frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}\right] \\
&\qquad \geq (1-\alpha_i)\left[\gamma_1\sum_j q_j\alpha_j C_s-\gamma_2\sum_j q_j\frac{\alpha_j\gamma_1}{\gamma_2}C_s\right]=0,
\end{aligned}
$$

where we used Condition 6, 1 and 2 for $k=m-1$. Thus Condition 6 for $k=m$ follows.

Proof of Condition 7: For $n \geq 2$, we have

$$
\begin{aligned}
V_m(2,n) - \sum_j q_j V_m(\alpha_j, n-1) = C_w + \lambda \left[V_{m-1}(2,n+1) - \sum_j q_j V_{m-1}(\alpha_j, n) \right] \\
+ \gamma_2 \left[V_{m-1}(2,n) - \sum_j q_j V_{m-1}(\alpha_j, n-1) \right] \\
+ \gamma_1 \sum_j q_j \left[V_{m-1}(\alpha_j, n-1) - (1-\alpha_j) \sum_k q_k V_{m-1}(\alpha_k, n-2) - \alpha_j V_{m-1}(3, n-1) \right] \\
+ \gamma_2 \sum_j q_j \left[V_{m-1}(\alpha_j, n-1) \right. \\
\left. - \min\{ V_{m-1}(\alpha_j, n-1) + \frac{\alpha_j \gamma_1}{\gamma_2} C_s, V_{m-1}(2, n-1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p \} \right].
\end{aligned}
$$

Then, using Condition 7, 6 and 2 (with $k = m-1$), we can conclude that the right hand side of the above equation is non-decreasing in n for $n \geq 2$ and thus $V_m(2,n) - \sum_j q_j V_m(\alpha_j, n-1)$ is also non-decreasing in n for $n \geq 2$. To complete the proof for Condition 7, we need to show that $V_m(2,1) - \sum_j q_j V_m(\alpha_j, 0) \geq 0$ and $V_m(2,2) - \sum_j q_j V_m(\alpha_j, 1) \geq V_m(2,1) - \sum_j q_j V_m(\alpha_j, 0)$.

To establish the first inequality, first, using (2.7) for $n = 1$, we can write

$$
V_m(2,1) = C_w + \lambda V_{m-1}(2,2) + \gamma_1 \sum_j q_j V_{m-1}(\alpha_j, 0) + \gamma_2 V_{m-1}(2,1),
$$

and using (2.5), we can write

$$
\sum_j q_j V_m(\alpha_j, 0) = V_m(0) = \lambda \sum_j q_j V_{m-1}(\alpha_j, 1) + (\gamma_1 + \gamma_2) \sum_j q_j V_{m-1}(\alpha_j, 0),
$$

where we used the fact that $\sum_j q_j V_m(\alpha_j, 0) = V_m(0)$. Thus, we have

$$
\begin{aligned}
V_m(2,1) - \sum_j q_j V_m(\alpha_j, 0) = C_w + \lambda \left[V_{m-1}(2,2) - \sum_j q_j V_{m-1}(\alpha_j, 1) \right] \\
+ \gamma_2 \left[V_{m-1}(2,1) - \sum_j q_j V_{m-1}(\alpha_j, 0) \right].
\end{aligned}
$$

From Condition 7 (with $k = m-1$), we can then see that $V_m(2,1) - \sum_j q_j V_m(\alpha_j, 0) \geq 0$. To establish the second inequality, first we can write

$$\begin{aligned}
&\left[V_m(2,2) - \sum_j q_j V_m(\alpha_j,1)\right] - \left[V_m(2,1) - \sum_j q_j V_m(\alpha_j,0)\right] = \\
&\qquad \lambda \left\{ \left[V_{m-1}(2,3) - \sum_j q_j V_{m-1}(\alpha_j,2)\right] - \left[V_{m-1}(2,2) - \sum_j q_j V_{m-1}(\alpha_j,1)\right] \right\} \\
&\qquad + \gamma_2 \left\{ \left[V_{m-1}(2,2) - \sum_j q_j V_{m-1}(\alpha_j,1)\right] - \left[V_{m-1}(2,1) - \sum_j q_j V_{m-1}(\alpha_j,0)\right] \right\} \\
&\qquad + \gamma_1 \sum_j q_j \left[V_{m-1}(\alpha_j,1) - (1-\alpha_j)\sum_j q_j V_{m-1}(\alpha_j,0) - \alpha_j V_{m-1}(3,1)\right] \\
&\qquad + \gamma_2 \sum_j q_j \left[V_{m-1}(\alpha_j,1) - \min\{V_{m-1}(\alpha_j,1) + \frac{\alpha_j\gamma_1 C_s}{\gamma_2}, V_{m-1}(2,1) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p\}\right] \geq \\
&\qquad \gamma_1 \sum_j q_j \left[V_{m-1}(\alpha_j,1) - (1-\alpha_j)\sum_j q_j V_{m-1}(\alpha_j,0) - \alpha_j V_{m-1}(3,1)\right] - \gamma_2 \sum_j q_j \frac{\alpha_j\gamma_1 C_s}{\gamma_2} \\
&\qquad\qquad \geq \gamma_1 \sum_j q_j \alpha_j C_s - \sum_j q_j \alpha_j \gamma_1 C_s = 0,
\end{aligned}$$

where we used Conditions 7, 6 and 2, which we know holds for $k = m-1$. Thus Condition 7 for $k = m$ follows.

Proof of Condition 8: First, we can write

$$\begin{aligned}
&\left[V_m(\alpha_i, n+1) - \sum_j q_j V_m(\alpha_j, n)\right] - \left[V_m(\alpha_i, n) - \sum_j q_j V_m(\alpha_j, n-1)\right] = \\
&\qquad \{[V_m(\alpha_i, n+1) - V_m(2, n+1)] - [V_m(\alpha_i, n) - V_m(2, n)]\} \\
&\qquad + \left\{ \left[V_m(2, n+1) - \sum_j q_j V_m(\alpha_j, n)\right] - \left[V_m(2, n) - \sum_j q_j V_m(\alpha_j, n-1)\right] \right\}.
\end{aligned}$$

Then, using Conditions 4 and 7 (with $k = m$), which we have already established, we can conclude that for all $n \geq 1$, $V_m(\alpha_i, n+1) - \sum_j q_j V_m(\alpha_j, n) \geq V_m(\alpha_i, n) - \sum_j q_j V_m(\alpha_j, n-1)$. It then remains to show that $V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0) \geq \alpha_i C_s$. We are left to show $V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0) \geq$

$\alpha_i C_r$. Using (2.6) (with $n = 1$), we have

$$V_m(\alpha_i, 1) = C_w + \lambda V_{m-1}(\alpha_i, 2) + (1 - \alpha_i)\gamma_1 \sum_j q_j V_{m-1}(\alpha_i, 0)$$
$$+\alpha_i \gamma_1 V_{m-1}(3, 1) + \gamma_2 \min\{V_{m-1}(\alpha_i, 1) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2, 1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\},$$

and using (2.5), we have

$$\sum_j q_j V_m(\alpha_j, 0) = V_m(0) = \lambda \sum_j q_j V_{m-1}(\alpha_j, 1) + (\gamma_1 + \gamma_2) \sum_j q_j V_{m-1}(\alpha_j, 0),$$

where we used the fact that $\sum_j q_j V_m(\alpha_j, 0) = V_m(0)$. Then, we have

$$\begin{aligned} V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0) = C_w + \lambda \left[V_{m-1}(\alpha_i, 2) - \sum_j q_j V_{m-1}(\alpha_j, 1) \right] \\ + \alpha_i \gamma_1 \left[V_{m-1}(3, 1) - \sum_j q_j V_{m-1}(\alpha_j, 0) \right] \\ + \gamma_2 \left[\min\{V_{m-1}(\alpha_i, 1) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2, 1) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\} - \sum_j q_j V_{m-1}(\alpha_j, 0) \right]. \end{aligned}$$

Hence

$$\begin{aligned} V_m(\alpha_i, 1) - \sum_j q_j V_m(\alpha_j, 0) \geq \lambda \alpha_i C_s + 0 + \gamma_2 \cdot \frac{\alpha_i(\gamma_1 + \gamma_2)}{\gamma_2} C_s \\ = \lambda \alpha_i C_s + (\gamma_1 + \gamma_2)\alpha_i C_s = \alpha_i C_s, \end{aligned}$$

where we used Conditions 8, 5 and 1 for $k = m - 1$. Thus, Condition 8 for $k = m$ follows.

Proof of Condition 9: Using (2.7), we have, for $n \geq 1$

$$\begin{aligned} V_m(\alpha_i, n) = \lambda V_{m-1}(\alpha_i, n + 1) + \gamma_1 V_{m-1}(3, n) \\ + \alpha_i \gamma_1 \left[V_{m-1}(3, n) - \sum_j q_j V_{m-1}(\alpha_j, n - 1) \right] \\ + \gamma_2 \min\{V_{m-1}(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V_{m-1}(2, n) + \frac{\gamma_1 + \gamma_2}{\gamma_2} C_p\}, \end{aligned}$$

where the first term is non-decreasing in i using Condition 9 (with $k = m - 1$). The second term is invariant in i. The third term is increasing in i because α_i is increasing in i and $V_{m-1}(3, n) - \sum_j q_j V_{m-1}(\alpha_j, n-1)$ is non-negative from Condition 5 (with $k = m - 1$). The last term is also non-decreasing in i from the fact that minimization preserves monotonicity and using Condition 9 (with $k = m - 1$). Then, we can conclude that $V_m(\alpha_i, n)$ is non-decreasing in i for all $n \geq 1$. This completes the proof of the lemma.

A.4 Proof of Lemma 4

We show this via two steps. First, we argue that for two policies which start applying $\overline{\gamma}$ at the same time, the one that applies for a longer interval is superior than the other. Second, we argue that it is suboptimal to start applying $\overline{\gamma}$ after $T - \overline{\delta}$. Let $\delta = t_s - t_0$, i.e., δ represents the total length of time one BeRTs.

1. Consider two policies π and γ with $t_0^\pi = t_0^\gamma = t_0$ and $\delta^\pi < \delta^\gamma$. We are going to show that $A^\pi \geq A^\gamma$. First, since $t_0^\pi = t_0^\gamma$ and $\delta^\pi < \delta^\gamma$, $x^\pi(t) = x^\gamma(t)$ for $0 \leq t \leq t_0 + \delta^\pi$. And thus $\int_0^{t_0+\delta^\pi} x^\pi(t)dt = \int_0^{t_0+\delta^\pi} x^\gamma(t)dt$. Hence

$$\begin{aligned} A^\pi - A^\gamma &= \int_{t_0+\delta^\pi}^{T} x^\pi(t)dt - \int_{t_0+\delta^\pi}^{T} x^\gamma(t)dt. \\ &= \left[\int_{t_0+\delta^\pi}^{t_0+\delta^\gamma} \left(x^\pi(t)dt - x^\gamma(t)\right) dt\right] + \left[\int_{t_0+\delta^\gamma}^{T} \left(x^\pi(t)dt - x^\gamma(t)\right) dt\right] \\ &:= \Delta_1 + \Delta_2. \end{aligned}$$

We evaluate the positivity of Δ_1 and Δ_2 respectively next. For $t_0 + \delta^\pi \leq t \leq t_0 + \delta^\gamma$, we have

$$x^\pi(t) = \max\{0, x(t_0 + \delta^\pi) + (\lambda - s\underline{\gamma})(t - t_0)\},$$

$$x^\gamma(t) = \max\{0, x(t_0 + \delta^\pi) + (\lambda - s\overline{\gamma})(t - t_0)\}.$$

Hence $x^\pi(t) \geq x^\gamma(t)$. And thus $\Delta_1 \geq 0$. For $t_0 + \delta^\gamma \leq t \leq T$, we have

$$x^\pi(t) = \max\{0, x^\pi(t_0 + \delta^\gamma) + (\lambda - s\underline{\gamma})(t - t_0)\},$$

$$x^\gamma(t) = \max\{0, x^\gamma(t_0 + \delta^\gamma) + (\lambda - s\underline{\gamma})(t - t_0)\}.$$

Since $x^\pi(t_0 + \delta^\gamma) \geq x^\gamma(t_0 + \delta^\gamma)$ we have $x^\pi(t) \geq x^\gamma(t)$. Thus $\Delta_2 \geq 0$. Consequently, we have $A^\pi \geq A^\gamma$. Thus we have completed our first step. A natural consequence of step 1 is that it is optimal to let $t_s = t_0 + \overline{\delta}$. Hence our decision variable reduces to t_0 alone.

2. It is suboptimal to let $t_0 \geq T - \overline{\delta}$, i.e., to start after $T - \overline{\delta}$.

Consider two policies π and γ where $t_0^\pi = T - \overline{\delta}$ and $T - \overline{\delta} < t_0^\gamma < T$. Based on step 1. we can assume that $\delta^\pi = \overline{\delta}$ and $\delta^\gamma = T - t_0^\gamma < \overline{\delta}$. We are going to show that $A^\pi \leq A^\gamma$. First since $t_0^\pi < t_0^\gamma$, $x^\pi(t) = x^\gamma(t)$ for $0 \leq t \leq t_0^\pi$. And thus $\int_0^{t_0^\pi} x^\pi(t)dt = \int_0^{t_0^\pi} x^\gamma(t)dt$. We have

$$\begin{aligned} A^\pi - A^\gamma &= \int_{t_0^\pi}^T x^\pi(t)dt - \int_{t_0^\pi}^T x^\gamma(t)dt \\ &= \int_{t_0^\pi}^{t_0^\gamma} \left[x^\pi(t) - x^\gamma(t)\right] dt + \int_{t_0^\gamma}^T \left[x^\pi(t) - x^\gamma(t)\right] dt \\ &:= \Delta_1 + \Delta_2. \end{aligned}$$

Next we evaluate the positivity of Δ_1 and Δ_2 respectively. For $t_0^\pi \leq t \leq t_0^\gamma$ we have

$$x^\pi(t) = \max\{0, x(t_0^\pi) + (\lambda - s\overline{\gamma})(t - t_0^\pi)\},$$

$$x^\gamma(t) = \max\{0, x(t_0^\pi) + (\lambda - s\underline{\gamma})(t - t_0^\pi)\}.$$

Hence $x^\pi(t) \leq x^\gamma(t)$, and $\Delta_1 \leq 0$. For $t_0^\gamma \leq t \leq T$ we have

$$x^\pi(t) = \max\{0, x^\pi(t_0^\gamma) + (\lambda - s\overline{\gamma})(t - t_0^\gamma)\},$$

$$x^\gamma(t) = \max\{0, x^\gamma(t_0^\gamma) + (\lambda - s\overline{\gamma})(t - t_0^\gamma)\}.$$

Since $x^\pi(t_0^\gamma) \leq x^\gamma(t_0^\gamma)$ we have $x^\pi(t) \leq x^\gamma(t)$, thus $\Delta_2 \leq 0$. Thus we have completely step 2.

Combining the above two results we can easily see that it is suboptimal to start after $T - \overline{\delta}$. And as long as one starts before $T - \overline{\delta}$ the maximum length of time one can apply $\overline{\gamma}$ is $\overline{\delta}$, and by step 1 we know that it is suboptimal to let $\delta < \overline{\delta}$, or $t_s - t_0 < \overline{\delta}$. This completes the proof of the lemma.

A.5 Proof of Theorem 4

Case 1 $s\underline{\gamma} < s\overline{\gamma} < \lambda$

For any $0 \leq t_0 \leq T - \overline{\delta}$, we have

$$A = \int_0^{t_0} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_0+\overline{\delta}} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt$$
$$+ \int_{t_0+\overline{\delta}}^{T} \left[x(t_0 + \overline{\delta}) + (\lambda - s\underline{\gamma})(t - t_0 - \overline{\delta})\right] dt,$$

where

$$x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0,$$

$$x(t_0 + \overline{\delta}) = x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})\overline{\delta},$$

subsituiting into the expression for A we get

$$A = \frac{1}{2}\left\{-s(\underline{\gamma} - \overline{\gamma})\overline{\delta}(2t_0 + \overline{\delta}) + 2T(x_0 + s(\underline{\gamma} - \overline{\gamma})\overline{\delta}) + T^2(-s\underline{\gamma} + \lambda)\right\}$$
$$= 2s(\overline{\gamma} - \underline{\gamma})\overline{\delta}t_0 + 2T(x_0 + s(\underline{\gamma} - \overline{\gamma}) + \overline{\delta}) + \frac{1}{2}T^2(-s\underline{\gamma} + \lambda),$$

which is minimized at $t_0^* = 0$. Hence we have

$$t_0^* = 0.$$

Case 2 $\lambda < s\underline{\gamma} < s\overline{\gamma}$

Let t_1 be the time it takes to clear the system if one starts applying $\overline{\gamma}$ from time 0 and is allowed to apply as long as possible. Then t_1 satisfies

$$x_0 + \int_0^{t_1} (\lambda - s\overline{\gamma})dt = 0,$$

which gives us $t_1 = \frac{x_0}{s\overline{\gamma} - \lambda}$.

First, if $\overline{\delta} > t_1$ then whenever we start, i.e., whatever t_0 we choose, we will be able to clear out the system before $t_0 + \overline{\delta}$. Second, if $\overline{\delta} \leq t_1$, then there exists an interval $[0, t_2]$ with $t_2 > 0$ such

that for any $0 \le t_0 < t_2$ we will have $x(t_0+\overline{\delta}) > 0$, and for any $t_0 \ge t_2$ we will have $x(t_0+\overline{\delta}) = 0$. Hence t_2 satisfies

$$x_0 + \int_0^{t_2} (\lambda - s\underline{\gamma})dt + \int_{t_2}^{t_2+\overline{\delta}} (\lambda - s\overline{\gamma})dt = 0,$$

which gives us

$$t_2 = \frac{x_0 - (s\overline{\gamma} - \lambda)\overline{\delta}}{s\underline{\gamma} - \lambda}.$$

Lastly, let t_3 be the time it takes to clear the system if one uses $\underline{\gamma}$ at all times, then t_3 satisfies

$$x_0 + \int_0^{t_1} (\lambda - s\underline{\gamma})dt = 0,$$

which gives us $t_3 = \frac{x_0}{s\underline{\gamma}-\lambda}$. We consider three scenarios based on the definition of t_1 and t_2.

Case 2.1 $\overline{\delta} < t_1$ **and** $T < t_2 + \overline{\delta}$**, i.e.,** $\overline{\delta} < \min\{\frac{x_0}{s\overline{\gamma}-\lambda}, \frac{x_0}{s(\overline{\gamma}-\underline{\gamma})} - \frac{T(s\underline{\gamma}-\lambda)}{s(\overline{\gamma}-\underline{\gamma})}\}$**.**

In this situation, whatever t_0 one chooses, one will not be able to clear the system by T. Hence, for any $0 \le t_0 \le T - \overline{\delta}$, we have

$$A = \int_0^{t_0} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_0+\overline{\delta}} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt \\ + \int_{t_0+\overline{\delta}}^{T} \left[x(t_0+\overline{\delta}) + (\lambda - s\underline{\gamma})(t - t_0 - \overline{\delta})\right] dt,$$

where

$$x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0,$$

$$x(t_0+\overline{\delta}) = x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})\overline{\delta},$$

subsituiting into the expression for A we get

$$A = \frac{1}{2}\left\{-s(\underline{\gamma} - \overline{\gamma})\overline{\delta}(2t_0 + \overline{\delta}) + 2T(x_0 + s(\underline{\gamma} - \overline{\gamma})\overline{\delta}) + T^2(-s\underline{\gamma} + \lambda)\right\} \\ = 2s(\overline{\gamma} - \underline{\gamma})\overline{\delta}t_0 + \frac{1}{2}\left\{-s(\underline{\gamma} - \overline{\gamma})\overline{\delta}^2 + 2T(x_0 + s(\underline{\gamma} - \overline{\gamma})\overline{\delta}) + T^2(-s\underline{\gamma} + \lambda)\right\},$$

which is minimized at $t_0^* = 0$. Hence we have $t_0^* = 0$.

Case 2.2 $\overline{\delta} < t_1$ **and** $T > t_2 + \overline{\delta}$**, i.e.,** $\frac{x_0}{s(\overline{\gamma}-\underline{\gamma})} - \frac{T(s\underline{\gamma}-\lambda)}{s(\overline{\gamma}-\underline{\gamma})} < \overline{\delta} < \frac{x_0}{s\overline{\gamma}-\lambda}$**.**

This is the situation where if $t_0 < t_2$ then $x(t_0 + \overline{\delta}) > 0$ while if $t_0 \geq t_2$ the fluid will drop to zero before $t_0 + \overline{\delta}$ and stays there until T.

First we argue that it is suboptimal to let $t_0 > t_2$. Consider two policies π and γ. $t_0^\pi = t_2$ and $t_2 < t_0^\gamma \leq \min\{T - \overline{\delta}, t_3\}$ (It is obvious that one shall not choose t_0 to be later than t_3). And for both π and γ we have $\delta^\pi = \delta^\gamma = \overline{\delta}$. We are going to show that $A^\pi \leq A^\gamma$. Since $t_0^\gamma > t_0^\pi = t_2$, $x^\pi(t) = x^\gamma(t)$ for $0 \leq t \leq t_2$. Thus

$$A^\pi - A^\gamma = \int_{t_2}^{T} x^\pi(t)dt - \int_{t_2}^{T} x^\gamma(t)dt.$$

Let t_4 be the time when the fluid drops to zero under policy γ, then

$$x^\gamma(t_4) = x_0 + (\lambda - s\underline{\gamma})t_0^\gamma + (\lambda - s\overline{\gamma})(t_3 - t_0^\gamma) = 0,$$

which gives us

$$t_4 = \frac{x_0 + (s\overline{\gamma} - s\underline{\gamma})t_0^\gamma}{s\overline{\gamma} - \lambda},$$

and

$$\int_{t_2}^{T} x^\pi(t)dt = \int_{t_2}^{t_2+\overline{\delta}} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt,$$

$$\int_{t_2}^{T} x^\gamma(t)dt = \int_{t_2}^{t_0^\gamma} \left[x(t_2) + (\lambda - s\underline{\gamma})(t - t_2)\right] dt + \int_{t_0^\gamma}^{t_4} \left[x^\gamma(t_0^\gamma) + (\lambda - s\overline{\gamma})(t - t_0^\gamma)\right] dt.$$

Since by definition of t_2, $x^\pi(t) = 0$ for $t \geq t_2 + \overline{\delta}$, and by definition of t_4, $x^\gamma(t) = 0$ for $t \geq t_4$, we also have

$$x^\gamma(t_0^\gamma) = x(t_2) + (\lambda - s\underline{\gamma})(t_0^\gamma - t_2),$$

using this in $\int_{t_2}^{T} x^{\gamma}(t)dt$ we get

$$\begin{aligned}\int_{t_2}^{T} x^{\gamma}(t)dt =& \int_{t_2}^{t_0^{\gamma}} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt + \int_{t_2}^{t_0^{\gamma}} (s\overline{\gamma} - s\underline{\gamma})(t - t_2)dt \\ &+ \int_{t_0^{\gamma}}^{t_4} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt + \int_{t_0^{\gamma}}^{t_4} \left[(\lambda - s\underline{\gamma})(t_0^{\gamma} - t_2)\right] dt \\ &+ \int_{t_0^{\gamma}}^{t_4} \left[(\lambda - s\overline{\gamma})(t_2 - t_0^{\gamma})\right] dt \\ =& \left(\int_{t_2}^{t_4} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt\right) + \frac{(s\overline{\gamma} - s\underline{\gamma})(t_0^{\gamma} - t_2)^2}{2} \\ &+ (s\overline{\gamma} - s\underline{\gamma})(t_0^{\gamma} - t_2)(t_3 - t_0^{\gamma}) > \int_{t_2}^{t_4} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt,\end{aligned}$$

comparing this with the expression for $\int_{t_2}^{T} x^{\pi}(t)dt$, if we can show that $t_4 > t_2 + \overline{\delta}$ we are done. Now

$$\begin{aligned}t_4 =& \frac{x_0 + (s\overline{\gamma} - s\underline{\gamma})t_0^{\gamma}}{s\overline{\gamma} - \lambda} > \frac{x_0 + (s\overline{\gamma} - s\underline{\gamma})t_2}{s\overline{\gamma} - \lambda} = t_2 + \frac{x_0 + (\lambda - s\underline{\gamma})t_2}{s\overline{\gamma} - \lambda} \\ =& t_2 + \frac{1}{s\overline{\gamma} - \lambda}\left[x_0 + (\lambda - s\underline{\gamma})\frac{x_0 - (s\overline{\gamma} - \lambda)\overline{\delta}}{s\underline{\gamma} - \lambda}\right] \\ =& t_2 + \overline{\delta}.\end{aligned}$$

This completes the argument. Thus it is suboptimal to start after t_2.

Now consider a policy that chooses $0 \leq t_0 \leq t_2$, let t_5 be the time the system is cleared, then t_5 satisfies

$$x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})\overline{\delta} + (\lambda - s\underline{\gamma})(t_5 - t_0 - \overline{\delta}) = 0,$$

which gives us

$$t_5 = \overline{\delta} + \frac{x_0 - \overline{\delta}(s\overline{\gamma} - \lambda)}{s\underline{\gamma} - \lambda},$$

and hence

$$\begin{aligned}A = \int_{t_0}^{t_1} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_0+\overline{\delta}} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt \\ + \int_{t_0+\overline{\delta}}^{t_5} \left[x(t_0 + \overline{\delta}) + (\lambda - s\underline{\gamma})(t - t_0 - \overline{\delta})\right],\end{aligned}$$

where

$$x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0,$$

$$x(t_0 + \overline{\delta}) = x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})\overline{\delta},$$

subsituiting into the expression for A we get

$$A = \frac{x_0^2 + 2sx_0(\underline{\gamma} - \overline{\gamma})\overline{\delta} - s(\underline{\gamma} - \overline{\gamma})\overline{\delta}\left[s(2t_0\underline{\gamma} + \overline{\gamma}\overline{\delta}) - (2t_0 + \overline{\delta})\lambda\right]}{2s\underline{\gamma} - 2\lambda}$$
$$= s(\overline{\gamma} - \underline{\gamma})\overline{\delta}t_0 + \frac{x_0^2 + 2sx_0(\underline{\gamma} - \overline{\gamma})\overline{\delta} - s(\underline{\gamma} - \overline{\gamma})\overline{\delta}^2(s\overline{\gamma} - \lambda)}{2s\underline{\gamma} - 2\lambda},$$

which is minimized at $t_0^* = 0$.

Case 2.3 $\overline{\delta} > t_1 = \frac{x_0}{s\overline{\gamma} - \lambda}$**.**

This is the situation where whatever $0 \le t_0 \le T - \overline{\delta}$ one chooses, one will be able to clear the system before $t_0 + \overline{\delta}$. Let t_6 be the time when the fluid is cleared if one starts at t_0, then t_6 satisfies

$$x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})(t_6 - t_0) = 0,$$

which gives us $t_6 = \frac{x_0 + (s\overline{\gamma} - s\underline{\gamma})t_0}{s\overline{\gamma} - \lambda}$. Hence for any $0 \le t_0 \le \min\{T - \overline{\delta}, t_3\}$, we have

$$A = \int_0^{t_0} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_6} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt,$$

where

$$x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0,$$

substituiting into the expression of A we get

$$A = \frac{x_0^2 + 2st_0x_0(-\underline{\gamma} + \overline{\gamma}) + st_0^2(\underline{\gamma} - \overline{\gamma})(s\underline{\gamma} - \lambda)}{2s\overline{\gamma} - 2\lambda},$$

which is minimized at $t_0^* = 0$.

Case 3 $s\underline{\gamma} < \lambda < s\overline{\gamma}$

Let $t_1 \geq 0$ denote the time it takes to clear the system if one starts to apply $\overline{\gamma}$ from 0 and is allowed to apply it for as long as possible. Then t_1 satisfies

$$x_0 + (\lambda - s\overline{\gamma})t_1 = 0,$$

which gives us $t_1 = \frac{x_0}{s\overline{\gamma}-\lambda}$.

Let $t_2 \geq 0$ be the point such that if one applies $\underline{\gamma}$ from $[0, t_2]$ and $\overline{\gamma}$ from $[t_2, t_2 + \overline{\delta}]$ then the fluid will drop to zero at $t_2 + \overline{\delta}$. Then t_2 satisfies

$$x(t_2 + \overline{\delta}) = x_0 + (\lambda - s\underline{\gamma})t_2 + (\lambda - s\overline{\gamma})\overline{\delta} = 0,$$

which gives us $t_2 = \frac{\overline{\delta}(s\overline{\gamma}-\lambda)-x_0}{\lambda-s\underline{\gamma}}$. We consider three scenarios based on the definitions of t_1 and t_2.

Case 3.1 $T < t_1$ **or** $T > t_1$ **and** $\overline{\delta} < t_1$**.**

This is the situation where no matter where one starts to apply $\overline{\gamma}$ for a maximum length of $\overline{\delta}$ one will not be able to clear the system before T. For any $0 \leq t_0 \leq T - \overline{\delta}$, we have

$$A = \int_0^{t_0} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_0+\overline{\delta}} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt \\ + \int_{t_0+\overline{\delta}}^{T} \left[x(t_0 + \overline{\delta}) + (\lambda - s\underline{\gamma})(t - t_0 - \overline{\delta})\right] dt,$$

where

$$x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0,$$

$$x(t_0 + \overline{\delta}) = x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})\overline{\delta},$$

subsituiting into the expression for A we get

$$A = \frac{1}{2}\left\{-s(\underline{\gamma} - \overline{\gamma})\overline{\delta}(2t_0 + \overline{\delta}) + 2T\left[x_0 + s(\underline{\gamma} - \overline{\gamma})\overline{\delta}\right] + T^2(-s\underline{\gamma} + \lambda)\right\} \\ = 2s(\overline{\gamma} - \underline{\gamma})\overline{\delta}t_0 + T\left[x_0 + s(\underline{\gamma} - \overline{\gamma})\overline{\delta}\right] + \frac{1}{2}T^2(-s\underline{\gamma} + \lambda),$$

which is minimized at $t_0^* = 0$.

Case 3.2 $\overline{\delta} > t_1$ **(hence** $T > t_1$**) and** $T > t_2 + \overline{\delta}$

This is the case where if and only if one chooses $t_0 \leq t_2$ then one will be able to clear the system at some point before T. First we show that it is suboptimal to let $t_0 > t_2$. We consider two policies π and γ for which $t_0^\pi = t_2$ and $t_2 < t_0^\gamma < T - \overline{\delta}$. And for both π and γ we have $\delta^\pi = \delta^\gamma = \overline{\delta}$. We are going to show that $A^\pi < A^\gamma$. First, it is easy to see that $x^\pi(t) = x^\gamma(t)$ for $0 \leq t \leq t_0^\pi = t_2$. Hence

$$A^\pi - A^\gamma = \int_{t_2}^{T} x^\pi(t)dt - \int_{t_2}^{T} x^\gamma(t)dt.$$

Also, $x^\pi(t_2) = x^\gamma(t_2) = x_0 + (\lambda - s\underline{\gamma})t_2$. To compute $\int_{t_2}^{T} x^\pi(t)dt$ we write

$$\int_{t_2}^{T} x^\pi(t)dt = \int_{t_2}^{t_2+\overline{\delta}} \left[x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)\right] dt + \int_{t_2+\overline{\delta}}^{T} (\lambda - s\underline{\gamma})(t - t_2 - \overline{\delta})dt,$$

where we have used the fact that that $x^\pi(t_2 + \overline{\delta}) = 0$, by definition of t_2. To compute $\int_{t_2}^{T} x^\gamma(t)dt$ we write

$$\begin{aligned}\int_{t_2}^{T} x^\gamma(t)dt = \int_{t_2}^{t_0^\gamma} [x(t_2) + (\lambda - s\underline{\gamma})(t - t_2)]dt + \int_{t_0^\gamma}^{t_0^\gamma+\overline{\delta}} [x^\gamma(t_0^\gamma) + (\lambda - s\overline{\gamma})(t - t_0^\gamma)]dt \\ + \int_{t_0^\gamma+\overline{\delta}}^{T} [x^\gamma(t_0^\gamma + \overline{\delta}) + (\lambda - s\underline{\gamma})(t - t_0^\gamma - \overline{\delta})]dt,\end{aligned}$$

where

$$x^\gamma(t_0^\gamma) = x(t_2) + (\lambda - s\underline{\gamma})(t_0^\gamma - t_2) = x(t_2) + (\lambda - s\underline{\gamma})\Delta.$$

$$\begin{aligned}x^\gamma(t_0^\gamma + \overline{\delta}) =& x_0^\gamma(t_0^\gamma) + (\lambda - s\overline{\gamma})\overline{\delta} \\ =& x(t_2) + (\lambda - s\underline{\gamma})\Delta + (\lambda - s\overline{\gamma})\overline{\delta} \\ =& (\lambda - s\underline{\gamma})\Delta,\end{aligned}$$

where we denote $\Delta = t_0^\gamma - t_3$. Notice that we have used the fact that $x^\pi(t_2+\overline{\delta}) = x(t_2)+(\lambda-s\overline{\gamma})\overline{\delta} = 0$, by definition of t_2, and the fact that for $t_0^\gamma > t_2$, one must have $x^\gamma(t_0^\gamma + \overline{\delta}) > 0$. Substituiting the

above two equations into that of $\int_{t_2}^{T} x^{\gamma}(t)dt$ we have

$$\begin{aligned}
\int_{t_2}^{T} x^{\gamma}(t)dt &= \int_{t_2}^{t_0^{\gamma}} [x(t_2) + (\lambda - s\underline{\gamma})(t - t_2)dt + \int_{t_0^{\gamma}}^{t_0^{\gamma}+\overline{\delta}} [x(t_2) + (\lambda - s\underline{\gamma})\Delta + (\lambda - s\overline{\gamma})(t - t_0^{\gamma})]dt \\
&\quad + \int_{t_0^{\gamma}+\overline{\delta}}^{T} [(\lambda - s\underline{\gamma})\Delta + (\lambda - s\underline{\gamma})(t - t_0^{\gamma} - \overline{\delta})]dt \\
&= \int_{t_2}^{t_0^{\gamma}} [x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)]dt + \int_{t_2}^{t_0^{\gamma}} s(\overline{\gamma} - \underline{\gamma})(t - t_2)dt \\
&\quad + \int_{t_0^{\gamma}}^{t_2+\overline{\delta}} [x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)]dt + \int_{t_0^{\gamma}}^{t_2+\overline{\delta}} [(\lambda - s\underline{\gamma})\Delta - (\lambda - s\overline{\gamma})\Delta]dt \\
&\quad + \int_{t_2+\overline{\delta}}^{t_0^{\gamma}+\overline{\delta}} [x(t_2) + (\lambda - s\underline{\gamma})\Delta + (\lambda - s\overline{\gamma})(t - t_0^{\gamma})]dt \\
&\quad + \int_{t_2+\overline{\delta}}^{T} (\lambda - s\underline{\gamma})(t - t_2 - \overline{\delta})dt - \int_{t_0+\overline{\delta}}^{t_0^{\gamma}+\overline{\delta}} [(\lambda - s\underline{\gamma})\Delta + (\lambda - s\underline{\gamma})(t - t_0^{\gamma} - \overline{\delta})]dt \\
&= \int_{t_2}^{t_2+\overline{\delta}} [x(t_2) + (\lambda - s\overline{\gamma})(t - t_2)]dt + \int_{t_2+\overline{\delta}}^{T} (\lambda - s\underline{\gamma})(t - t_2 - \overline{\delta})dt \\
&\quad + s(\overline{\gamma} - \underline{\gamma})\frac{\Delta^2}{2} + s(\overline{\gamma} - \underline{\gamma})\Delta(\overline{\delta} - \Delta) + x(t_2)\Delta + (\lambda - s\underline{\gamma})\Delta^2 \\
&\quad + (\lambda - s\overline{\gamma})\Delta(\overline{\delta} - \frac{\Delta}{2}) - (\lambda - s\underline{\gamma})\Delta^2 + (\lambda - s\underline{\gamma})\frac{\Delta^2}{2} \\
&= \int_{t_2}^{T} x^{\pi}(t)dt + \Delta[(\lambda - s\underline{\gamma})\overline{\delta} + x(t_2)] > \int_{t_2}^{T} x^{\pi}(t)dt,
\end{aligned}$$

and thus $A^{\pi} < A^{\gamma}$. It is suboptimal to start after t_2.

Now We have shown that one shall choose $0 \leq t_0 \leq t_2$. We find out the optimal t_0 by expressing A as a function of t_0 and minimize it. Since $t_0 \leq t_2$, by definition of t_2 we will be able to clear the system at some point before T. Let t_3 be such point, then t_3 satisfies

$$x_0 + (\lambda - s\underline{\gamma})t_0 + (\lambda - s\overline{\gamma})(t_3 - t_0) = 0,$$

which gives us $t_3 = \frac{(s\overline{\gamma} - s\underline{\gamma})t_0 + x_0}{s\overline{\gamma} - \lambda}$. It is easy to see that $t_3 - t_0 < \overline{\delta}$, hence we will be able to keep the fluid level at 0 for a positive amount of time until t_4, and $t_4 = t_3 + (\overline{\delta} - (t_3 - t_0)) = t_0 + \overline{\delta} < t_2 + \overline{\delta} < T$. Hence from t_4 and on until T the fluid level will build up by a rate of $\lambda - s\underline{\gamma}$ per unit of time. Based on the discussion we can write

$$A = \int_0^{t_0} \left[x_0 + (\lambda - s\underline{\gamma})t\right] dt + \int_{t_0}^{t_3} \left[x(t_0) + (\lambda - s\overline{\gamma})(t - t_0)\right] dt + \int_{t_4}^{T} (\lambda - s\underline{\gamma})(t - t_4)dt,$$

where $x(t_0) = x_0 + (\lambda - s\underline{\gamma})t_0$. Substituiting into the expression of A we get

$$A = at_0^2 + bt_0 + c,$$

where

$$a = -s\underline{\gamma} + \lambda + \frac{(-s\underline{\gamma} + \lambda)^2}{2(s\overline{\gamma} - \lambda)} = (\lambda - s\underline{\gamma})\left(1 + \frac{\lambda - s\underline{\gamma}}{2(s\overline{\gamma} - \lambda)}\right).$$

$$b = x_0 - T(-s\underline{\gamma} + \lambda) + \overline{\delta}(-s\underline{\gamma} + \lambda) + \frac{x_0(-s\underline{\gamma} + \lambda)}{s\overline{\gamma} - \lambda} = \frac{s\overline{\gamma} - s\underline{\gamma}}{s\overline{\gamma} - \lambda}x_0 - (\lambda - s\underline{\gamma})(T - \overline{\delta}).$$

Hence we have

$$-\frac{b}{2a} = \frac{T - \overline{\delta} - \frac{s\overline{\gamma} - s\underline{\gamma}}{(s\overline{\gamma} - \lambda)(\lambda - s\underline{\gamma})}x_0}{2 + \frac{\lambda - s\underline{\gamma}}{s\overline{\gamma} - \lambda}}.$$

If $-\frac{b}{2a} < 0$, then

$$t_0^* = 0.$$

If $-\frac{b}{2a} > t_2$, then

$$t_0^* = t_2 = \frac{\overline{\delta}(s\overline{\gamma} - \lambda) - x_0}{\lambda - s\underline{\gamma}}.$$

else

$$t_0^* = -\frac{b}{2a} = \frac{T - \overline{\delta} - \frac{s\overline{\gamma} - s\underline{\gamma}}{(s\overline{\gamma} - \lambda)(\lambda - s\underline{\gamma})}x_0}{2 + \frac{\lambda - s\underline{\gamma}}{s\overline{\gamma} - \lambda}}.$$

Case 3.3 $\overline{\delta} > t_1$ and $T < t_2 + \overline{\delta}$.

By the condition $T < t_2 + \overline{\delta}$ we get $T - \overline{\delta} < t_2$. Since we have already shown that it is suboptimal to $t_0 > T - \overline{\delta}$. For any $0 \le t_0 \le T - \overline{\delta} < t_2$, let t_3 and t_4 be as defined in Case 3.2, they shall have the same expressions. The fluid will drop to zero at t_3, stay there until t_4 and then build up at a rate of $\lambda - s\underline{\gamma}$ until T, which is because $t_4 = t_0 + \overline{\delta} < T - \overline{\delta} + \overline{\delta} = T$. And hence A take the same expression as in Case 3.2. Hence if $-\frac{b}{2a} < 0$, then $t_0^* = 0$; If $-\frac{b}{2a} > T - \overline{\delta}$, then $t_0^* = T - \overline{\delta}$; else $t_0^* = -\frac{b}{2a}$, where a and b take the same expressions as in Case 3.2.

Summarizing the results from all cases considered, we end up with the result of the theorem. This completes the proof.

A.6 Single-Server Clearing Model

We consider a single-server clearing model for which all model assumptions are the same as those for the queueing model discussed in Chapter 2, except that there is no external arrivals to the system and the system starts with a fixed number of customers, denoted by N, where $N \geq 1$. We also assume without loss of generality that the first customer in queue enters service at the beginning of time. Our objective is to minimize total expected cost until the system is emptied.

We formulate the problem as a Markov decision process. The state space $\mathbb{X}$ can be described as $\mathbb{X} = \{0\} \cup \{(m, n) \mid m \in \{\alpha_i\}_{i=1}^{\infty} \cup \{2, 3\},\ 1 \leq n \leq N\}$. The state definition has the same format and meanings as that for the main model in discussion except that the number of customers in the system cannot exceed N, the initial number of customers presented, since there is no external arrivals from time 0 and onwards. As before, We restrict ourselves to the policy set Π, where any $\pi \in \Pi$ is a stationary, non-idling, state-dependent policy, and is a mapping from the system state $\mathbb{X}$ to the action space $\mathcal{A} = \{0, 1\}$ where 0 corresponds to the decision of starting a sequential service and 1 corresponds to the decision of starting a parallel service under the constraint no action is available in state (0) and action 1 is only available in states $x = (\alpha_i, n)$, for some i and $1 \leq n \leq N$. Again, here the policies we consider can be seen as preemptive in the sense that the system controller can switch from "parallel service" to "service in sequence" at a decision epoch, which can correspond to either an arrival time or a service completion time, as long as neither primary nor secondary service is complete for the customer or from "service in sequence" to "parallel service" as long as the primary service of the customer is still in progress.

We apply uniformization with uniformization constant $\beta = \gamma_1 + \gamma_2$. Without loss of generality we assume $\beta = 1$. Thus, instead of considering a continuous-time problem defined above we can look at a discrete-time equivalent. The *finite horizon total cost optimality equations* (FHTCOE) for the single server clearing model are

$$V(0) = 0. \tag{A.1}$$

For $1 \le n \le N$,

$$V(\alpha_i, n) = nC_w + (1-\alpha_i)\gamma_1 \sum_j q_j V(\alpha_j, n-1) + \alpha_i \gamma_1 V(3,n) + \gamma_2 \min\{V(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s, V(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p\}, \quad \text{(A.2)}$$

where we define $V(\alpha_i, 0) = V(0) = 0$ for all i for convenience of representation. And $(\alpha_i, 0)$s are not in the state space.

$$V(2,n) = nC_w + \gamma_1 \sum_j q_j V(\alpha_j, n-1) + \gamma_2 V(2,n). \quad \text{(A.3)}$$

$$V(3,n) = nC_w + \gamma_2 \sum_j q_j V(\alpha_j, n-1) + \gamma_1 V(3,n). \quad \text{(A.4)}$$

Next we show that the optimal policy can be characterized by a simple formula. Notice that from the optimality equations, it is optimal to use the parallel service option if and only if $V(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s > V(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p$ at state (α_i, n) where the system has n customers and the customer who is currently receiving stage-1 service and for whom no stage-2 service has ever been completed has type-2 probability being α_i. The next theorem shows when the inequality is satisfied and thus prescribes the structure of the optimal policy.

Theorem 5. *For the single-server clearing model, it is optimal to call ahead if and only if* $nC_w \frac{\alpha_i}{\gamma_1+\gamma_2} \ge C_p - \alpha_i C_s$.

Proof. We only need to show that $V(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s \ge V(2,n) + \frac{\gamma_1+\gamma_2}{\gamma_2} C_p$, or $V(\alpha_i, n) - V(2,n) \ge \frac{\gamma_1+\gamma_2}{\gamma_2} C_p - \frac{\alpha_i \gamma_1}{\gamma_2} C_s$ if and only if $nC_w \frac{\alpha_i}{\gamma_1+\gamma_2} \ge C_p - \alpha_i C_s$. Subtracting (A.3) from (A.2) we get

$$\begin{aligned} V(\alpha_i, n) - V(2,n) =& \alpha_i \gamma_1 \left[V(3,n) - \sum_j q_j V(\alpha_j, n-1) \right] \\ &+ I_{\{V(\alpha_i,n) - V(2,n) \le \frac{\gamma_1+\gamma_2}{\gamma_2} C_p - \frac{\alpha_i \gamma_1}{\gamma_2} C_s\}} \gamma_2 \left[V(\alpha_i, n) + \frac{\alpha_i \gamma_1}{\gamma_2} C_s - V(2,n) \right] \\ &+ I_{\{V(\alpha_i,n) - V(2,n) > \frac{\gamma_1+\gamma_2}{\gamma_2} C_p - \frac{\alpha_i \gamma_1}{\gamma_2} C_s\}} (\gamma_1+\gamma_2) C_p. \end{aligned}$$

Due to (A.4) we have $V(3,n) - \sum_j q_j V(\alpha_j, n-1) = \frac{nC_w}{\gamma_2}$, substituiting into the above we get

$$\begin{aligned} V(\alpha_i, n) - V(2,n) =& \frac{n\alpha_i\gamma_1}{\gamma_2} C_w \\ &+ I_{\{V(\alpha_i,n)+\frac{\alpha_i\gamma_1}{\gamma_2}C_s \le V(2,n)+\frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}} \gamma_2 \left[V(\alpha_i,n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s - V(2,n) \right] \\ &+ I_{\{V(\alpha_i,n)+\frac{\alpha_i\gamma_1}{\gamma_2}C_s > V(2,n)+\frac{\gamma_1+\gamma_2}{\gamma_2}C_p\}} (\gamma_1+\gamma_2) C_p. \end{aligned}$$

This means if $V(\alpha_i, n) - V(2,n) \le \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\alpha_i\gamma_1}{\gamma_2}C_s$, then

$$\begin{aligned} V(\alpha_i, n) - V(2,n) = \frac{n\alpha_i\gamma_1}{\gamma_2}C_w + \gamma_2 \left[V(\alpha_i,n) + \frac{\alpha_i\gamma_1}{\gamma_2}C_s - V(2,n) \right] \\ \le \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\alpha_i\gamma_1}{\gamma_2}C_s \iff \\ V(\alpha_i, n) - V(2,n) = \frac{n\alpha_i}{\gamma_2}C_w + \alpha_i C_s \le \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\alpha_i\gamma_1}{\gamma_2}C_s \iff \\ nC_w \frac{\alpha_i}{\gamma_1+\gamma_2} < C_p - \alpha_i C_s. \end{aligned}$$

While if $V(\alpha_i, n) - V(2,n) > \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\alpha_i\gamma_1}{\gamma_2}C_s$ then

$$\begin{aligned} V(\alpha_i, n) - V(2,n) = \frac{n\alpha_i\gamma_1}{\gamma_2}C_w + (\gamma_1+\gamma_2)C_p \ge \frac{\gamma_1+\gamma_2}{\gamma_2}C_p - \frac{\alpha_i\gamma_1}{\gamma_2}C_s \\ \iff nC_w \frac{\alpha_i}{\gamma_1+\gamma_2} < C_p - \alpha_i C_s. \end{aligned}$$

□

A.7 Tables for Chapter 5

Table A.1: Odds ratios of Prob(high acuity)/Prob(low or medium acuity) = Prob(medium or high acuity)/Prob(low acuity) for chief complaint (contrast: other) from the model for the association between ED census and triage decisions. (A model where the two odds ratios were not necessarily the same for chief complaints provided similar results.)

	P(high)/P(low or medium)	95% CI for OR
abdominal pain	1.982	[1.858,2.114]
abdominal swelling	1.486	[.678,3.258]
		Continued on next page

Table A.1 – continued from previous page

	P(high)/P(low or medium)	95% CI for OR
abnormal electrocardiogram	3.331	[1.589,6.982]
abnormal laboratory test	2.287	[1.594,3.282]
altered mental status	10.690	[9.126,12.522]
anorexia	.959	[.566,1.625]
atrial fibrillation	6.865	[4.502,10.470]
back pain	.212	[.193,.234]
blood in stool	1.416	[1.004,1.997]
cancer	4.550	[3.560,5.817]
chest pain	3.798	[3.522,4.096]
confusion	2.776	[1.264,6.097]
cough	.447	[.377,.530]
crohns flare	2.245	[1.120,4.502]
dehydration	2.074	[1.413,3.044]
dialysis	1.118	[.695,1.800]
dyspnea	2.962	[2.393,3.666]
fever	1.292	[1.181,1.412]
gastrointestinal bleed	3.223	[2.039,5.092]
headache	1.167	[1.051,1.297]
hemoptysis	1.564	[.882,2.774]
high blood sugar	2.463	[1.795,3.379]
hypotension	11.655	[6.642,20.452]
jaundice	1.406	[.679,2.913]
lethargic	2.869	[1.466,5.615]
loss of consciousness	2.229	[1.061,4.683]
overdose	68.729	[37.241,126.839]
palpitations	2.107	[1.323,3.354]

Continued on next page

Table A.1 – continued from previous page

	P(high)/P(low or medium)	95% CI for OR
pancreatitis	1.795	[.978,3.295]
pneumonia	2.433	[1.716,3.450]
pulmonary embolus	7.302	[3.453,15.439]
rapid heart rate	6.345	[3.820,10.540]
rectal bleed	1.852	[1.424,2.409]
respiratory distress	17.610	[12.175,25.471]
seizure	3.854	[3.012,4.932]
shortness of breath	3.324	[2.993,3.692]
slurred speech	3.958	[1.967,7.963]
stroke	14.250	[9.885,20.544]
syncope	2.541	[2.155,2.995]
tachycardia	5.246	[2.609,10.549]
transient ischemic attack	4.825	[2.330,9.991]
unable to walk	1.547	[0.666,3.591]
vomiting blood	2.278	[1.520,3.415]
weakness	2.129	[1.771,2.559]
wheezing	1.671	[1.144,2.442]

Table A.2: Odds ratios of Prob(admit) versus Prob(discharge) for chief complaint (contrast: other) from the model for the association between ED census and disposition decisions.

	Prob(admit)/Prob(discharge)	95% CI for OR
abdominal pain	1.155	[1.068,1.248]
abdominal swelling	1.913	[.816,4.488]
abnormal electrocardiogram	.668	[.292,1.528]
abnormal laboratory test	3.273	[2.185,4.902]

Continued on next page

Table A.2 – continued from previous page

	Prob(admit)/Prob(discharge)	95% CI for OR
altered mental status	1.974	[1.625,2.398]
anorexia	3.987	[2.001,7.946]
atrial fibrillation	2.911	[1.627,5.207]
back pain	.450	[.367,.551]
blood in stool	1.303	[.888,1.912]
cancer	2.977	[2.210,4.010]
chest pain	1.511	[1.387,1.645]
confusion	1.078	[.454,2.561]
cough	.676	[.503,.910]
crohns flare	3.252	[1.523,6.942]
dehydration	1.792	[1.182,2.716]
dialysis	2.241	[1.314,3.820]
dyspnea	1.226	[.945,1.590]
fever	1.429	[1.260,1.620]
gastrointestinal bleed	4.417	[2.386,8.178]
headache	.452	[.380,.539]
hemoptysis	5.894	[2.892,12.012]
high blood sugar	1.136	[.807,1.601]
hypotension	1.534	[.781,3.014]
jaundice	3.263	[1.331,8.000]
lethargic	1.774	[.826,3.809]
loss of consciousness	.812	[.341,1.934]
overdose	1.866	[1.180,2.951]
palpitations	.379	[.204,.701]
pancreatitis	8.739	[4.062,18.804]
pneumonia	3.281	[2.175,4.949]

Continued on next page

Table A.2 – continued from previous page

	Prob(admit)/Prob(discharge)	95% CI for OR
pulmonary embolus	1.315	[.559,3.091]
rapid heart rate	.716	[.396,1.295]
rectal bleed	1.619	[1.217,2.155]
respiratory distress	4.801	[2.766,8.335]
seizure	.799	[.598,1.067]
shortness of breath	2.096	[1.857,2.365]
slurred speech	1.947	[.877,4.322]
stroke	1.801	[1.131,2.869]
syncope	1.074	[.896,1.288]
tachycardia	1.233	[.552,2.752]
transient ischemic attack	1.078	[.487,2.386]
unable to walk	2.389	[.989,5.772]
vomiting blood	1.678	[1.069,2.634]
weakness	1.892	[1.548,2.311]
wheezing	1.511	[.913,2.502]

Table A.3: Odds ratios of Prob(admit) versus Prob(discharge) for interaction terms between ESI and age group (contrast: ESI3 and Age Group 18 to 40) from the model for the association between ED census and disposition decisions. (Some of the interaction terms are omitted due to small sample sizes.)

	Prob(admit)/Prob(discharge)	95% CI for OR
ESI2 and 3m below	1.150	[.642,2.059]
ESI4 and 3m below	.760	[.350,1.649]
ESI2 and 3m to 3	1.382	[.993,1.923]
ESI4 and 3m to 3	.481	[.307,.753]
ESI5 and 3m to 3	.743	[.076,7.213]

Continued on next page

Table A.3 – continued from previous page

	Prob(admit)/Prob(discharge)	95% CI for OR
ESI1 and 3 to 8	.971	[.197,4.775]
ESI2 and 3 to 8	1.119	[.808,1.550]
ESI4 and 3 to 8	.829	[.544,1.262]
ESI1 and 8 to 18	3.197	[.403,25.367]
ESI2 and 8 to 18	.731	[.590,.907]
ESI4 and 8 to 18	.811	[.555,1.187]
ESI5 and 8 to 18	3.170	[.630,1.596]
ESI1 and 40 to 55	.564	[.260,1.228]
ESI2 and 40 to 55	.854	[.732,.995]
ESI4 and 40 to 55	.832	[.617,1.123]
ESI5 and 40 to 55	.944	[.157,5.684]
ESI1 and 55 to 70	.644	[.266,1.560]
ESI2 and 55 to 70	.888	[.751,1.050]
ESI4 and 55 to 70	1.101	[.812,1.494]
ESI5 and 55 to 70	.667	[.069,6.465]
ESI1 and 70 and above	1.015	[.331,3.119]
ESI2 and 70 and above	.871	[.717,1.059]
ESI4 and 70 and above	1.317	[.905,1.915]
ESI5 and 70 and above	6.992	[1.113,4.394]

We used three criteria to check multicollinearity between control variables as suggested in (Belsley et al., 2005). First, we calculated the Variance Inflation Factors (VIF), which are the diagonal entries of the standardized design matrix. It is generally considered that variance inflation factors greater than 10 are indicative of significant multicollinearity. In our data, the largest VIF is only 0.002. Second, we checked the condition indices as denoted by $\eta_k = \frac{\mu_{\mathbf{max}}}{\mu_k}$, where μ_k's are the singular values of the standardized design matrix and max is their maximum. According to

(Belsley et al., 2005), a rough guide is that condition indices of the order 5-10 are associated with weak dependencies, but those in the range 30-100 imply moderate to strong association. In our data, η_k's corresponding to the interaction terms between acuity and age, and are generally within the range of 5-10, hence indicative of a weak dependency between the age and acuity interaction terms but those for all the other variables are very small not indicative of any multicollinearity. Finally, we checked the quantity π_{kj}'s, which measure the proportion of the variance of the j_{th} parameter estimate that is accounted for by the k_{th} singular value. (Belsley et al., 2005) suggest to look for instances in which a k with large η_k gives rise to at least two large values of π_{kj} and that proportions of the order of 0.999 are not uncommon in cases of serious multicollinearity. In our data, we did not find any such instance. In summary, based on the three criteria we checked, we did not find any severe multicollinearity between any control variables of interest in this study.

BIBLIOGRAPHY

Agresti, A. (2003). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Ahalt, V., Argon, N. T., Ziya, S., Strickler, J., and Mehrotra, A. (2016). Comparison of emergency department crowding scores: A discrete-event simulation approach. *Health Care Management Science*, pages 1–12.

Barak-Corren, Y., Israelit, S. H., and Reis, B. Y. (2017). Progressive prediction of hospitalisation in the emergency department: Uncovering hidden patterns to improve patient flow. *Emergency Medicine Journal*, pages emermed–2014.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Buzacott, J. A. (1996). Commonalities in reengineered business processes: Models and issues. *Management Science*, 42(5):768–782.

Crabill, T. B. (1972). Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Science*, 18(9):560–566.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

Fogarty, E., Saunders, J., and Cummins, F. (2014). The effect of boarders on emergency department process flow. *The Journal of Emergency Medicine*, 46(5):706–710.

Freeman, M., Robinson, S., and Scholtes, S. (2017). Gatekeeping under congestion: An empirical study of referral errors in the emergency department.

Freeman, M., Savva, N., and Scholtes, S. (2016). Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*.

George, F. and Evridiki, K. (2015). The effect of emergency department crowding on patient outcomes. *Health Science Journal*, 9(1):1.

George, J. M. and Harrison, J. M. (2001). Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731.

Gilboy, N., Tanabe, P., Travers, D., Rosenau, A. M., et al. (2012). Emergency severity index (esi): a triage tool for emergency department care, version 4. *Implementation handbook*, pages 12–0014.

Gorski, J. K., Batt, R. J., Otles, E., Shah, M. N., Hamedani, A. G., and Patterson, B. W. (2017). The impact of emergency department census on the decision to admit. *Academic Emergency Medicine*, 24(1):13–21.

Harrison, J. (1985). Brownian motion and stochastic flow systems.

Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136.

Horwitz, L. I., Green, J., and Bradley, E. H. (2010). Us emergency department performance on wait time and length of visit. *Annals of Emergency Medicine*, 55(2):133–141.

Hwang, U., Weber, E. J., Richardson, L. D., Sweet, V., Todd, K., Abraham, G., and Ankel, F. (2011). A research agenda to assure equity during periods of emergency department crowding. *Academic Emergency Medicine*, 18(12):1318–1323.

Kang, H., Nembhard, H. B., Rafferty, C., and DeFlitch, C. J. (2014). Patient flow in the emergency department: a classification and analysis of admission process policies. *Annals of Emergency Medicine*, 64(4):335–342.

Kc, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

Kc, D. S. and Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65.

Koçağa, Y. L., Armony, M., and Ward, A. R. (2015). Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, 24(7):1101–1117.

Kuntz, L., Mennicken, R., and Scholtes, S. (2014). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771.

LaMantia, M. A., Platts-Mills, T. F., Biese, K., Khandelwal, C., Forbach, C., Cairns, C. B., Busby-Whitehead, J., and Kizer, J. S. (2010). Predicting hospital admission and returns to the emergency department for elderly patients. *Academic Emergency Medicine*, 17(3):252–259.

McCarthy, M. L., Zeger, S. L., Ding, R., Levin, S. R., Desmond, J. S., Lee, J., and Aronsky, D. (2009). Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine*, 54(4):492–503.

McCusker, J., Vadeboncoeur, A., Lévesque, J.-F., Ciampi, A., and Belzile, E. (2014). Increases in emergency department occupancy are associated with adverse 30-day outcomes. *Academic Emergency Medicine*, 21(10):1092–1100.

Miller, J. G. (1960). Information input overload and psychopathology. *American Journal of Psychiatry*, 116(8):695–704.

Morganti, K. G., Bauhoff, S., Blanchard, J. C., Abir, M., Iyer, N., Smith, A., Vesely, J. V., Okeke, E. N., and Kellermann, A. L. (2013). The evolving role of emergency departments in the united states. *Rand Health Quarterly*, 3(2).

Olshaker, J. S. (2009). Managing emergency department overcrowding. *Emergency Medicine Clinics of North America*, 27(4):593–603.

Peck, J., Benneyan, J., Gaehde, S., Nightingale, D., and Boston, V. (2012a). Models for using predictions to facilitate hospital patient flow. In *Healthcare Systems Process Improvement Conference*, pages 1–6.

Peck, J. S., Benneyan, J. C., Nightingale, D. J., and Gaehde, S. A. (2012b). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, 19(9).

Qiu, S., Chinnam, R. B., Murat, A., Batarse, B., Neemuchwala, H., and Jordan, W. (2015). A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. *Health Care Management Science*, 18(1):67–85.

Richardson, D. (1998). No relationship between emergency department activity and triage categorization. *Academic Emergency Medicine*, 5(2):141–145.

Richardson, D. B. et al. (2006). Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medical Journal of Australia*, 184(5):213–216.

Sennott, L. I. (2009). *Stochastic Dynamic Programming and the Control of Queueing Systems*, volume 504. John Wiley & Sons.

Stidham Jr, S. and Weber, R. R. (1989). Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations Research*, 37(4):611–625.

Travers, D., Mehrotra, A., Chen, W., Lopiano, K., Bohrmann, T., Argon, N., Ziya, S., Strickler, J., and Linthicum, B. (2017). Starting with a clear endpoint: Development of a tool to predict admission at triage. *Academic Emergency Medicine*, 24(S1):18.

Van Der Aalst, W. M. (2013). Business process management: A comprehensive survey. *ISRN Software Engineering*, 2013.

Weber, R. R. and Stidham, S. (1987). Optimal control of service rates in networks of queues. *Advances in Applied Probability*, 19(1):202–218.

Welch, S. J., Asplin, B. R., Stone-Griffith, S., Davidson, S. J., Augustine, J., Schuur, J., and Alliance, E. D. B. (2011). Emergency department operational metrics, measures and definitions: Results of the second performance measures and benchmarking summit. *Annals of Emergency Medicine*, 58(1):33–40.