Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments

Soroush Saghafian
Industrial Engineering, School of Computing, Informatics and Decision Systems Engineering,
Arizona State University, Tempe, Arizona 85281, soroush.saghafian@asu.edu

Wallace J. Hopp
Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109, whopp@umich.edu

Mark P. Van Oyen
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109, vanoyen@umich.edu

Jeffrey S. Desmond, Steven L. Kronick
Emergency Department, University of Michigan Hospital, Ann Arbor, Michigan 48109
{jdesmo@med.umich.edu, skronick@med.umich.edu}

Crisis-level overcrowding conditions in emergency departments (EDs) have led hospitals to seek out new patient-flow designs to improve both responsiveness and safety. One approach that has attracted attention and experimentation in the emergency medicine community is a system in which ED beds and care teams are segregated and patients are "streamed" based on predictions of whether they will be discharged or admitted to the hospital. In this paper, we use a combination of analytic and simulation models to determine whether such a streaming policy can improve ED performance, where it is most likely to be effective, and how it should be implemented for maximum performance. Our results suggest that the concept of streaming can indeed improve patient flow, but only in some situations. First, ED resources must be shared across streams rather than physically separated. This leads us to propose a new "virtual-streaming" patient flow design for EDs. Second, this type of streaming is most effective in EDs with (1) a high percentage of admitted patients, (2) longer care times for admitted patients than discharged patients, (3) a high day-to-day variation in the percentage of admitted patients, (4) long patient boarding times (e.g., caused by hospital "bed-block"), and (5) high average physician utilization. Finally, to take full advantage of streaming, physicians assigned to admit patients should prioritize upstream (new) patients, whereas physicians assigned to discharge patients should prioritize downstream (old) patients.

1. Introduction

Between 1996 and 2006, annual visits to emergency departments (EDs) in the United States increased by 32% (from 90.3 million to 119.2 million), whereas the number of hospital EDs decreased from 4,019 to 3,833 (Pitts et al. 2008). This trend has elevated ED overcrowding to crisis levels in many U.S. hospitals. Similar trends have intensified pressure on EDs around the world.

The consequences of ED overcrowding can be tragic. For example, in 2006, 49-year-old Beatrice Vance arrived at the busy ED of Vista Medical Center East in Waukegan, Illinois, complaining of nausea, shortness of breath, and chest pain. Triaged and sent to the ED waiting room, Mrs. Vance waited there for two hours without further attention. When she was finally called, she failed to respond and was found dead of an acute myocardial infarction (SoRelle 2006).

Other less tragic but still important consequences of ED overcrowding include patient "elopement" (i.e., leaving without being seen), ambulance diversions, and treatment delays (Hoot and Aronsky 2008). The ED overcrowding

situation has become so dire that the American College of Emergency Physicians (ACEP) in its 2006 report gave a failing mark to emergency care in 41 of 50 states in the United States, and a D− nationally for access to care (see American College of Emergency Physicians 2006). Some experts believe that the recent healthcare bill will exacerbate the already serious overcrowding problem in U.S. EDs (SoRelle 2010).

This situation has prompted researchers to investigate a variety of methods for alleviating ED overcrowding, including (1) personnel staffing, (2) hospital bed access control, (3) nonurgent and low acuity patient referrals, (4) ambulance diversion, (5) destination control, and (6) improved resource utilization (Hoot and Aronsky 2008).

The most direct way to alleviate crowding and improve responsiveness is by adding resources. However, because this is also the most expensive approach, it is generally not the preferred option. Recognizing this, Richardson (2003, p. 517) concluded, "the debate is no longer about the level of resources our EDs deserve, but rather about how

to ensure that ED resources are directed to those who need them—the patients in the waiting room." To achieve this, some practitioners have recently suggested streaming patients based on their likelihood of being admitted to the hospital. In one pioneering effort, Flinders Medical Center in Australia implemented a system in which ED patients and resources are divided into two streams: one for those likely to be discharged (hereafter "Discharge" or "*D*" patients) and one for those likely to be admitted to the hospital (hereafter "Admit" or "*A*" patients) (King et al. 2006, Ben-Tovim et al. 2008). They reported a 48-minute reduction in average time spent by the patients in the ED. Although Flinders is an Australian hospital, basic ED practices are very similar to those in the United States and other developed countries. However, because Flinders represented a single uncontrolled experiment in a specific environment in which other changes (e.g., lean initiatives) were implemented along with the streaming system, it is impossible to infer that their results are purely due to streaming and/or that they will translate to other EDs. Nevertheless, motivated by positive reports from Flinders, other hospitals such as Bendigo Health (Kinsman et al. 2008) have begun implementing similar strategies.

Although streaming patients based on the likelihood of being admitted to the hospital is new, patient streaming is not. By the 1980s most EDs (although not Flinders) had adopted separate "fast tracks" for patients with minor injuries (Welch 2009). In the 1990s, many EDs also established "observation units" for patients requiring lengthy diagnosis. However, as Welch (2009, p. 1) noted, "these innovations were the tip of the iceberg, and performance-driven emergency departments have been experimenting with models that segment patients into streams for more efficient health care delivery."

For clarity, we will use the term "streaming" to refer specifically to the newly proposed policy that separates patients (and resources) into different streams according to anticipated disposition (*A* or *D*). We label the conventional policy that treats both types of patients together (with pooled resources) as "pooling." It is well known from the operations management literature that pooling offers efficiency benefits resulting from improved resource utilization. This means that in order for streaming to be effective, it must offer advantages that offset its inherent "anti-pooling" disadvantage. The Flinders results suggest that this may be possible. However, because their results could be due to (a) specific conditions (e.g., high percentage of admits, the fact that they did not yet have a separate stream (fast track) for low acuity patients, etc.), (b) other changes (e.g., lean), or (c) a Hawthorne effect halo, we cannot say without a careful analysis.

In this paper, we use a combination of analytical and simulation models to perform a systematic study of the attractiveness of streaming. Specifically, we address the following questions:

1. *Whether streaming (or a variation on it) can improve ED performance?*

2. *Where (i.e., in what hospital environments) is streaming (or an effective variation on it) most attractive?*

3. *How should Admit/Discharge information be implemented for maximum effectiveness?*

The remainder of the paper is organized as follows. Section 2 summarizes previous research relevant to the above questions. Section 3 describes ED flows and performance metrics in order to construct models with which to understand them. Section 4 considers a simple clearing model with a single-stage service process, in which patients can be classified (*A* or *D*) without error. This analysis provides insight into the relative effectiveness of streaming and pooling with respect to sequencing patients into the examination rooms. Although this suggests that sequencing alone is not enough to overcome the antipooling disadvantage of streaming, it also indicates that streaming is more robust to patient mix variation and classification errors than is pooling, which can lead to streaming outperforming pooling in real-world settings. In §5, we consider another analytic clearing model, with perfect patient classification but with multistage service processes, in order to understand the impact of patient sequencing within the exam rooms (i.e., the order in which physicians visit the patients assigned to them) on the streaming versus pooling comparison. We find that prioritizing downstream (i.e., near service completion) *D* patients and upstream (i.e., recent arrivals) *A* type patients enhances the advantage of streaming over pooling. In §6, we use a simulation model of a realistic ED environment that includes dynamic patient arrivals, multistage service processes, and patient misclassification error to test the conjectures made from our analytic models. Taken together, our results suggest that, implemented properly in the right environment, streaming can significantly improve overall ED performance by substantially reducing wait times for *D* patients at the expense of only a modest increase in wait times for *A* patients. We conclude in §7 with a summary of our overall insights about *whether, where*, and *how* streaming can be a potentially attractive strategy for improving ED responsiveness. An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1096.

2. Literature Survey

There are two main streams of research related to the work of this paper: (1) empirical studies of the ED overcrowding problem (published in medical journals), and (2) general queueing systems research (published in operations research journals) that deal with pooling and/or customer sequencing. We highlight key contributions from each of these below.

For an excellent survey of empirical studies of ED overcrowding, see Hoot and Aronsky (2008). Some of these studies have examined the nature and extent of the problem. For example, Liew et al. (2003) showed that there is a strong correlation between ED length of stay and inpatient

length of stay and concluded that "strategies to reduce the length of stay in the ED may significantly reduce healthcare expenditures and patient morbidity" (Liew et al. 2003, p. 524). The Centers for Disease Control and Prevention (CDC) estimated that 379,000 deaths occurred in U.S. EDs in 2000 (McCaig and Ly 2002). Other studies have found that long waiting times are linked to patient mortality as well as elevated risks of errors and adverse events (e.g., Thomas et al. 2000, Gordon et al. 2001, and Trzeciak and Rivers 2003). One such study estimated that long waiting times and high occupancies caused 13 deaths per year in one Australian ED (Richardson 2006). Thus, reducing waiting times is a means for promoting higher levels of patient safety. Because admit patients typically include the most critical cases that need more rapid attention, some researchers have focused specifically on studying mortality among admit patients. For instance, Sprivulis et al. (2006) associated a combined measure of hospital and ED crowding (which causes long waiting times) with an increased risk of mortality among admitted patients.

Other studies have evaluated the factors that influence overcrowding. Miro et al. (2003) evaluated different internal factors that affect patient flow and concluded that ED overcrowding is driven by both external pressure and internal factors such as how flow across the ED is measured. Schull et al. (2007) studied the effect of low complexity ED patients on the waiting times of other patients and concluded that the impact is negligible. Still other papers have examined the impact of various reorganizations. The papers on the Flinders experiment with streaming (King et al. 2006, Ben-Tovim et al. 2008) fall into this category. Another example is Howell et al. (2004), which considered a new ED admission process in which ED physicians admit patients directly to the general medical unit after a telephone consultation with a hospitalist.

A subcategory of empirical research on the ED deals with developing metrics with which to address the issues of ED crowding. Solberg et al. (2003) provided an overview of the various metrics that have been proposed. We focus on two important measures in our study: length of stay (LOS), which measures total time in the ED from arrival to discharge/admit, and time to first treatment (TTFT), which measures the time from arrival to the first meaningful interaction with the physician.

Finally, a stream of empirical ED research involves time studies that characterize how caregivers spend their time in the ED, as well as the nature and duration of treatments. Examples of this type of research include Hollingsworth et al. (1998) and Graff et al. (1993). We will make use of these results to calibrate our models.

A number of researchers (e.g., Cochran and Roche 2009, Green et al. 2006, and Allon et al. 2012) have used queueing models to study various aspects of the ED. Within the large literature on queueing, studies that consider resource pooling, customer partitioning, or customer sequencing/prioritizing are most relevant to our work.

The standard insight from studies of pooling in queueing systems is that when two classes of customers in a queueing system become sufficiently different, pooling becomes ineffective and may even be harmful (see Mandelbaum and Reiman 1997, Tekin et al. 2009, Van Dijk and Van Der Sluis 2008). This suggests that a significant difference in treatment times between A- and D-type patients may be one way for streaming to overcome the antipooling disadvantage. However, verifying this requires an extension of known results because in the ED patient misclassification is inevitable, service is a complex process involving several physician-patient interactions, different streams of patients have different performance metrics, and the system has limited buffers (i.e., examination rooms/beds).
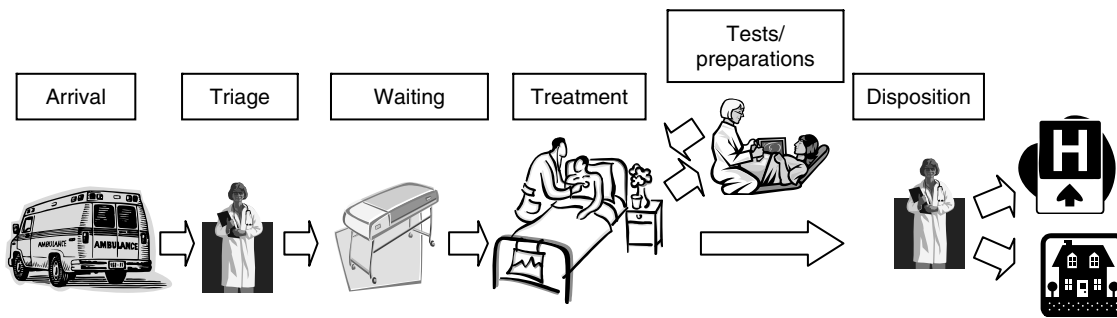
A related stream of queueing systems research considers effective ways of partitioning resources (e.g., Rothkopf and Rech 1987, Whitt 1999, Hu and Benjaafar 2009). An important observation from these studies is that separating fast and slow customers can protect customers with short processing times from waiting behind customers with long processing times. Note, however, that the same effect can be achieved by assigning priorities to customers with shorter processing times (Hu and Benjaafar 2009). However, for either partitioning or prioritizing to work effectively, we must be able to classify customers with a high level of accuracy. Analyses of priority queueing systems under misclassification errors (which are inevitable in EDs) suggest that these insights may not hold when classification is imperfect (e.g., Argon and Ziya 2009).

One last line of queueing research relevant to our work is the one that studies sequencing. In queueing systems where multiple customers are in the system at the same time (e.g., serial production lines with jobs at different stages of completion or an ED with multiple patients in the exam rooms awaiting physician attention), the server (physician) faces a customer sequencing problem. Related studies of serial systems can be found in Duenyas et al. (1998), Hopp et al. (2005), and Van Oyen et al. (2001), whereas related studies of parallel queueing systems can be found, for instance, in Andradóttir et al. (2003), Saghafian et al. (2011), and the references therein. In particular, Van Oyen et al. (2001) proposed a "pick-and-run" policy for servers in a serial system that favors working on the most downstream (old) jobs. We find that a similar policy can help physicians assigned to the D stream to choose their next patient in a manner that reduces average LOS.

3. Modeling Flows and Performance in the ED

To develop a modeling framework with which to address the *whether*, *where*, and *how* questions stated above, we must first describe the key characteristics of ED operations. We start by representing the general flow of patients in Figure 1. Patients arrive to the ED in a nonstationary, stochastic manner. Upon arrival, patients first go to the triage stage

Figure 1. The general flow of patients in an ED.

where each patient is assigned an emergency severity index (ESI), usually by a nurse, but sometimes by a doctor. ESI is an integer between 1 to 5, where clinical urgency decreases in ESI level. ESI 1 patients (who constitute a small percentage of total patient volume) are subject to high mortality risk if not treated immediately. Hence, they are always given high priority. As such, they are generally tracked separately from the rest of the patients through an "acute care" or "resuscitation" track. In American hospitals, ESI 4 and 5 patients are also often tracked separately through a "fast track" because their treatment needs are relatively simple and straightforward. Hence, in this paper, we focus on the ESI 2 and 3 patients who make up the bulk (about 80% at the University of Michigan) of the patients in the main ED.

In addition to assigning ESI levels, Flinders Medical Center has reported that, at the time of triage, nurses can predict whether a patient is *A* or *D* with roughly 80% accuracy (King et al. 2006). Empirical studies in other medical centers have reported similar results (e.g., Holdgate et al. 2007, Kronick and Desmond 2009).

After a patient has been triaged, he/she waits in a waiting area and is eventually called to an examination room. There he/she goes through one or more phases of interaction (treatment) with the same physician, as shown in Figure 1. (While caregivers may be nonphysicians—e.g., physician assistants—we use the term "physician" for simplicity.) Each physician-patient interaction (treatment stage) lasts a stochastic amount of time and is followed by testing (MRI, CT scan, etc.) or processing activities (e.g., wound cleaning) by a nurse that do not involve the physician. During testing or processing stages, which are also stochastic in duration, the patient is unavailable to the physician. The final processing stage after the last physician interaction is "disposition," in which the patient is either discharged or admitted to the hospital by staff based on the physician's final instructions.

Note that a patient is usually assigned to a single physician and so must wait for his/her physician to return for each treatment phase. Also, in most EDs, a patient is assigned to an exam room and holds that room, even when he/she is sent to a test facility, until he/she is disposed (discharged or admitted). Because physicians and exam rooms are limited, both of these resources can be bottlenecks.

The flow of patients in the ED is impacted by two phases of sequencing decisions. *Phase 1 sequencing* decisions determine the order/priority in which patients are initially taken from the waiting area to an examination room. Phase 1 decisions are usually made by a nurse in consideration of ESI levels and patient arrival orders. In theory, it could also make use of A/D predictions. Once patients are in examination rooms, *Phase 2 sequencing* is done to determine the order in which patients are seen. Individual physicians make the Phase 2 sequencing decisions by choosing the patients assigned to them in consideration of ESI levels, patient comfort, time in system, experience, etc. We have observed wide variance in the Phase 2 sequencing logic of individual physicians working within the same ED. Furthermore, physicians tend to limit the number of patients they have at any given time—seven seems to be a typical upper limit.

It is impossible to capture all of the abovementioned complexities of the ED in a single tractable analytic model. Of course, we can use simulation, but it is difficult to draw clear insights from purely numerical studies. Therefore, to probe the *whether*, *where*, and *how* questions, we will first examine a series of analytic models that represent simplified versions of the ED flow and then test the resulting conclusions under realistic conditions with a high-fidelity simulation calibrated with hospital data.

To compare streaming and pooling strategies, we must model the flows under each protocol. In a typical ED, which uses a pooling protocol, patients are not classified into A/D categories and all (ESI 2 and 3) patients are served by a set of pooled/shared resources (exam rooms, physicians, etc.), with priority given to ESI 2 patients. Under the streaming protocol, resources are divided into two groups: one for the *A* stream and one for the *D* stream, and A/D predictions are used to direct patients to the appropriate stream.

To compare the pooling and streaming protocols, we also need a performance criterion. Two commonly used metrics in the ED are length of stay (LOS) and time to first treatment (TTFT). For *D* patients, LOS is the key metric because it correlates with both convenience and safety (because a low LOS also guarantees a low TTFT). However, for *A* patients, LOS in the ED is usually a small fraction of their total LOS in the hospital, which on average

extends for days beyond their time in the ED. For these patients, safety is of much greater importance than amount of time they spend in the ED rather than in a hospital bed. Because safety is enhanced by starting treatment as soon as possible, TTFT is the most important metric for *A* patients.

We let α denote the percentage of *A* patients and define $T_A^\pi(\alpha)$ and $L_D^\pi(\alpha)$ to be the (average) TTFT of *A* patients and (average) LOS of *D* patients under policy $\pi \in \Pi$, respectively, where $\Pi = \{PA$(Pooling with priority to *A*s), *PD*(Pooling with priority to *D*s), *S*(Streaming)$\}$ represents the set of admissible policies. More specifically, letting *N* denote the total number of patients who visit the ED during a sufficiently long period (e.g., a year), we define $T_A^\pi(\alpha) = \mathbb{E}^\pi[(1/(\alpha N)) \sum_{i=1}^{\alpha N} T_{A,i}]$ and $L_D^\pi(\alpha) = \mathbb{E}^\pi[(1/((1-\alpha)N)) \sum_{i=1}^{(1-\alpha)N} L_{D,i}]$, where αN $((1-\alpha)N)$ is the number of *A*s (*D*s) during the period, $\mathbb{E}^\pi$ denotes expectation with respect to the probability measure defined by policy $\pi \in \Pi$, and $T_{A,i}$ and $L_{D,i}$ are random variables denoting the TTFT and LOS of the *i*th *A* and the *i*th *D* patient, respectively. Note that we are restricting attention only to pooling and streaming policies in keeping with the *whether* question raised in the Introduction. We acknowledge that a more complex state-dependent policy might outperform the policies in set Π. However, how much improvement is possible and whether such policies can be made practical in actual ED settings are open questions. In this paper, we restrict our attention to the potential for improvement through demonstrably implementable streaming policies.

To construct a single objective function, we let β represent the relative weight placed on the TTFT of *A* patients and define $f^\pi(\alpha, \beta) = \beta T_A^\pi(\alpha) + (1-\beta)L_D^\pi(\alpha)$ as the performance metric under policy $\pi \in \Pi$. We note that this performance metric can also be derived from a cost perspective. To see this, suppose c_A and c_D represent the per-patient cost of increasing the TTFT of *A* patients and LOS of *D* patients by one unit of time, respectively. If $\beta = (c_A \alpha)/(c_A \alpha + c_D(1-\alpha))$, then $f^\pi(\alpha, \beta)$ represents the average cost per patient under policy π. For instance, setting $\beta = \alpha$ implies an objective in which increasing TTFT of *A* patients and LOS of *D* patients by one minute is equally costly. We also note that although other metrics are used to evaluate the performance of an ED, most of these are highly correlated with our objective function. For example, the percentage of patients who leave without being seen (LWBS) is commonly tracked in EDs, but studies such as Fernandes et al. (1994) have indicated that the majority of such patients leave the ED because of prolonged waiting times. Hence, improvements in our objective function can be expected to result in reduced LWBS as well. We will examine the impact of streaming on LWBS in §6.

A closer look at the empirical results reported by Flinders (King et al. 2006) indicates that streaming reduced the LOS of *D* patients but increased TTFT of *A* patients. Hence, if streaming is attractive, it is because it strikes a better balance between these potentially conflicting objectives. Our combined objective enables us to examine this trade-off.

4. Phase 1 Implications of Streaming and Pooling

Realistic models of ED flow described in the previous section would be too complex for anything other than simulation. So, to get some clear insights into *whether*, *where*, and *how* streaming can outperform pooling, we start with a stylized patient flow model in which (1) all patients are available at the beginning of each day (i.e., static arrivals), (2) there are only two physicians, who work in parallel under the pooling protocol and are assigned to the *A* and *D* streams in the streaming protocol, (3) patient diagnosis/treatment occurs in a merged single service stage, where X_A (X_D) is a random variable with mean μ_A (μ_D) representing the service time of an *A* (*D*) patient, (4) *A/D* classification is perfect (i.e., error free), and (5) to avoid inefficient underutilization, the $A(D)$ physician switches to serve $D(A)$ patients when there is no other $A(D)$ patient available. Because we model service as a single stage, we eliminate the Phase 2 sequencing decisions. Hence, this model only offers insights into the performance of pooling and streaming via their impact on Phase 1 sequencing.

The above assumptions (most of which will be relaxed in subsequent sections) allow us to represent the ED with a *clearing* queueing model, in which a fixed number (*n*) of patients is available at the beginning of the day. Because the overall performance of the ED is heavily influenced by performance during periods of overload (which occur during predictably in the mid afternoon), the clearing model approximates ED behavior better than the more conventionally used steady-state queueing model.

We start by examining the relative effectiveness of the three policies in the admissible space Π for extreme cases where $\beta = 1$ or 0 (i.e., when the objective function is either merely TTFT for *A*s or LOS for *D*s).

PROPOSITION 1 (EXTREME CASES). *With* $\Pi = \{PA, PD, S\}$, *the following hold for the clearing model* (*with arbitrary distributions of* X_A *and* X_D):

(i) *For every* $\alpha \in [0, 1]$ *and every sample path* ω, $\arg\min_{\pi \in \Pi} T_A^\pi(\alpha, \omega) = PA$. *That is, if only TTFT of As matters* (*i.e., when* $\beta = 1$), *then pooling with priority to As is the best policy in* Π (*in the almost sure sense*).

(ii) *For every* $\alpha \in [0, 1]$ *and every sample path* ω, $\arg\min_{\pi \in \Pi} L_D^\pi(\alpha, \omega) = PD$. *That is, if only LOS of Ds matters* (*i.e., when* $\beta = 0$), *then pooling with priority to Ds is the best policy in* Π (*in the almost sure sense*).

This intuitive proposition suggests that streaming is not attractive unless we care about both TTFT for *A*s and LOS for *D*s. Therefore, we now analyze the optimal strategy when the objective function is a convex combination of these two metrics. To do this, we first formally define a *strategy* for our problem.

DEFINITION 1 (STRATEGY). A *strategy* is a map $\pi: [0, 1] \times [0, 1] \to \Pi$ that defines the policy $\pi(\alpha, \beta)$ for each α, β. An optimal strategy is the one that defines an optimal policy $\pi^*(\alpha, \beta)$ for every (α, β).

A useful property that allows us to establish the structure of the optimal strategy is β-convexity, which we define in two steps as follows.

DEFINITION 2 (π REGION). For policy $\pi \in \Pi$, the π region, denoted by $\mathscr{A}^\pi$, is the collection of (α, β) for which policy π is optimal. That is, the π region is $\mathscr{A}^\pi = \{(\alpha, \beta) \in [0, 1] \times [0, 1]: \pi^*(\alpha, \beta) = \pi\}$.

DEFINITION 3 (β-CONVEXITY). The optimal strategy $\pi^*: [0, 1] \times [0, 1] \to \Pi$ is said to be β-convex if all the π regions (i.e., sets $\mathscr{A}^\pi$ ($\forall \pi \in \Pi$)) are convex in β for every $\alpha \in [0, 1]$.
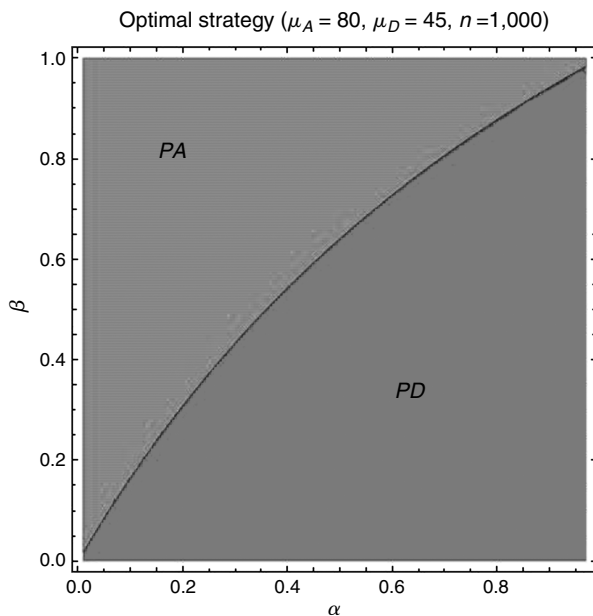
LEMMA 1 (β-CONVEXITY). *The optimal strategy $\pi^*(\alpha, \beta)$ is β-convex.*

Using the above lemma, we can establish the structure of the optimal strategy.

PROPOSITION 2 (DOUBLE THRESHOLD POLICY). *For every fixed $\alpha \in [0, 1]$, there exist double thresholds $\underline{\beta}(\alpha), \bar{\beta}(\alpha)$ such that streaming is the best policy in Π if, and only if, $\beta \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. If $\beta < \underline{\beta}(\alpha)$, then pooling with priority to Ds is the best policy in Π. If $\beta > \bar{\beta}(\alpha)$, then pooling with priority to As is the best policy in Π.*

Because EDs vary in their percentage of As (α) and relative weight of TTFT of Ds (β), the appeal of streaming depends on the width of the gap between $\underline{\beta}$ and $\bar{\beta}$. Unfortunately, our numerical experiments suggest that this gap is very narrow for the stylized model of this section. Indeed, Figure 2 illustrates an example with deterministic

Figure 2. An example of the optimal strategy with three admissible policies (PA, PD, S) and deterministic service times for which streaming is almost never optimal.

Optimal strategy ($\mu_A = 80$, $\mu_D = 45$, $n = 1,000$)

service times in which there is no region of optimality for streaming (it can, however, appear with stochastic service times). Although the optimality region for streaming can appear when service times are stochastic, it is generally small when α is constant and known. Knowing the exact proportion of As enables a fixed priority policy to strike an effective balance between the waiting costs of As and Ds.

This is no longer the case under the (highly realistic) assumption that α is uncertain. If the percentage of A patients varies from day to day, then a pooling policy that prioritizes either A or D patients can be quite ineffective. The reason is that we must pick which patient type to prioritize before the mix of A and D patients is known for the day. If we choose the wrong policy for the mix that actually occurs, performance could be very poor. We illustrate this in Figure 3, which plots the optimality gap (i.e., difference between the objective function of a given policy and that of the optimal policy) for the S, PA, and PD policies. These results show that whereas PA is optimal for small α, it is very poor for large α. Conversely, PD is optimal for large α and very poor for small α. In contrast, the streaming policy, S, is almost never optimal but is also never poor. Hence, we can make the following observation.
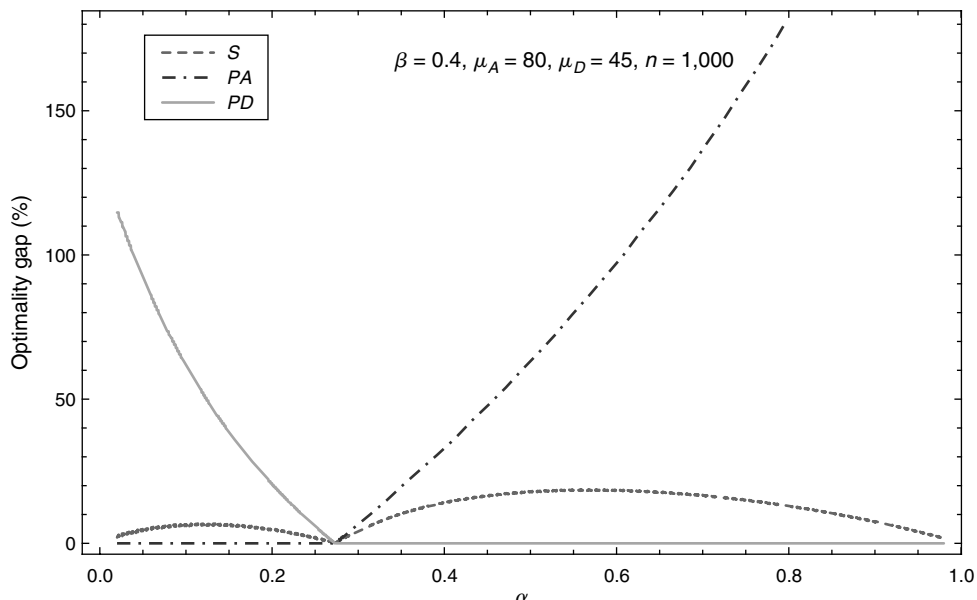
OBSERVATION 1. Streaming is much more *robust* to changes in patient mix (α) than is pooling.

The reason is that streaming mimics a dynamic policy with the simplicity of a static rule. By allocating some capacity to both patient types, it never results in a few patients of one type waiting for many patients of the other type.

To examine the impact of uncertainty in α, we assume α is chosen from a family of Beta distributions given by Beta($f(x), 2f(x)$), where $f(x) = (2 - 9x)/(27x)$, $x \in (0, 2/9)$. This results in $\mu_\alpha = 1/3$, which approximates the fraction of As in the University of Michigan Emergency Department (UMED), and $\sigma_\alpha^2 = x$, so we can generate a range of uncertainty of α by varying x. We choose the Beta distribution because (1) it is the most common distribution for a random variable that takes values between 0 and 1, and it includes the other well-known distribution, the uniform, as a special case, and (2) it seems to well represent our data from UMED. Figure 4 uses our analytical model of the ED along with the Beta distribution to illustrate the impact of varying σ_α^2 on the optimal strategy. This figure offers two insights: (1) As noted before, when there is no uncertainty ($\sigma_\alpha^2 = 0$), streaming is not optimal for any value of β. (2) As the level of uncertainty (measured by σ_α^2) increases, streaming becomes optimal for an increasingly broad range of β values.

From Figures 3 and 4, we can make an important conjecture (which we will test in §6): streaming is more robust than pooling to variation in patient mix. The intuition behind this robustness result is that a pooling system that completely prioritizes one type of patients can sequence them far from the optimal order (e.g., putting

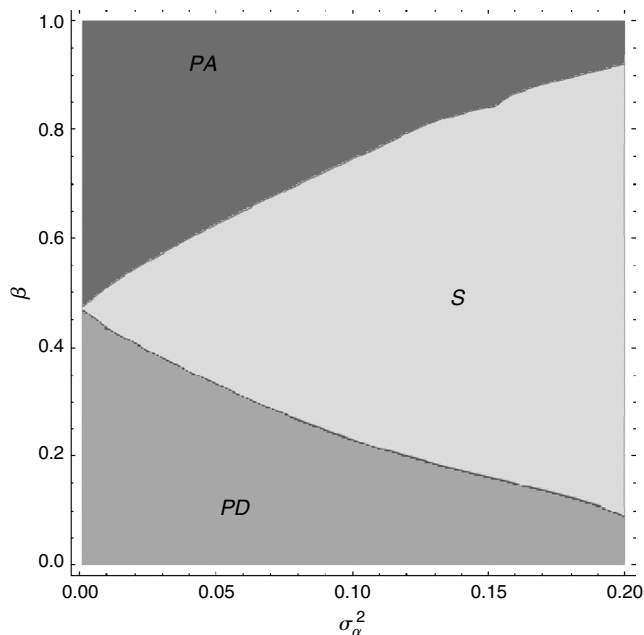**Figure 3.**   Sensitivity of policies to changes in α.

Note. Streaming is more robust to changes in patient mix than are the pooling policies.

D patients at the end of the line on a day in which they should have been at the beginning of the line). In contrast, a streaming system always gives some priority to both types of patients by "reserving" some capacity for each type. Although the proportion of capacity assigned to A and D patients may not be optimal on any given day (depending on the mix of patients), the fact that the two streams

Figure 4.   When the level of uncertainty in the percentage of A patients (measured by σ_α^2) increases, streaming becomes the optimal policy for an increasingly wide range of β values.

"back each other up" makes such suboptimalities much less disruptive than the "reverse prioritization" that can occur under pooling. Hence, altering the mix of patient types has a much more modest impact on performance in the streaming system. We relegate discussion of the model behind Figure 4 to Online Appendix B, in the electronic companion, for the sake of brevity. We will test another important conjecture that streaming is more robust than pooling to misclassification errors in §6.
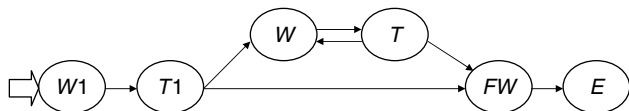
5. Phase 2 Implications of Streaming and Pooling

By modeling patient care as a single-stage service process, the above model focused attention exclusively on Phase 1 sequencing. However, as we noted earlier, ED patients typically receive multiple visits from physician (designated as "treatment" states), interspersed with tests, waiting for test results and intermediate processing (designated as "wait" states), during which the patient is not available for interaction with the physician. To examine the Phase 2 sequencing decisions of which patient to see next whenever a physician completes a treatment stage, we now relax the single-stage service assumption and consider a multistage treatment process. Note that we still face the Phase 1 sequencing decision concerning the order in which to bring the patients back into the examination rooms. In both Phase 1 and Phase 2 sequencing, we can make use of ESI information and, if available, A/D information. In Phase 2 sequencing, a physician can also potentially consider the number of past interactions with the patient. For instance, he/she could prioritize patients that have completed more treatment stages because they may be closer to completion.

Figure 5. Multistage ED service: ($W(1)$: (initial) wait; $T(1)$: (initial) treatment; FW: final wait; E: exit).

To explore the Phase 2 sequencing problem and its impact on the streaming versus pooling comparison, we consider a static arrival (clearing) model with two physicians, where one physician is assigned to each stream under the streaming protocol. However, we represent the service process by the multistage model in Figure 5. In this model, after an initial wait state labeled as $W1$, patients go through an initial treatment (direct or indirect interaction with the physician) labeled as state $T1$ (so TTFT is the time between the start of $T1$ and the arrival of the patient). After $T1$, the patient oscillates between a stochastic number of "wait" (labeled as W) and "treatment" (labeled as T) states. We note that the treatment states start only if both the physician and the patient are available and the physician elects to work on that patient. After the final treatment by the physician, the patient experiences a final wait state (labeled as FW) that involves final processing by a nurse and a delay specific to admission (e.g., assignment to a bed) or discharge (e.g., final paper work and follow-up instructions), and then the patient exits the ED (to state E). To allow the distribution of physician interactions per patient to match observed data, we let the probability of a transition to the final stage (FW) depend on the number of previous interactions.

Because our focus here is on Phase 2 sequencing, we simplify some other aspects of the system to construct a tractable model. First, without loss of generality, we consider a single ESI level for patients. We do this because, in a clearing system, all ESI 2 patients will be served before ESI 3 patients (due to their Phase 1 sequencing priority). Hence, distinguishing between these patient classes will have little effect on system performance. Second, to permit maximum opportunity for Phase 2 sequencing, we assume there are enough examination rooms to hold all of the patients. Third, we assume that times in "wait" states (i.e., times spent for tests, waiting for test results, and intermediate processing) are i.i.d. (independent and identically distributed) and exponentially distributed. For convenience, we also assume that times in the treatment states are i.i.d. and exponentially distributed and are independent of the duration of wait states. The i.i.d. assumption glosses over any queueing for test equipment or nurses that could serve to correlate the times in the wait states. However, because these states account for many different activities, we would not expect such correlation to be large. The exponential assumption reflects the unpredictability of the activities between physician interactions. Finally, to avoid the minor complications injected if preemption is disallowed,

we allow preemption. For instance, when a patient returns from a test, the physician has the option of preempting the current patient and switching to the returning patient. We will relax these assumptions in the next section.

Because A and D patients have different performance metrics, it makes sense to treat them differently in Phase 2 sequencing. For D patients, LOS matters most. The work of Van Oyen et al. (2001) (which considers a manufacturing system with multiple phases of worker/product interaction) suggests that a "pick-and-run" policy can be effective when the performance criterion is average time spent in the system. Under this policy, the goal is to serve the most downstream job. In the ED, the equivalent policy would be for physicians to work on the patient closest to completion and try to complete this "old" patient's service (to the extent possible) before initiating a service for a "new" patient. We refer to this policy as *Prioritize Old* (*PO*). In contrast, for A patients, TTFT is the key performance metric. Hence, for them, physicians should give preference to patients that have not yet been seen, unless constrained by the availability of exam rooms or the patient per physician limit. (Thus, in our simulation framework of the next section, where such additional constraints are also considered, a physician at his/her capacity should be directed to clear out in-process patients as quickly as possible by following the PO policy.) We refer to the policy that favors unseen patients as the *Prioritize New* (*PN*) policy.

We can show that these policies are optimal in the context of our simplified model. (See Online Appendix A for a proof, where a Markov decision process is developed to analyze the underlying multiarmed restless bandit model.) We also refer interested readers to Dobson et al. (2012) for some other related results regarding Phase 2 sequencing.

PROPOSITION 3 (WHO TO SEE NEXT?). *In the clearing model of a streaming ED flow with one physician assigned to each stream and multistage exponential treatment and wait stages modeled as in Figure* 5:

(i) *If the probability of completion increases in the number of previous physician-patient interactions, the Prioritize Old* (*PO*) *rule is optimal* (*in the expected sense*) *for the D stream.*

(ii) *The Prioritize New* (*PN*) *rule is optimal* (*in the almost sure sense*) *for the A stream.*

The implication of the above result is that instructing D physicians to work on the most downstream (old) patient and A physicians to work on the most upstream (new) patient should further improve the effectiveness of streaming. This addresses the *how* question we posed in §1 by suggesting a policy simple enough to be implemented in EDs. It also partially corresponds to what was done at Flinders (see King et al. 2006), where physicians assigned to the D stream were instructed that, in the absence of a threat to life/limb, need for time-critical intervention, or severe pain, they were to see patients in the order of arrival (i.e., a FCFS (first-come-first-served) mechanism).

Moreover, "the staff were further encouraged to attempt as far as possible to complete one patient's journey before bringing the next patient out of the waiting room into a cubicle" (King et al. 2006, p. 393). However, physicians assigned to the *A* stream were instructed to continue to prioritize patients according to ESI categories and to use FCFS within each category. Our results suggest that Flinders sequencing policies within the ED are reasonable but not optimal.

We will confirm the conjecture that implementing the *PO* and *PN* rules for Phase 2 sequencing in the ED can enhance the effectiveness of streaming in the next section.

6. A Simulation-Based Comparison of Streaming and Pooling

We now test the conjectures suggested by our simple analytic models by means of a detailed simulation model of the ED. This simulation incorporates many realistic features discussed earlier, including dynamic nonstationary arrivals, multistage service, multiple physicians and exam rooms, inaccuracy in disposition prediction, and bed-block by the hospital, among others. Our base-case model was calibrated using a year of data from UMED plus time study data from the literature. Below, we highlight key features of the model. A more detailed description of our modeling assumptions is presented in Online Appendix C.
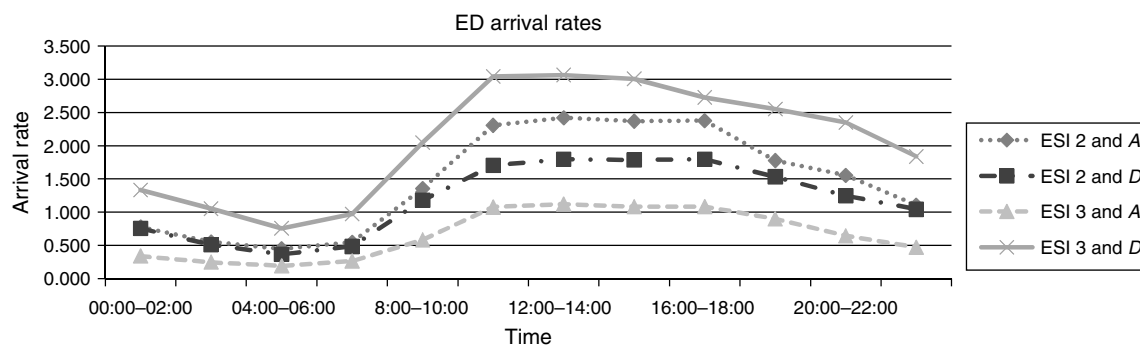
Patient Classes. As discussed earlier, patients are classified according to both ESI level (2 or 3) and ultimate disposition (*A* or *D*). This is done at the triage stage and results in patient classes 2*A*, 2*D*, 3*A*, and 3*D*. However, *A*/*D* prediction at triage is imperfect, resulting in misclassification errors. The true type of a patient is not revealed until the admit/discharge decision is made. Misclassification errors may vary from hospital to hospital, but achievable levels seem to be in the range of 20%–25% (King et al. 2006, Holdgate et al. 2007, Kronick and Desmond 2009).

Arrival Process. Arrivals for patient classes are modeled using nonstationary Poisson processes (which closely approximate the data) with arrival rates by class (obtained from a year of UMED data), depicted in Figure 6. The general pattern is similar to those reported in other studies (e.g., Green et al. 2006).

Service Process. The ED service process is depicted in Figure 5. Each patient goes through several phases of patient-physician interactions/treatment followed by tests and preparations. The duration of each interaction is random, and its average may depend on the class of the patient and the number of previous interactions. For instance, the first and last interactions are usually longer than intermediate ones. The number of interactions with a physician ranges from 1 to 7 and depends on the class of the patient, as well as several other factors. Based on the class of the patient, we draw the number of such interactions from a distribution constructed from a detailed time study (see Table 3 of Graff et al. 1993) after modifying the data to represent our four patient classes (see Online Appendix C for details). The simulated service process is noncollaborative (an ED physician rarely transfers his/her patients to another physician) and nonpreemptive (an ED physician rarely moves to another patient in the middle of his/her current interaction). The nonpreemptive framework rules out impractical policies that, for instance, instruct physicians to visit each patient for a short time and then move to the other patient before finishing the interaction with the current patient. Such preemptive policies are generally avoided by physicians because they are inefficient for the physician (who will need to re-review patient history and condition upon the next return), as well as irritating to patients.

Physician-Patient Assignments. As noted earlier, the process of connecting patients with physicians involves two phases. In Phase 1, patients are assigned to available exam rooms, usually by the charge nurse, based on a Phase 1 sequencing rule. In Phase 2, whenever a physician becomes available, he/she chooses the next patient (among those available/ready in the exam rooms) to see based on a Phase 2 sequencing rule. Under all patient flow designs, prioritizing ESI 2 patients over ESI 3 patients in Phase 1 is a constraint for safety reasons. For Phase 1 sequencing under streaming, patients are first streamed according to *A*/*D* information and then prioritized within streams with ESI 2 patients before ESI 3 patients (ties are broken with a FCFS rule). Under pooling, Phase 1 sequencing may or may not make use of *A*/*D* information, depending on the scenario under consideration. If *A*/*D* information is not available, then Phase 1 sequencing only considers

Figure 6. Class-dependent arrival rates to the ED for an average day (obtained from a year of data in UMED).

ESI levels by prioritizing ESI 2 over ESI 3 with an FCFS rule to break ties. If A/D information is available under pooling, then Phase 1 sequencing prioritizes patients in the following order: $2A, 2D, 3A, 3D$, with FCFS to break ties within a class.

In keeping with practice in UMED and elsewhere, we assume physicians do not take on more than seven patients at any time. We consider the following Phase 2 sequencing rules: (1) Service-In-Random-Order (SIRO), in which when a physician becomes available, s/he selects a patient at random from the pool of available (i.e., those not under a preparation or test) patients assigned to him and the new patients in the examination rooms waiting for a physician, provided that his/her total patient load does not exceed seven. This SIRO policy approximates current practice in many EDs in which physicians are not specifically encouraged to follow any specific rule, and hence, exogenous factors (changes in patient urgency level, patient discomfort, physician preference and experience, anticipation of interactions with testing facilities, access to newly available information, etc.) override systematic sequencing of patients. (We note, however, that although exogenous factors may make it appear that patients are sequenced according to SIRO, the decisions of physicians are not actually random. They are just based on criteria other than flow efficiency.) (2) First-Come-First-Served (FCFS), in which a physician selects his/her next patient in order of their arrival to the ED. This is an implementable policy to which many EDs aspire. (3) Prioritize-New-Prioritize-Old (PNPO), in which the prioritize new (PN) policy is used by physicians assigned to the A stream, and the prioritize old (PO) policy is used by physicians assigned to the D stream. That is, physicians in the A stream take an unassigned new patient whenever one is available in an exam room and the physician's patient load does not exceed seven. In contrast, physicians assigned to the D stream are instructed to prioritize the most downstream patient assigned to them, to free up rooms and minimize LOS by completing patient journeys as quickly as possible. If a physician is handling seven patients s/he is asked to serve the most downstream patient assigned to him regardless of the stream s/he is working in (in an effort to free up a room and lower his/her workload). Ties are always broken using a FCFS rule. Although new to EDs, PNPO is an implementable policy that our previous analytic results suggest should be effective.

Naming Convention. To distinguish between patient flow designs, we adopt a naming convention that labels each design as Protocol/Phase 1/Phase 2. "Protocol" designates the type of system: pooling (P), streaming (S), and virtual streaming (VS). The difference between the S and VS protocols is that S represents an implementation of streaming in which resources (rooms and physicians) are physically segregated and hence, idle resources assigned to one stream cannot be used by the patients of the other stream. In contrast, in VS, resources are only logically segregated and thus can be shared across streams.

The "Phase 1" and "Phase 2" parts in the naming convention designate the Phase 1 and 2 sequencing rules described earlier. Phase 1 sequencing under streaming is done by separating patients based on their ultimate disposition (A or D) and prioritizing each stream by ESI level (2 before 3). Hence, we label all S and VS cases with "$AD + ESI$" to indicate the Phase 1 rule. Similarly, for "Phase 1" under pooling, we use "ESI" to denote the case where Phase 1 sequencing is based only on ESI information, and we use "$AD + EDI$" to denote the case where, in addition to ESI levels, A/D information is used to sequence patients in the order: $2A, 2D, 3A, 3D$. For phase 2 sequencing rules, we use SIRO, FCFS, and PNPO. SIRO and FCFS can be used under either pooling or streaming, but PNPO can only be implemented in S and VS systems where physicians and patient classes are segregated into A and D streams. Table 1 summarizes this notation and the possible patient flow designs.

In the following subsections, we present our main findings from the simulation experiments. For each patient flow design described above, the objective function $(\beta TTFT(A) + (1 - \beta)LOS(D))$ is computed as an average over 5,000 replications of a week of operation, where the result for each replication is obtained after a warm-up period of one week. Further details about our simulation framework can be found in Online Appendix C.

6.1. ED Flow Design: Pooling, Physical Streaming, or Virtual Streaming?

We start with a comparison between the current practice of pooling in the EDs and physical streaming (where, unlike virtual streaming and our analytical clearing model, capacity sharing is not possible).

OBSERVATION 2. *Comparing simulations of the $S/AD + ESI/SIRO$ and $P/ESI/SIRO$ systems shows that pooling is more effective than physical streaming, with a 77% lower objective value.*

The inefficiency of physical streaming results from the imbalanced and low utilization of resources (which leads to intervals in which physicians are starved for lack of a patient or bed, even though a patient and/or bed is available in the opposite stream). In other words, physical streaming exhibits an "antipooling effect," which occurs because physical separation in either physicians or beds prevents capacity sharing. To place the observed magnitude of the antipooling effect of physical streaming (77%) in context, we make use of Kingman's formula for a $G/G/s$ queueing system with $s = 8$ physicians and two parallel $G/G/s$ queueing systems with $s = 4$ physicians each, with a server utilization matching our base case. The pooling benefit of having a $G/G/8$ queue versus two parallel $G/G/4$ queues on the average waiting time and the average system time is 79% and 7%, respectively. Because our objective function is a weighted average of TTFT (queue time) and LOS (system time), we would expect the antipooling penalty to

Table 1. Different patient flow designs under consideration and the notation implemented.

Protocol	Phase 1	Phase 2	Notation
Streaming (S)	ESI only (ESI)	Service in Random Order (SIRO)	$S/ESI/SIRO$
		First Come First Served (FCFS)	$S/ESI/FCFS$
		Prioritize New Prioritize Old (PNPO)	$S/ESI/PNPO$
	A/D Info and ESI ($AD + ESI$)	Service in Random Order (SIRO)	$S/AD + ESI/SIRO$
		First Come First Served (FCFS)	$S/AD + ESI/FCFS$
		Prioritize New Prioritize Old (PNPO)	$S/AD + ESI/PNPO$
Pooling (P)	ESI only (ESI)	Service in Random Order (SIRO)	$P/ESI/SIRO$
		First Come First Served (FCFS)	$P/ESI/FCFS$
	A/D Info and ESI ($AD + ESI$)	Service in Random Order (SIRO)	$P/AD + ESI/SIRO$
		First Come First Served (FCFS)	$P/AD + ESI/FCFS$
Virtual streaming (VS)	ESI only (ESI)	Service in Random Order (SIRO)	$VS/ESI/SIRO$
		First Come First Served (FCFS)	$VS/ESI/FCFS$
		Prioritize New Prioritize Old (PNPO)	$VS/ESI/PNPO$
	A/D Info and ESI ($AD + ESI$)	Service in Random Order (SIRO)	$VS/AD + ESI/SIRO$
		First Come First Served (FCFS)	$VS/AD + ESI/FCFS$
		Prioritize New Prioritize Old (PNPO)	$VS/AD + ESI/PNPO$

fall between these values, as it does. This simple example illustrates that even when capacity is perfectly balanced, the inability to share capacity between streams can be very damaging to performance. In the ED, this effect is particularly pronounced (i.e., toward the higher end of the range indicated by the $G/G/s$ model) because (1) it is not possible to balance utilization in the two streams exactly due to the discreteness of physicians and beds, and the fact that the average mix of A and D patients fluctuates according to the time of day (see Figure 7), and (2) the limited number of beds in the ED means that patients can be held in the waiting room even when physicians are idle, an effect that becomes more pronounced when beds are separated into two smaller systems under physical streaming. (The magnitude of this effect becomes apparent when we observe that the antipooling penalty falls to 17% in the simulation model when the number of beds is made infinite.) As a result, physical streaming is decidedly worse for performance than is a conventional pooling protocol. This leads us to suspect that Flinders does not rigidly adhere to a complete physical separation of streams, even though they described their system as such.

Because physical streaming is so unattractive, we do not consider it further, and instead we investigate whether virtual streaming (VS) can improve ED performance. We start by considering the SIRO Phase 2 sequencing rule (as an approximation of the status quo in most EDs) and compare $VS/AD + ESI/SIRO$ (basic virtual streaming) and $P/AD + ESI/SIRO$ (improved pooling) with $P/ESI/SIRO$ (current pooling practice in most EDs). Figure 7 depicts the simulation results. The graph on the left depicts the percentage improvement in the combined objective function (with $\beta = 0.5$). The graph on the right illustrates the improvement (in hours) achieved for each class of patients separately. The significant improvement shown in Figure 7 (left) is achieved because VS dramatically
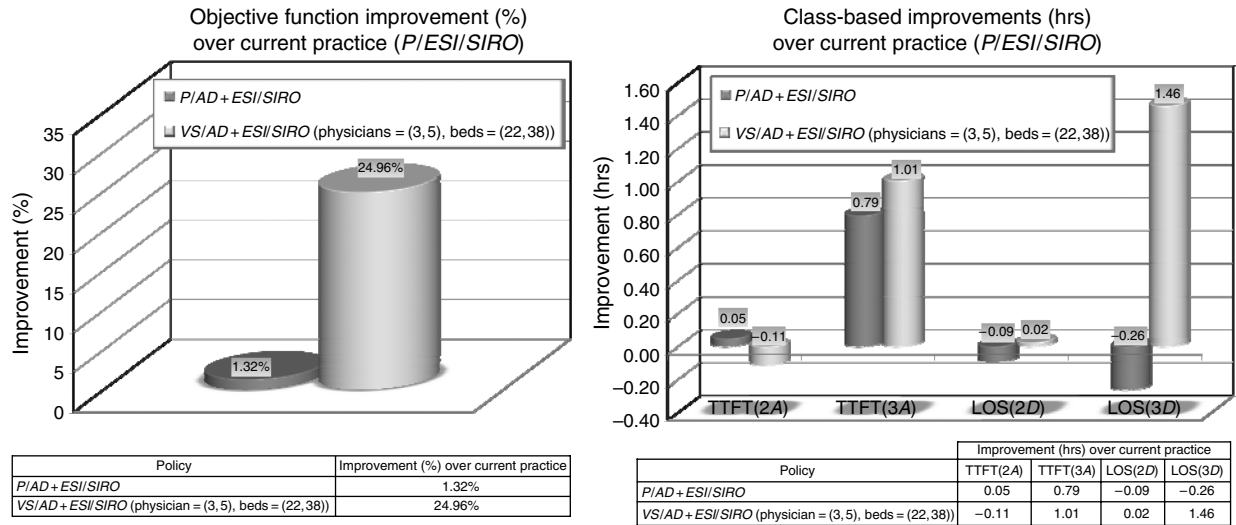
decreases LOS of $3D$s while only slightly increasing TTFT of $2A$s (see Figure 7 (right)).

OBSERVATION 3. Virtual streaming significantly outperforms both pooling and improved pooling by striking a better balance between TTFT of As and LOS of Ds.

Because VS does not require any physical reconfiguration of the ED, this finding provides strong evidence that virtual streaming can be an attractive and practical option for improving ED responsiveness. Because there is only a minor drop in performance for $2A$ patients (Figure 7 (right)), this attractiveness is also very robust to the weights assigned to our two main metrics, TTFT for As and LOS for Ds.

To further confirm this insight, we also compare the performance of the proposed virtual streaming ($VS/AD + ESI/PNPO$) with the current practice ($P/ESI/SIRO$) using all four metrics (i.e., TTFT and LOS for both As and Ds). Table 2 presents these four metrics in hours for our base-case scenario under pooling and streaming. To examine the robustness of streaming, we consider a weighted average of all these four metrics defined as $TTFT(A) + \beta_1 TTFT(D) + \beta_2 LOS(A) + \beta_3 LOS(D)$, where the weight for $TTFT(A)$ is assumed to be 1 and other weights represent the relative priorities of the remaining metrics to that of $TTFT(A)$. Our analysis reveals that pooling is only preferred in unrealistic cases where (a) almost no weight is placed on $LOS(D)$ (i.e., β_3 is small), (b) $LOS(A)$ is weighted more heavily than $TTFT(A)$ (i.e., $\beta_2 > 1$), and (c) $LOS(A)$ is more heavily weighted than $TTFT(D)$ (i.e., $\beta_2 > \beta_1$). Condition (a) is problematic, because (as we discussed previously) $LOS(D)$ is of great concern for EDs. Conditions (b) and (c) are particularly unrealistic because As remain in the hospital well beyond their stay in the ED, and hence, LOS in the ED is not that important for them. These provide further evidence that (1) the benefit of the proposed streaming policy (over the current pooling

Figure 7. Virtual streaming significantly outperforms pooling and improved pooling.

Policy	Improvement (%) over current practice
P/AD + ESI/SIRO	1.32%
VS/AD + ESI/SIRO (physician = (3, 5), beds = (22, 38))	24.96%

Policy	Improvement (hrs) over current practice			
	TTFT(2A)	TTFT(3A)	LOS(2D)	LOS(3D)
P/AD + ESI/SIRO	0.05	0.79	−0.09	−0.26
VS/AD + ESI/SIRO (physician = (3, 5), beds = (22, 38))	−0.11	1.01	0.02	1.46

Notes. The reason is that *VS* dramatically decreases LOS for 3*D* patients with only a minor increase in TTFT of 2*A* patients (results for an ED with eight physicians and 60 beds, a 20% misclassification error rate, and a weight for TTFT of *A* patients of $\beta = 0.50$).

policy) is robust with respect to weights assigned, and (2) considering an objective function made up of the two most important metrics, $TTFT(A)$ and $LOS(D)$, is a reasonable approximation of the full multiobjective optimization problem. Hence, for the reminder of of our analyses, we will make use of the two-dimensional objective function involving only $TTFT(A)$ and $LOS(D)$. However, it is worth noting that, based on the results presented in Table 2, we also expect the percentage of the left-without-being-seen (LWBS) metric to be improved by the proposed streaming design, because it improves the TTFT of both *A*s and *D*s.

Because patients who abandon the ED are not tracked in detail, we do not have enough data (e.g., how long they waited before leaving) to characterize the exact effect of streaming on LWBS. However, we can get an estimate using the following method. First, we assume that patients may leave after an exponentially distributed amount of time if they have not yet been seen. This is a reasonable

Table 2. Performance (in hours) of the proposed streaming design ($VS/AD + ESI/PNPO$) and current pooling practice ($P/ESI/SIRO$) under four metrics as well as the associated LWBS (%).

Policy	TTFT (A)	TTFT (D)	LOS (A)	LOS (D)	LWBS (%)
P/ESI/SIRO	0.88532	1.07893	7.4458	3.51401	3
VS/AD + ESI/PNPO (exogenous LWBS)	0.67253	0.95437	7.7389	2.60942	3
VS/AD + ESI/PNPO (endogenous LWBS)	0.74601	1.01349	7.8134	2.67707	1.4

Note. For the streaming design the physician and bed split have been optimized at physicians = (3, 8) and beds = (22, 38) for the *A* and *D* sides, respectively.

approximation of reality if there are multiple factors leading to a patient abandonment, each occurring according to a Poisson process. Under these conditions, the patient abandonment process is a superposition of Poisson processes that is itself Poisson. To estimate the rate of this process, we note that the current LWBS percentage in the UMED is 3%. Moreover, based on Table 2, the TTFT for an average patient (*A* or *D*) is about one hour. Thus, we need to find the exponential distribution that has a cdf equal to 0.03 at $TTFT = 1$. This leads to an exponential distribution with rate 0.031. Next, augmenting the arrival rates in the simulation by the current percentage of LWBS, 3%, and having patients leave after this exponential time, we observe that the LWBS (when made endogenous) under the streaming scenario is 1.04% compared to that of 3% in the current pooling system. Because the LWBS is reduced, the arrival rate to the ED is increased, which in turn slightly increases the TTFT relative to what it would be without the LWBS improvement. Nevertheless, streaming still significantly improves TTFT compared to current pooling practice in addition to achieving a significant reduction in the percentage LWBS. The bottom line is that streaming can reduce overall TTFT, LOS, and LWBS relative to pooling. However, as illustrated in Table 2, it does this by allowing a slight increase in LOS for *A* patients in order to achieve substantial improvements in all other metrics.

Having answered the *whether* question, we now seek to answer *how VS* should be implemented for maximum benefit. Proposition 3 suggests that following the PNPO rule for Phase 2 sequencing may further improve performance. Using our simulation test bed we observe that this conjecture is true. However, we also observe that improved Phase 2 sequencing does not make as large an improvement as that achieved by virtual streaming.

OBSERVATION 4. Using the PNPO rule for Phase 2 sequencing improves the performance, but performance of *VS* is relatively insensitive to the Phase 2 sequencing rule, indicating that most of the benefit of streaming is attributable to Phase 1.

The insensitivity of performance to Phase 2 sequencing is due to the fact that ED physicians frequently do not have many patients to choose among, because patients are often unavailable while waiting for test results. In EDs with shorter test times, higher physician utilization, and larger case loads (patients per physician), there would be more choice among in-process patients, and hence more benefit from an improved Phase 2 sequencing policy.

To get a sense of the maximum achievable value of the PNPO policy, we considered an ED with 50% shorter test times than UMED, as well as higher maximum case loads (12 versus 7) and very high dedicated utilization (up to 88% compared to 44% in the base case). "Dedicated utilization" refers to the fraction of the time that a physician is involved in activities that will not be interrupted to see another patient. These include direct care of patients and some indirect activities (e.g., reading patient test results). However, physicians also engage in many indirect activities (e.g., staff management, paper work, discussions with colleagues) that are preemptible and hence do not contribute to patient queueing. Studies report that direct care activities occupy 32% of ED physician time (Hollingsworth et al. 1998), so the 44% value for dedicated utilization in our base model is plausible. Of course, total ED physician utilization, which includes all direct and indirect activities, is much higher; ED physicians are busy. However, here we are only concerned with dedicated utilization, because this is what drives congestion.

The percentage improvement due to implementing the PNPO policy is shown in Figure 8 for a range of dedicated utilization values. This figure confirms that implementing PNPO becomes more effective when (1) the dedicated utilization of physicians is high, (2) the number of patients allowed per physician is large, and (3) patient test times are short. This suggests a practical limit of 4% on the amount of improvement possible via better Phase 2 sequencing. When combined with the benefit of virtual streaming, this results in a 29% improvement in the overall objective function compared to the current pooling practice ($P/ESI/SIRO$).
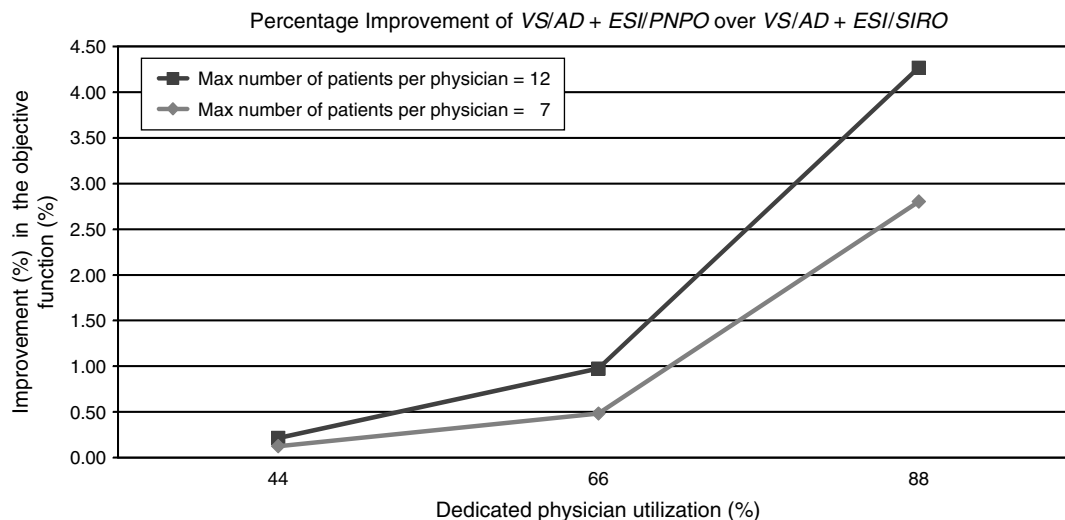
6.2. Sensitivity Analyses: Where to Implement Virtual Streaming?

Having addressed the *whether* and *how* questions we raised in §1, we now turn to the question of *where* virtual streaming is likely to be most attractive. We address this by performing sensitivity analyses on environmental characteristics to identify key factors that amplify the advantage of implementing virtual streaming over pooling.

To this end, in addition to using $V/AD + ESI/PNPO$ as a good candidate for virtual streaming, we select $P/AD + ESI/FCFS$ as a good candidate for pooling because (a) it makes use of A/D information in Phase 1 sequencing, and (b) FCFS is an implementable policy that was used at Flinders, and showed a small improvement over SIRO for Phase 2 sequencing in our simulation experiments. However, as we observed previously, the effect of a Phase 2 sequencing rule is small compared to the benefit obtained from virtual streaming, so we do not expect the results to be sensitive to the Phase 2 sequencing rule.
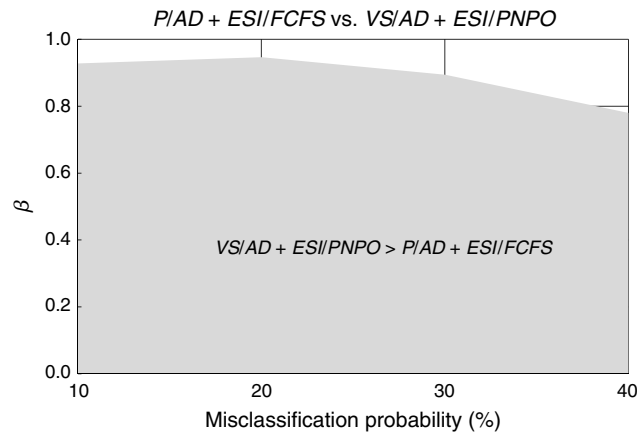
We start by examining the role of misclassification errors and β (the relative weight given to TTFT of A patients compared to LOS of D patients) on the relative benefit of virtual streaming over pooling. Based on our earlier clearing model, we conjectured that a higher β should favor

Figure 8. The benefit of implementing PNPO sequencing rule.

Percentage Improvement of *VS/AD + ESI/PNPO* over *VS/AD + ESI/SIRO*

Note. EDs with a higher physician utilization or with a higher maximum number of patients allowed per physician benefit more from PNPO.

Figure 9. Sensitivity of virtual streaming and pooling designs.

P/AD + ESI/FCFS vs. *VS/AD + ESI/PNPO*

VS/AD + ESI/PNPO > P/AD + ESI/FCFS

Misclassification probability (%)

Note. Lower weight on TTFT of *A* patients (β) or misclassification probability favors virtual streaming over pooling.

pooling. Common sense suggests that A/D information is less valuable if it is inaccurate, so we also expect a higher misclassification probability to also favor pooling. Figure 9 confirms these conjectures and shows that unless an ED gives an extremely heavy weight to the TTFT of *A* patients (high β) or has a very high misclassification error rate, virtual streaming is preferred to pooling.

Next we consider the effect of the percentage of *A* patients (α). Our analytical model in §4 led us to conjecture that a higher mean or a higher day-to-day variance in the percentage of *A* patients increases the attractiveness of virtual streaming. Figure 10 (left) shows simulation results indicating that virtual streaming is indeed more attractive in EDs with a higher percentage of *A* patients. Figure 10 (right) shows the effect of increasing day-to-day

variation in the mix of patients by drawing α from a family of beta distributions, Beta($f(x), 2f(x)$) where $f(x) = (2 - 9x)/(27x)$, $x \in (0, 2/9)$. Recall that doing this holds the mean at $\mu_\alpha = 1/3$ (which approximates UMED data), but allows the variance, $\sigma_\alpha^2 = x$, to range from 0 to 2/9. Figure 10 (right) indicates that higher variability in α also makes virtual streaming more attractive, as our analytic models predicted.

OBSERVATION 5. A higher fraction of *A* patients and a higher variance in the day-to-day fraction of *A* patients both favor (virtual) streaming relative to pooling.
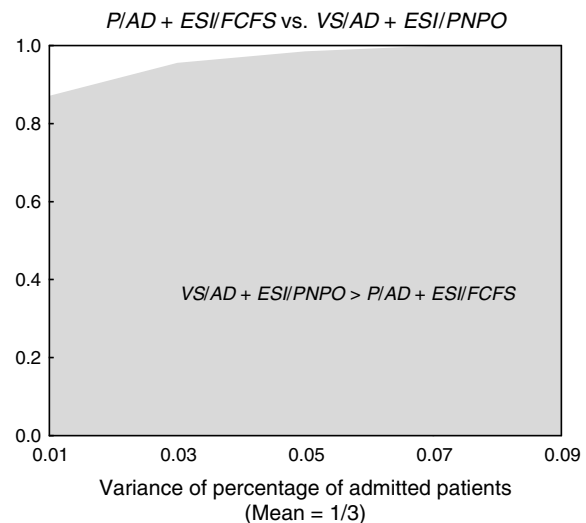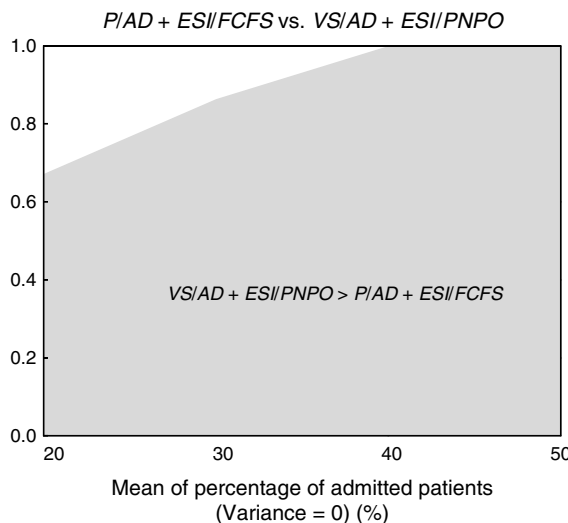
It is worth nothing that the percentage of *A* patients at Flinders is relatively high ($\alpha = 43\%$) compared to the average rate of admission in the U.S. EDs, which was $\alpha = 12.8\%$ in 2006 (Pitts et al. 2008). This may be one reason that streaming was considered a success at Flinders.

Another environmental factor that affects the (virtual) streaming versus pooling comparison is the relative test and treatment times of *A*s versus *D*s. In Figure 11, we examine the sensitivity of the $VS/AD + ESI/PNPO$ and $P/AD + ESI/FCFS$ configurations to increases in the test times of *A* (left) and *D* (right) patients. In Figure 12, we similarly consider the sensitivity of these two configurations to increases in the treatment times of *A* (left) and *D* (right) patients.

OBSERVATION 6. Increasing the difference between the test and/or treatment times of *A* and *D* patients increases the attractiveness of virtual streaming relative to pooling.
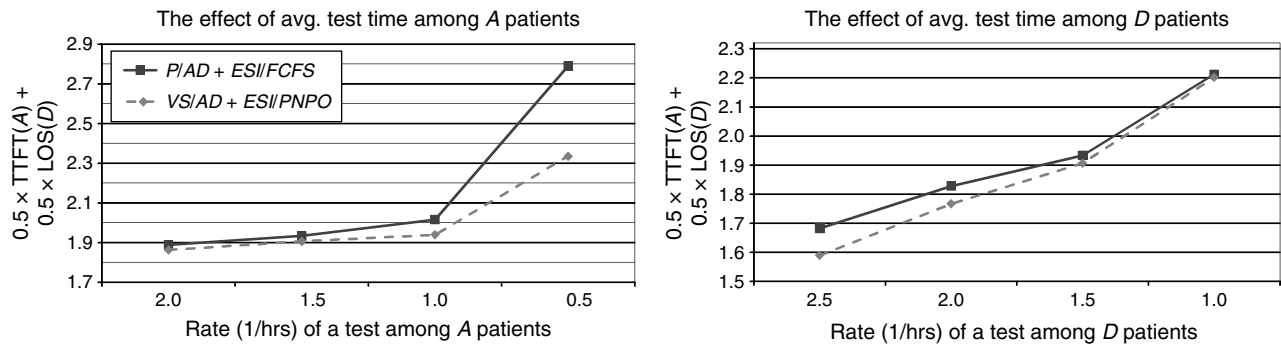
This observation has potentially important consequences for where virtual streaming is likely to be effective. First, EDs with congested or slow test facilities (which are used more frequently by *A*s than by *D*s) are likely to benefit more from virtual streaming than EDs with fast or ample

Figure 10. Sensitivity of virtual streaming (VS) and pooling (P) designs with respect to mean μ_α (left) and variance σ_α^2 (right) of the percentage of *A* patients.

P/AD + ESI/FCFS vs. *VS/AD + ESI/PNPO*

VS/AD + ESI/PNPO > P/AD + ESI/FCFS

Mean of percentage of admitted patients
(Variance = 0) (%)

P/AD + ESI/FCFS vs. *VS/AD + ESI/PNPO*

VS/AD + ESI/PNPO > P/AD + ESI/FCFS

Variance of percentage of admitted patients
(Mean = 1/3)

Note. VS dominates P in the shaded region.

Figure 11. The effect of average patient test time (MRI, CT scan, etc.) on the relative performance of two virtual streaming and pooling configurations.

Note. As test time for *A* patients increases (left) or decreases for *D* patients (right), virtual streaming becomes more attractive compared to pooling.

test facilities. Second, EDs that handle serious/complex patients among their *A*s (e.g., Level 1 trauma centers and teaching hospitals) are more likely to benefit from virtual streaming than EDs with less extreme *A*s (e.g., community hospitals), because the former is likely to have a larger gap between treatment times of *A*s and *D*s.

To further answer the *where* question, we consider the impact of a common phenomenon in EDs, the so-called "bed-block" process, which occurs when *A* patients are boarded in the ED while they wait for a hospital bed. Decreasing bed-block times has been shown to be one of the most significant factors (even more significant than increasing the number of beds) in reducing LOS (Khare et al. 2008). However, its impact on streaming has not been studied. Figure 13 compares the performance of the $VS/AD + ESI/PNPO$ and $P/AD + ESI/FCFS$ configurations for various values of the average boarding time of an *A* patient.

OBSERVATION 7. The relative attractiveness of virtual streaming over pooling increases with the average boarding time of *A* patients.

The implication is that EDs with higher frequency of bed-block or longer waits for hospital beds can benefit more from virtual streaming.

Finally, we consider the effect of the average dedicated utilization of physicians on the attractiveness of virtual streaming. Figure 14 (left) depicts the objective function for policies $VS/AD + ESI/PNPO$ and $P/AD + ESI/FCFS$, whereas Figure 14 (right) shows the improvement in the objective function from implementing $VS/AD + ESI/PNPO$ instead of $P/AD + ESI/FCFS$.
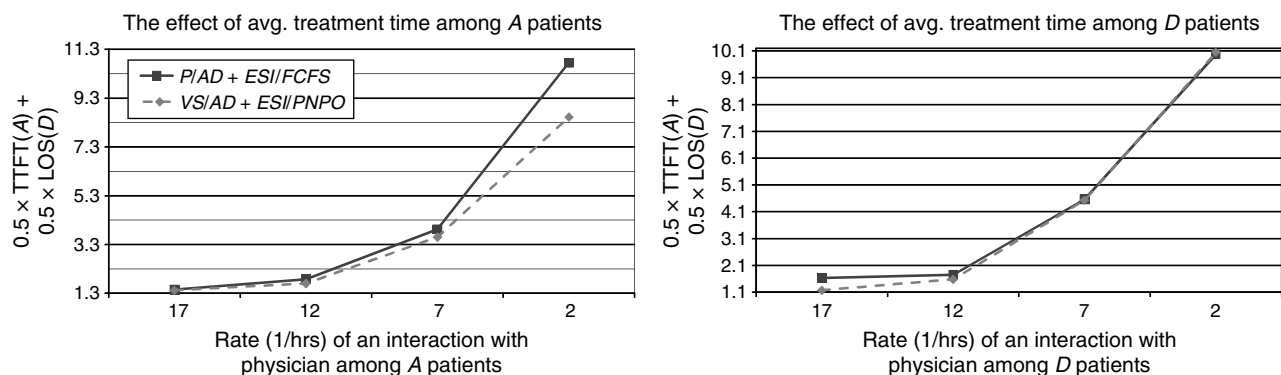
OBSERVATION 8. The relative attractiveness of virtual streaming over pooling increases with average dedicated utilization of physicians.

The implication is that congested EDs with high arrival rates or a low number of physicians can benefit more from virtual streaming. Furthermore, we did not explicitly account for physician interruptions, such as treating ESI-1 patients or dealing with other nonpatient issues, which would add to physician's nonpreemptible activities (and hence dedicated utilization). Thus, our estimates of the benefits of virtual streaming are probably conservative.
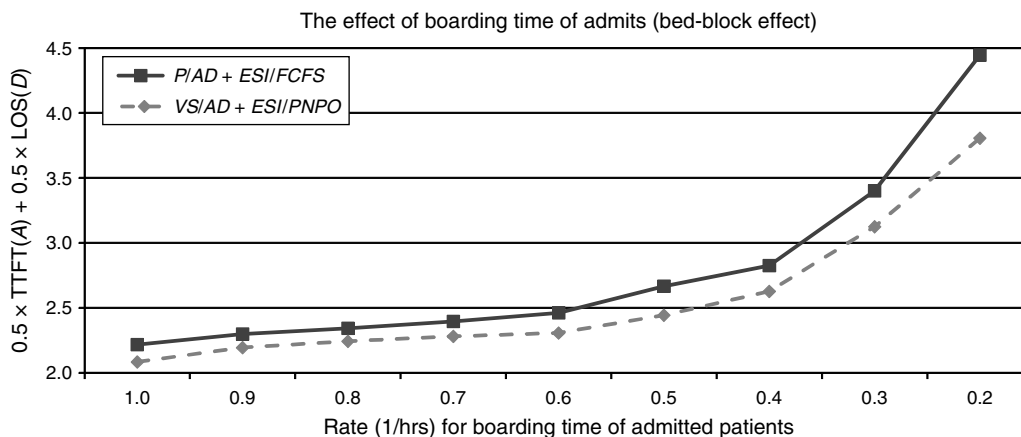
7. Conclusion

This paper describes our investigation of a new approach to managing patient flows in EDs: streaming, which separates patients based on an up-front prediction on their final

Figure 12. The effect of treatment times on the relative performance of two virtual streaming and pooling configurations.
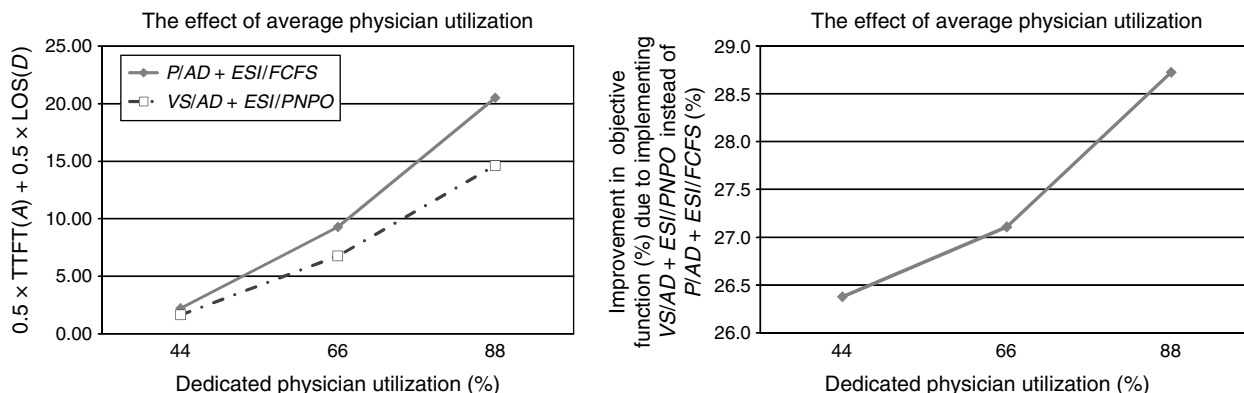

Note. As treatment time for *A* patients increases (left) or decreases for *D* patients (right), virtual streaming becomes more attractive compared to pooling.

Figure 13. The effect of the average boarding time on the performance of two virtual streaming and pooling configurations.

The effect of boarding time of admits (bed-block effect)

Note. EDs with longer boarding of *A*s benefit more from virtual streaming.

Figure 14. The effect of average physician utilization on the attractiveness of virtual streaming.

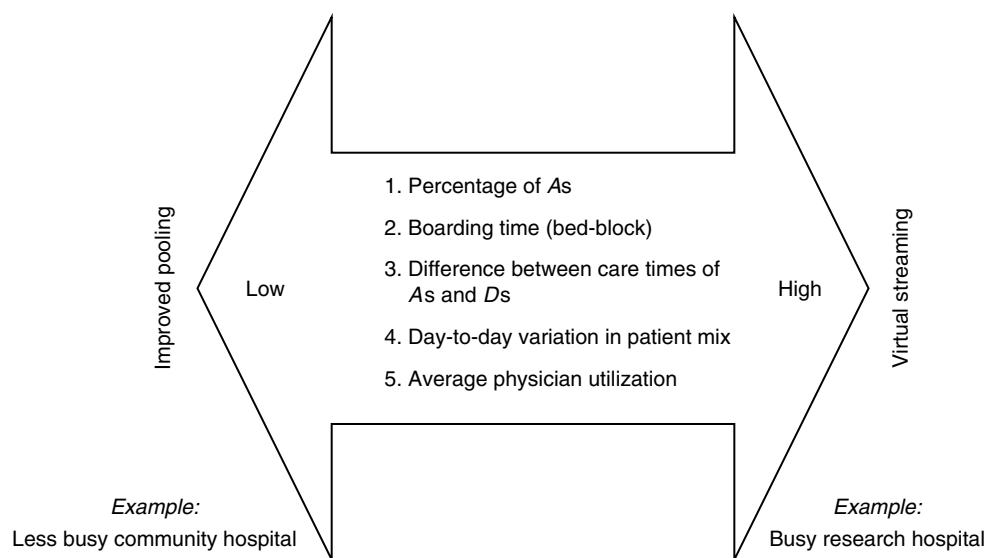*Note.* EDs with higher average physician utilization benefit more.

disposition (admission or discharge). Streaming has been popularized by Flinders Medical Center, where it has been credited with dramatically reducing patient length of stay (LOS). While the empirical results reported by Flinders have stimulated substantial interest among ED professionals, they are not conclusive because (1) the Flinders experiment was not a controlled study, so a Hawthorne effect cannot be ruled out, (2) other changes (e.g., lean) were implemented along with streaming, and (3) the environment at Flinders may not reflect other EDs (e.g., the fraction of *A* patients at Flinders is substantially above the norm). Indeed, our results suggest that the physical streaming approach as described by the Flinders may actually degrade ED performance because of an "antipooling" effect caused by separating resources into segments. Hence, we suspect that the Flinders success is partly due to informal capacity sharing to overcome the antipooling effect and partly due to other process improvements.

To avoid the antipooling effect of physical streaming, we proposed virtual streaming, in which physicians and rooms are only logically separated and, hence, excess capacities can be shared. Using simple analytical models, we found that virtual streaming can strike a better balance between the TTFT of *A* patients and the LOS of *D* patients by devoting some capacity to each patient type, rather than giving full priority to one. These analytic models also led to several conjectures about the environmental factors that should make virtual streaming more attractive.

We tested these conjectures with a realistic simulation and found that virtual streaming can indeed significantly improve ED performance (by 25% in a case designed to represent the ED of a busy academic hospital). Because implementing virtual streaming does not require a physical layout redesign in the ED, it provides a practical option to improve ED responsiveness.

We also found that the information used to stream patients (i.e., *A* or *D* classification) can be used by physicians to sequence patients within exam rooms and achieve additional performance improvements (up to 4% beyond the improvement due to virtual streaming alone). To achieve this, physicians assigned to the *A* stream should use (to the extent possible) a "prioritize new" rule that favors seeing new patients before finishing patients already in progress, while physicians assigned to the *D* stream

Figure 15. ED patient flow design strategy based on key environmental characteristics of the ED.

should use (to the extent possible) a "prioritize old" rule that favors completing patient journeys before initializing new ones.

Our results also indicate that while virtual streaming can be effective, it is not uniformly attractive to all EDs. Figure 15 summarizes the results of our sensitivity analyses, which suggest that virtual streaming is best suited for EDs with (1) a high percentage of *A* patients, (2) longer service times for *A*s than *D*s, (3) long patient boarding times due to bed-block, (4) high day-to-day variations in patient mix, and (5) high average physician utilization. Using a PNPO Phase 2 sequencing rule is more effective in EDs with (1) high average physician utilization, (2) large patient case load, and (3) short waits for test results.

In broad terms, our results indicate that better triage information about patients (e.g., A/D classification) can be leveraged to improve ED performance. One question to be answered in future research is whether other types of pre-treatment information (e.g., case complexity, type of testing required, etc.) are possible to obtain and yield additional benefit. Given the crisis levels of ED congestion, it is critical to find out.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1096.

Acknowledgments

References

American College of Emergency Physicians (2006) American College of Emergency Physicians national report card on the state of emergency medicine. http://www.acep.org/assets/0/16/648/1994/00FA9DFA-9B89-4DA8-A3D8-5FBD37DD858D.pdf.

Andradóttir S, Ayhan H, Down DG (2003) Dynamic server allocation for queueing networks with flexible servers. *Oper. Res.* 51(6):952–968.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management* 11(4):674–693.

Ben-Tovim DI, Bassham JE, Bennett DM, Dougherty ML, Martin MA, O'Neill SJ, Sincock JL, Szwarcbord MG (2008) Redesigning care at the Flinders Medical Centre: Clinical process redesign using "lean thinking." *Medical J. Australia* 188(6):27–31.

Cochran JK, Roche KT (2009) A multi-class queueing network analysis methodology for improving hospital emergency department performance. *Comput. Oper. Res.* 36(1):1497–1512.

Dobson G, Tezcan T, Tilson V (2012) Optimal workflow decisions for investigators in systems with interruptions. *Management Sci.* Forthcoming.

Duenyas I, Gupta D, Olsen T (1998) Control of a single-server tandem queueing system with setups. *Oper. Res.* 46(2):218–230.

Fernandes CM, Daya MR, Barry S, Palmer N (1994) Emergency department patients who leave without seeing a physician: The Toronto hospital experience. *Ann. Emergency Medicine* 24(6):1092–1096.

Gordon JA, Billings J, Asplin BR, Rhodes KV (2001) Safety net research in emergency medicine: Proceedings of the Academic Emergency Medicine Consensus Conference on "the unraveling safety net." *Acad. Emergency Medicine* 8(11):1024–1029.

Graff LG, Wolf S, Dinwoodie R, Buono D, Mucci D (1993) Emergency physician workload: A time study. *Ann. Emergency Medicine* 22(7):1156–1163.

Green LV, Soares J, Giglio JF, Green RA (2006) Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

Holdgate A, Morris J, Fry M, Zecevicl M (2007) Accuracy of triage nurses in predicting patient disposition. *Emergency Medicine Australasia* 19:341–345.

Hollingsworth JC, Chisholm CD, Giles BK, Cordell WH, Nelson DR (1998) How do physicians and nurses spend their time in the emergency department? *Ann. Emergency Medicine* 31(1):87–91.

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

Hopp WJ, Iravani SMR, Shou B (2005) Serial agile production systems with automation. *Oper. Res.* 53(5):852–866.

Howell EE, Bessman ES, Rubin HR (2004) Hospitalists and an innovative emergency department admission process. *J. General Internal Medicine* 19(3):266–268.

Hu B, Benjaafar S (2009) Partitioning of servers in queueing systems during rush hour. *Manufacturing Service Oper. Management* 11(3):416–428.

Khare RK, Powell ES, Reinhardt G, Lucenti M (2008) Adding more beds to the emergency department or reducing admitted patient boarding times: Which has a more significant influence on emergency department congestion? *Ann. Emergency Medicine* 53(5):575–585.

King DL, Ben-Tovim DI, Bassham J (2006) Redesigning emergency department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia* 18:391–397.

Kinsman L, Champion R, Lee G, Martin M, Masman K, May E, Mills T, Taylor MD, Thomas P, Williams RJ, Zalstein S (2008) Assessing the impact of streaming in a regional emergency department. *Emergency Medicine Australasia* 20:221–227.

Kronick SL, Desmond JS (2009) Blink: Accuracy of physician estimates of patient disposition at the time of ED triage. *SAEM Midwest Regional Meeting, Ann Arbor, MI*.

Liew D, Liew D, Kennedy MP (2003) Emergency department length of stay independently predicts excess inpatient length of stay. *Medical J. Australia* 179(17):524–526.

Mandelbaum A, Reiman MI (1997) On pooling in queueing networks. *Management Sci.* 44(7):971–981.

McCaig LF, Ly N (2002) National hospital ambulatory medical care survey: 2000 emergency department summary. *National Health Statist. Report*. Adv. Data 2001, 321, 1–31.

Miro O, Sanchez M, Espinosa G, Coll-Veient B, Bragulat E, Milla J (2003) Analysis of patient flow in the emergency department and the effect of an extensive reorganisation. *Emergency Medicine J.* 20:143–148.

Pitts SR, Niska RW, Xu J, Burt CW (2008) National hospital ambulatory medical care survey: 2006 emergency department summary. *Natl. Health Statist. Rep.* 7:1–39.

Richardson D (2003) Reducing patient times in emergency department. *Medical J. Australia* 179(17):516–517.

Richardson D (2006) Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medical J. Australia* 184:213–216.

Rothkopf MH, Rech P (1987) Perspective on queues: Combining queues is not always beneficial. *Oper. Res.* 35(6):906–909.

Saghafian S, Van Oyen MP, Kolfal B (2011) The "W" network and the dynamic control of unreliable flexible servers. *IIE Trans.* 43(12):893–907.

Schull MJ, Kiss A, Szali J-P (2007) The effect of low complexity patients on emergency department waiting times. *Ann. Emergency Medicine* 49(3):257–264.

Solberg LI, Asplin BR, Magid DJ, Weinick RM (2003) Emergency department crowding: Consensus development of potential measures. *Ann. Emergency Medicine* 42(6):824–834.

SoRelle R (2006) Homicide charges against Illinois ED stun EM. *Emergency Medicine News* 28(12):1, 35.

SoRelle R (2010) Breaking news: Health reform and the ED: Prepare for the surge. *Emergency Medicine News* 32(5):1, 20.

Sprivulis PC, Da Silva JA, Jacobs IG, Frazer RL, Jelinek GA (2006) The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Medical J. Australia* 184:208–212.

Tekin E, Hopp WJ, Van Oyen MP (2009) Pooling strategies for service center agent cross-training. *IIE Trans.* 41:546–561.

Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, Howard KM, Weiler PC, Brennan TA (2000) Incidence and type of adverse events and negligent care in Utah and Colorado. *Medical Care* 38(3):261–271.

Trzeciak S, Rivers EP (2003) Emergency department overcrowding in the United States: An emerging threat to patient safety and public health. *Emergency Medical J.* 20(5):402–405.

Van Dijk NM, Van Der Sluis E (2008) To pool or not to pool in call centers. *Production Oper. Management* 17(3):296–305.

Van Oyen MP, Gel EGS, Hopp WJ (2001) Opportunity for workforce agility in collaborative and non-collaborative work systems. *IIE Trans.* 33(9):761–777.

Welch SJ (2009) Patient segmentation: Redesigning flow. *Emergency Medicine News* 31(8).

Whitt W (1999) Partitioning customers into service groups. *Management Sci.* 45:1579–1592.

Soroush Saghafian is an assistant professor of industrial engineering at Arizona State University. His research focuses on the application and development of operations research methods in modeling and control of stochastic systems with specific applications in healthcare operations, control of queuing systems, and operations management. He has been awarded the 2012 Wilson Prize, 2011 University of Michigan College of Engineering Outstanding Ph.D. Research Award, 2010 Pierskalla Award (from the Healthcare Applications Society of INFORMS), 2010 POMS College of Healthcare Operations best student paper award (second place), the 2010 Murty Prize, and the 2007 IOE Bonder Fellowship award.

Wallace J. Hopp is the Alessi Professor and Associate Dean for Faculty and Research in the Ross School of Business at the University of Michigan. His current research interests center on manufacturing, healthcare, and innovation. He is a Fellow of INFORMS, MSOM, POMS, IIE, and SME.

Mark P. Van Oyen is an associate professor of industrial and operations engineering (IOE) at the University of Michigan. His core interests focus on the analysis, design, control, and management of operational systems, with emphasis on performance and flexibility. He also contributes to applied probability and the control and performance analysis of queueing networks. His awards include the 2010 Pierskalla Award, coauthoring first- and second-prize–winning papers in the 2011 POMS paper competition, IOE Department Faculty of the Year for 2008–2009, ALCOA Manufacturing Systems Faculty Fellow, and Researcher of the Year from Loyola University Chicago's School of Business.

Jeffrey S. Desmond is a clinical assistant professor and the service chief of the Department of Emergency Medicine at the University of Michigan Hospital.

Steven L. Kronick is a clinical assistant professor, the associate service chief, and the director of advanced cardiac life support of the Department of Emergency Medicine at the University of Michigan Hospital.

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments" by Soroush Saghafian, Wallace J. Hopp, Mark P. Van Oyen, Jeffrey S. Desmond, and Steven L. Kronick, *Operations Research*, http://dx.doi.org/10.1287/opre.1120.1096.

Online Appendix A: Proofs.

Proof of Proposition 1. We use a sample path argument. Consider the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Let $CA_k^\pi(\omega)$ and $CD_k^\pi(\omega)$ denote the completion time of kth Admit and kth Discharge type patient (under policy π and along sample path $\omega \in \Omega$), respectively. Also, assume $T_A^\pi(\alpha, \omega)$ and $L_D(\alpha, \omega)$ denote the (average) TTFT of Admits and the (average) LOS of Discharges for a given $\alpha \in [0,1]$ and sample path $\omega \in \Omega$, respectively.

Proof of Part (i). To prove part (i), it is sufficient to show that for every α and every sample path ω: (a) $T_A^{PA}(\alpha, \omega) \le T_A^S(\alpha, \omega)$, and (b) $T_A^{PA}(\alpha, \omega) \le T_A^{PD}(\alpha, \omega)$. To prove (a), fix α and let $t(\omega) = \min\{CA_{n_A}^S(\omega), CD_{n_D}^S(\omega)\}$ denote the time that system moves to a pooling scenario under Streaming policy and over sample path ω. If $t(\omega) = CA_{n_A}^S(\omega)$ (i.e., if Streaming becomes Pooling when Admits are all served) then notice that under $\pi = S$, the kth Admit patient starts its treatment at $CA_{k-1}^S(\omega)$ but under $\pi = PA$, the kth Admit patient starts its treatment at $\min\{CA_{k-1}^{PA}(\omega), CA_{k-2}^{PA}(\omega)\} \le \min\{CA_{k-1}^S(\omega), CA_{k-2}^S(\omega)\} \le CA_{k-1}^S(\omega)$, where the first inequality can be easily shown using induction on k, and the second inequality trivially holds. Hence, under $\pi = PA$ each patient is seen no later than when s/he is seen under $\pi = S$, and therefore (a) holds. Now if $t(\omega) = CD_{n_D}^S(\omega)$ (i.e., if Streaming becomes Pooling when some Admits still have not been seen), assume the last Admit type patient that has been seen before or at time $t(\omega)$ under $\pi = S$ is the $n_t(\omega)$th patients of this type. Using the previous argument, none of first $n_t(\omega)$ patients under $\pi = S$ are seen before the time they would have been seen under $\pi = PA$. Moreover, under $\pi = S$ every remaining Admit patient is seen with a constant delay of at least $t(\omega) - CA_{n_t(\omega)-1}^{PA}(\omega) \ge 0$ compared to what it would have been seen under $\pi = PA$. Therefore, for every ω and every α, every Admit type patient is seen under $\pi = S$ no sooner than what it would have been seen under $\pi = PA$. Thus (a) holds. To show (b), fix α and notice that under $\pi = PD$ every Admit patient is seen with a constant delay of at least $CD_{n_D-1}^{PD}(\omega)$ compared to what it would have been seen under $\pi = PA$. Thus, (b) holds and the proof of (i) is complete.

Proof of Part (ii). To prove part (ii), it is sufficient to show that for every α and every sample path ω: (1) $L_D^{PD}(\alpha, \omega) \le L_D^S(\alpha, \omega)$, and (2) $L_D^{PD}(\alpha, \omega) \le L_D^{PA}(\alpha, \omega)$. To show (1), fixing α, we show

that $CD_k^{PD}(\omega) \leq CD_k^S(\omega)$ ($\forall k \in 1, 2, \cdots, n_D$). To show this notice that using the same argument as part (i) (and after swapping labels D and A) it is easy to show that TTFT of each Discharge patient under $\pi = PD$ is no more than its TTFT under $\pi = S$. That is, if $TD_k^\pi(\omega)$ denotes the TTFT of the kth Discharge patient under sample path ω, then $TD_k^{PD}(\omega) \leq TD_k^S(\omega)$ ($\forall k \in 1, 2, \cdots, n_D$). Next, if $SD_k(\omega)$ is the service time of kth Discharge patient under sample path ω, $CD_k^\pi(\omega) = TD_k^\pi(\omega) + SD_k(\omega)$. Thus, since $TD_k^{PD}(\omega) \leq TD_k^S(\omega)$, we have $CD_k^{PD}(\omega) \leq CD_k^S(\omega)$ ($\forall k \in 1, 2, \cdots, n_D$), and hence (1) holds. To show (2), fix α and notice that the completion time of every Discharge patient under PA is delayed at least for $CD_{n_A-1}^{PA}$ units of time compared to PD, and hence, the proof is complete. $\qquad\square$

<u>Proof of Lemma 1.</u> To prove this lemma, using the definition of β-convexity, we need to show that sets $\mathcal{A}^\pi$ ($\forall \pi \in \mathbf{\Pi}$) are convex in β for every α. Fix α and consider β_1 and β_2 such that $(\alpha, \beta_1) \in \mathcal{A}^\pi$ and $(\alpha, \beta_2) \in \mathcal{A}^\pi$. We then need to show that $(\alpha, \gamma\beta_1 + (1-\gamma)\beta_2) \in \mathcal{A}^\pi$ for every $\gamma \in [0, 1]$. Notice that as $(\alpha, \beta_1) \in \mathcal{A}^\pi$, for every other policy $\pi' \in \mathbf{\Pi}$ we have:

$$\beta_1 T_A^\pi(\alpha) + (1-\beta_1)L_D^\pi(\alpha) \leq \beta_1 T_A^{\pi'}(\alpha) + (1-\beta_1)L_D^{\pi'}(\alpha). \tag{EC.1}$$

Similarly, as $(\alpha, \beta_2) \in \mathcal{A}^\pi$, for every other policy $\pi' \in \mathbf{\Pi}$ we have:

$$\beta_2 T_A^\pi(\alpha) + (1-\beta_2)L_D^\pi(\alpha) \leq \beta_2 T_A^{\pi'}(\alpha) + (1-\beta_2)L_D^{\pi'}(\alpha). \tag{EC.2}$$

Now multiplying both sides of (EC.1) by γ and both sides of (EC.2) by $(1-\gamma)$ and adding up the resulting inequalities we get:

$$(\gamma\beta_1 + (1-\gamma)\beta_2) T_A^\pi(\alpha) + (1 - [\gamma\beta_1 + (1-\gamma)\beta_2])L_D^\pi(\alpha)$$
$$\leq (\gamma\beta_1 + (1-\gamma)\beta_2) T_A^{\pi'}(\alpha) + (1 - [\gamma\beta_1 + (1-\gamma)\beta_2])L_D^{\pi'}(\alpha).$$

Hence, since the above inequality holds for every $\pi' \in \mathbf{\Pi}$ and every $\gamma \in [0, 1]$, $(\alpha, \gamma\beta_1 + (1-\gamma)\beta_2) \in \mathcal{A}^\pi$ for every $\gamma \in [0, 1]$. Thus, the optimal strategy $\pi^*(\alpha, \beta)$ is convex in β. $\qquad\square$

<u>Proof of Proposition 2.</u> Define functions $\beta_1(\alpha)$ and $\beta_2(\alpha)$ as follows:

$$\beta_1(\alpha) = \inf\{\beta : f^S(\alpha, \beta) \leq f^{PD}(\alpha, \beta)\},$$

$$\beta_2(\alpha) = \sup\{\beta : f^S(\alpha, \beta) \leq f^{PA}(\alpha, \beta)\}.$$

We show that by setting $\underline{\beta}(\alpha) = \min\{\beta_1(\alpha), \beta_2(\alpha)\}$ and $\bar{\beta}(\alpha) = \max\{\beta_1(\alpha), \beta_2(\alpha)\}$, Streaming is optimal for a given α if, and only if, $\beta(\alpha) \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. To see the "if" part, fix α, suppose $\beta(\alpha) \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$, and write $\beta(\alpha)$ as a convex combination of extreme points $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$. Then notice that by definition of $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$, Streaming is optimal at both extreme points $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$. Hence, by Lemma 1 Streaming is also optimal at $\beta(\alpha)$. To see the "only if" part, fix α and suppose $\beta(\alpha) \notin [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. That is, suppose for some $\epsilon > 0$ either (a) $0 \le \beta(\alpha) \le \underline{\beta}(\alpha) - \epsilon$, or (b) $\bar{\beta}(\alpha) + \epsilon \le \beta(\alpha) \le 1$. If (a) holds, write $\beta(\alpha)$ as a convex combination of $\tilde{\beta}(\alpha) = 0$ and $\underline{\beta}(\alpha) - \epsilon$. Then notice that, from Proposition 1, $\pi = PD$ is optimal at $\tilde{\beta}(\alpha) = 0$. Also, $\underline{\beta}(\alpha) - \epsilon < \underline{\beta}(\alpha) \le \beta_1(\alpha)$. Therefore, from the definition of $\beta_1(\alpha)$, $\pi = PD$ is better than $\pi = S$ at $\underline{\beta}(\alpha) - \epsilon$. Moreover, $\pi = PA$ cannot be optimal at $\underline{\beta}(\alpha) - \epsilon$, since otherwise, choosing a β in $[\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$ and writing that as a convex combination of $\tilde{\tilde{\beta}}(\alpha) = 1$ (for which $\pi = PA$ is optimal by Proposition 1) and $\underline{\beta}(\alpha) - \epsilon$ will result in a contradiction. Thus, $\pi = PD$ is optimal at both extreme points $\tilde{\beta}(\alpha) = 0$ and $\underline{\beta}(\alpha) - \epsilon$. Hence, $\pi = PD$ is also optimal at their convex combination, $\beta(\alpha)$, by Lemma 1. If, on the other hand, (b) holds, write $\beta(\alpha)$ as a convex combination of $\overline{\beta}(\alpha) + \epsilon$ and $\tilde{\tilde{\beta}}(\alpha) = 1$. Then, similar to the discussion of part (a), notice that by definition of $\beta_2(\alpha)$, $\pi = PA$ is optimal at $\overline{\beta}(\alpha) + \epsilon$. Moreover, by Proposition 1, $\pi = PA$ is also optimal at $\tilde{\tilde{\beta}}(\alpha) = 1$. Thus, from Lemma 1 we see that $\pi = PA$ should be also optimal at $\beta(\alpha)$. This completes the proof. $\square$

<u>Proof of Proposition 3 - Part (i).</u> We develop a *Markov Decision Process (MDP)* model to show the optimality in the expected sense. It should be noted that the underlying problem is in the class of *multi-armed restless bandit problems*, which are usually hard to analyze. Since beds are not limited (e.g., larger than the number of patients in the clearing model), suppose, without loss of generality, that at the beginning all patients are in state W_1, i.e., in the initial waiting state depicted in Figure 5. The ith waiting stage, W_i, is followed by a treatment stage, T_i. The duration of waiting stages and treatment stages are independent of each other and exponentially distributed with rates denoted by γ and μ, respectively. Suppose the maximum number of interactions with the physician is denoted by $\bar{k}$, and $W_{\bar{k}+1}$ denotes the final nurse visit before disposition (i.e., stage FW in Figure 5). For the ease of notation, we also assume stage $T_{\bar{k}+1}$ represents the disposition stage. That is, we assume every patients who leaves the ED goes to (absorbing) stage $T_{\bar{k}+1}$. The LOS of

a patient in our clearing model is then equal to the time that s/he leaves stage $W_{\bar{k}+1}$ to enter $T_{\bar{k}+1}$.

Let p_k denote the probability that a patient who is in treatment stage k, T_k, is having its final treatment by the physician and will go to the final treatment by nurse, $W_{\bar{k}+1}$, afterwards. Assume p_k is increasing in k (that is being in a higher treatment stage is associated with a higher chance of being in the final treatment stage) and $p_{\bar{k}} = 1$. The state of the system then can be represented by $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} = (x_1, x_2, \cdots, x_{\bar{k}+1})$ and $\mathbf{Y} = (y_1, y_2, \cdots, y_{\bar{k}+1})$, where x_i and y_i denote the number of patients in ith stage of treatment and wait (T_i and W_i), respectively. Let N denote the total number of patients at time 0. The goal is to dynamically control the location of the physician, denoted by l, to go from state $(N, 0, \cdots, 0)$ to state $(0, 0, \cdots, N)$ with the minimum expected average LOS or equivalently with the minimum sum of patient completion times. Now, using uniformization with rate $\psi = N\gamma + \mu < \infty$, we can consider the discrete time version of the problem (where the times between consecutive events are i.i.d and exponentially distributed with rate ψ). Doing so and denoting the optimal remaining cost when the system is at state $(\mathbf{X}, \mathbf{Y})$ with $J(\mathbf{X}, \mathbf{Y})$, we have the following optimality equation (with the terminal condition $J(0, 0, \cdots, N) = 0$):

$$
\begin{aligned}
J(\mathbf{X}, \mathbf{Y}) = \frac{1}{\psi} \Big[&\sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \\
&+ \mu \min_{l \in \mathcal{L}(\mathbf{x})} \Big\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l = k\} \big[p_k J(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \big] \Big\} \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i J(\mathbf{X} + e_i, \mathbf{Y} - e_i) \\
&+ \big(\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\{ \sum_{ki=1}^{\bar{k}} x_i \geq 1 \} \big) J(\mathbf{X}, \mathbf{Y}) \Big],
\end{aligned}
\tag{EC.3}
$$

where e_k is a row vector of size $\bar{k} + 1$ with a one in kth element and zero everywhere else, and $\mathcal{L}(\mathbf{X}) = \{ i \leq \bar{k} : x_i \geq 1 \}$ is the set of possible locations to allocate the physician when $\mathbf{X}$ is the first part of the state. The first line in the above optimality equation represents the current cost (every patient's completion time who is still in the ED is delayed for one unit of uniformized time). The second line is the event related to treating a patient by the physician. The third line represents the event that a patient moves from a waiting stage to a treatment stage, and the fourth line represents the self-loop event. (Notice that since preemption is allowed, using a sample path argument, it can be easily shown that forced idling is suboptimal. Therefore, without loss of generality the term

in the self-loop with coefficient μ is independent of the control action, l.) Also, a finite horizon version of the above MDP can be considered using the following optimality equation with terminal condition $J_0(\mathbf{X}, \mathbf{Y}) = 0$ for every state $(\mathbf{X}, \mathbf{Y})$ and $n \in \mathbb{N}$:

$$
\begin{aligned}
J_{n+1}(\mathbf{X}, \mathbf{Y}) = \frac{1}{\psi} \Big[&\sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \\
&+ \mu \min_{l \in \mathcal{L}(\mathbf{x})} \Big\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l = k\} \big[p_k J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \big] \Big\} \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i J_n(\mathbf{X} + e_i, \mathbf{Y} - e_i) \\
&+ \big(\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \} \big) J_n(\mathbf{X}, \mathbf{Y}) \Big],
\end{aligned}
\tag{EC.4}
$$

where $J_n(\mathbf{X}, \mathbf{Y})$ denotes the optimal remaining cost when the state is $(\mathbf{X}, \mathbf{Y})$ and there are n periods to go. (Notice that $J_n(\mathbf{X}, \mathbf{Y}) \to J(\mathbf{X}, \mathbf{Y})$ as $n \to \infty$ since there is an absorbing state.) To show that the PO policy which prescribes serving the "old" patient in the most downstream stage is optimal, we use induction on n. First notice that for $n = 1$ all policies are the same considering the minimization in (EC.4), since $J_0(\mathbf{X}, \mathbf{Y}) = 0$ for every state $(\mathbf{X}, \mathbf{Y})$. Now, suppose it is optimal to follow PO policy at any state when in period n. We show that it is optimal to follow PO at any state in period $n + 1$ as well. To this end, consider period $n + 1$ and an arbitrary state $(\mathbf{X}, \mathbf{Y})$. Suppose in state $(\mathbf{X}, \mathbf{Y})$ treatment stage k^* is the the most downstream stage with an available patient. To show that allocating the physician to stage $1 \leq k^* \leq \bar{k}$ is optimal in $n + 1$, suppose there is also another stage $k < k^*$ with an available patient at state $(\mathbf{X}, \mathbf{Y})$ (i.e., with $x_k \geq 1$ and $x_{k^*} \geq 1$). Then considering the minimization in (EC.4), to show that serving stage k^* in period $n + 1$ is optimal, it is sufficient to show that for any such k, we have:

Property i: $\quad p_{k^*} J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*}) J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1})$

$$\leq p_k J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}). \tag{EC.5}$$

We show the above property of the optimal cost function along with the following property:

Property ii: $\quad p_k^* J_n(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_k^*) J_n(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y})$

$$p_k J_n(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_n(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}). \tag{EC.6}$$

In other words, we assume Properties i and ii hold for $n-1$, and show that they both hold for n as well. First, we show Property i. To do so, we build an upper bound for the LHS of (EC.5) using suboptimal actions and show that this upper bound is less than the RHS of this inequality. The upper bound for the LHS can be obtained by suboptimally allocating the physician to treatment stage k in period n and then following the optimal policy (i.e., PO) in the remaining periods. To this end, consider state $(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1})$ in period n and use the suboptimal but feasible (since $x_k \geq 1$) action $l = k$ to obtain an upper bound for $J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1})$. Doing so we have:

$$
\begin{aligned}
J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \leq \frac{1}{\psi} \Big[& \big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
& + \mu \big[p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1}) \\
& \qquad + (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \big] \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{\bar{k}+1}) \\
& + \gamma J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) \\
& + \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\{\sum_{i=1}^{\bar{k}} x_i \geq 1\}\big) J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \Big].
\end{aligned}
$$
(EC.7)

Similarly, using the suboptimal but feasible action $l = k$ at state $(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1})$, we obtain an upper bound for $J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1})$:

$$
\begin{aligned}
J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \leq \frac{1}{\psi} \Big[& \big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
& + \mu \big[p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{k^*+1}) \\
& \qquad + (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{k^*+1}) \big] \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{k*+1}) \\
& + \gamma J_{n-1}(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \\
& + \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\{\sum_{i=1}^{\bar{k}} x_i \geq 1\}\big) J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \Big].
\end{aligned}
$$
(EC.8)

Now multiplying both sides of (EC.7) by p_{k^*}, both sides of (EC.8) by $(1 - p_{k^*})$, and summing up the results we have:

$$
p_{k^*} J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*}) J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \leq
$$

$$
\frac{1}{\psi} \Big[\big(\sum_{i=1}^{\bar{k}} x_i - 1 \big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1 \big) + \mu \big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1})
$$

$$
+ p_{k^*}\,(1 - p_k)\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1})
$$

$$
+ (1 - p_{k^*})\, p_k\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{\bar{k}+1})
$$

$$
+ (1 - p_{k^*})\,(1 - p_k)\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \big]
$$

$$
+ \gamma \sum_{i=1}^{\bar{k}+1} y_i \big[p_{k^*}\, J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{\bar{k}+1}) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{k^*+1}) \big]
$$

$$
+ \gamma \big[p_{k^*}\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_{k^*})\, J_{n-1}(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \big]
$$

$$
+ \bar{\psi} \big(p_{k^*}\, J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*})\, J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \big) \Big], \tag{EC.9}
$$

where, for the ease of notation, we let $\bar{\psi}$ denote the self-loop rate, i.e., $\bar{\psi} = \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\big\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \big\} \big)$. Now in the above upper bound, using the induction hypothesis, we can replace the terms with coefficient γ to obtain another upper bound. Using Property i and ii for the first and second terms with coefficient γ, we have:

$$
p_{k^*}\, J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*})\, J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \leq
$$

$$
\frac{1}{\psi} \Big[\big(\sum_{i=1}^{\bar{k}} x_i - 1 \big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1 \big) + \mu \big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1})
$$

$$
+ p_{k^*}\,(1 - p_k)\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1})
$$

$$
+ (1 - p_{k^*})\, p_k\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{\bar{k}+1})
$$

$$
+ (1 - p_{k^*})\,(1 - p_k)\, J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \big]
$$

$$
+ \gamma \sum_{i=1}^{\bar{k}+1} y_i \big[p_k\, J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{\bar{k}+1}) + (1 - p_k)\, J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{k+1}) \big]
$$

$$
+ \gamma \big[p_k\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k)\, J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \big]
$$

$$
+ \bar{\psi} \big(p_k\, J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k)\, J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \big) \Big]. \tag{EC.10}
$$

Thus, we have obtained an upper bound for the LHS of (EC.5). Now consider the RHS of (EC.5) and first for state $(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1})$ use (EC.4) to obtain $J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1})$. Note that, by the induction hypothesis, PO is optimal in period n. Hence, it is optimal to assign the physician to treatment stage k^* in period n at state $(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1})$, since k^* is the most down-stream treatment stage with an available patient when state is $(\mathbf{X}, \mathbf{Y})$ (and hence when state is $(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1})$). Thus, using (EC.4) we have:

$$
\begin{aligned}
J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) = \frac{1}{\psi} \Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1 \big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1 \big) \\
&+ \mu \big[p_{k^*} J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1}) \\
&\quad\quad + (1 - p_{k^*}) J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{k^*+1} + e_{\bar{k}+1}) \big] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{\bar{k}+1}) \\
&+ \gamma J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \} \big) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) \Big].
\end{aligned}
$$
(EC.11)

Similarly, using (EC.4) to obtain $J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1})$ we have:

$$
\begin{aligned}
J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) = \frac{1}{\psi} \Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1 \big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1 \big) \\
&+ \mu \big[p_{k^*} J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \\
&\quad\quad + (1 - p_{k^*}) J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \big] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{k+1}) \\
&+ \gamma J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \} \big) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \Big].
\end{aligned}
$$
(EC.12)

Now multiplying both sides of (EC.11) by p_k, both sides of (EC.12) by $(1 - p_k)$, and summing up the results we have:

$$
\begin{aligned}
p_k \, &J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) \, J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) = \\
\frac{1}{\psi} \Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1 \big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1 \big) + \mu \big[p_{k^*} p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1}) \\
&\quad\quad + p_{k^*}(1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \\
&\quad\quad + (1 - p_{k^*}) p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{\bar{k}+1}) \\
&\quad\quad + (1 - p_{k^*})(1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \big] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i \big[p_k J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{\bar{k}+1}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{k+1}) \big] \\
&+ \gamma \big[p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \big] \\
&+ \bar{\psi} \big(p_k J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \big) \Big],
\end{aligned}
$$
(EC.13)

where, for the ease of notation, we again let $\bar{\psi} = \left(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\left\{\sum_{i=1}^{\bar{k}} x_i \geq 1\right\}\right)$. Notice that RHS of (EC.13) is equal to the upper bound of the LHS of (EC.5) derived in (EC.10). Thus, Property i holds for every n by induction, and hence the PO is optimal in every period.

To complete the proof, it remains to show Property ii. To do so, we use the same technique used to show Property i. First, notice that for $n = 0$ (or $n = 1$) this property is trivial. Next suppose it holds for $n - 1$. To show that it would also hold for n, we use suboptimal actions to obtain an upper bound for the LHS of (EC.6) and show that this upper bound is equal to its RHS. To do so, consider states $(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y})$ and $(\mathbf{X} + e_{k^*+1} - e_k, \mathbf{Y})$, and for each one, to obtain an upper bound, use the optimality equation (EC.4) but with suboptimal actions $l = k$. Then multiply the upper bound obtained for $J(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y})$ and $J(\mathbf{X} + e_{k^*+1} - e_k^*, \mathbf{Y})$ by p_{k^*} and $1 - p_{k^*}$, respectively. Summing up the results, we gain the following upper bound for the LHS of (EC.5):

$$
\begin{aligned}
p_{k^*} J_n(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_{k^*}) J_n(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \leq \\
\frac{1}{\psi} \Bigg[-p_{k^*} + \Big(\sum_{i=1}^{\bar{k}} x_i\Big) + \Big(\sum_{i=1}^{\bar{k}+1} y_i\Big) \\
+ \mu \Big[p_{k^*} p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \\
+ p_{k^*} (1 - p_k) J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1}) \\
+ (1 - p_{k^*}) p_k J_{n-1}(\mathbf{X} + e_{k^*+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \\
+ (1 - p_{k^*})(1 - p_k) J_{n-1}(\mathbf{X} + e_{k^*+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1}) \Big] \\
+ \gamma \sum_{i=1}^{\bar{k}+1} y_i \Big[p_{k^*} J_{n-1}(\mathbf{X} + e_{\bar{k}+1} + e_i - e_{k^*}, \mathbf{Y} - e_i) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} + e_{k^*+1} + e_i - e_{k^*}, \mathbf{Y} - e_i) \Big] \\
+ \psi \Big(p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \Big) \Bigg].
\end{aligned}
\tag{EC.14}
$$

Now, using the optimality equation (EC.4) to derive $J_n(\mathbf{X} + e_{\bar{k}+1} - \mathbf{e}_k, \mathbf{Y})$ and $J_n(\mathbf{X} + e_{k+1} - \mathbf{e}_k, \mathbf{Y})$, and then multiplying them by p_k and $1 - p_k$, respectively, and finally summing up the results we get the following equality for the RHS of (EC.5). (Notice that by the induction hypothesis assigning the physician to k^* is optimal when computing $J_n(\mathbf{X} + e_{\bar{k}+1} - \mathbf{e}_k, \mathbf{Y})$ and $J_n(\mathbf{X} + e_{k+1} - \mathbf{e}_k, \mathbf{Y})$.)

$$
\begin{aligned}
p_k J_n(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_n(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) = \\
\frac{1}{\psi} \Bigg[-p_k + \Big(\sum_{i=1}^{\bar{k}} x_i\Big) + \Big(\sum_{i=1}^{\bar{k}+1} y_i\Big)
\end{aligned}
$$

$$+ \mu \Big[p_{k^*} \, p_k \, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1})$$

$$+ p_{k^*} \, (1 - p_k) \, J_{n-1}(\mathbf{X} + e_{k+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1})$$

$$+ (1 - p_{k^*}) \, p_k \, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1})$$

$$+ (1 - p_{k^*}) \, (1 - p_k) \, J_{n-1}(\mathbf{X} + e_{k+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \Big]$$

$$+ \gamma \sum_{i=1}^{\bar{k}+1} y_i \Big[p_k \, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} + e_i - e_k, \mathbf{Y} - e_i) + (1 - p_k) \, J_{n-1}(\mathbf{X} + e_{k+1} + e_i - e_k, \mathbf{Y} - e_i) \Big]$$

$$+ \psi \left(p_k \, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) \, J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \right) \Big]. \qquad \text{(EC.15)}$$

Now, notice that since $k^* > k$, by assumption we have $p_{k^*} \geq p_k$. Next, using the induction hypothesis and since $p_{k^*} \geq p_k$, it is easy to show that the upper bound obtained in (EC.14) is less than or equal to (EC.15), which establishes Property ii for n and completes the proof. $\qquad \square$

Proof of Proposition 3 - Part (ii). We use a sample path argument to show the result in the almost sure sense. Consider the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and similar to the proof of part (i), without loss of generality, suppose at time 0, all of the N patients in the clearing model are in state W_1, i.e., in the initial waiting state depicted in Figure 5. Let $w_1^n(\omega)$ be the realized duration of the initial waiting stage, W_1, for patient $n \in \{1, \cdots, N\}$ under sample path $\omega \in \Omega$. Let G be the set of all admissible (Markovovian or non-Markovian) policies and $TTFT^{g,n}(\omega)$ be the Time To First Treatment of patient n under policy $g \in G$ and sample path $\omega \in \Omega$. Notice that $TTFT^{g,n}(\omega) \geq w_1^n(\omega)$ for every $g \in G$, every $\omega \in \Omega$, and every $n \in \{1, \cdots, N\}$, since a patient cannot been seen before s/he finishes stage W_1. Therefore, $\inf_{g \in G} TTFT^{g,n}(\omega) \geq w_1^n(\omega)$. Now notice that for the underlying Prioritize New (PN) policy, which instructs the physician to initialize a new patient journey whenever possible (perhaps by preempting other tasks), $TTFT^{PN,n}(\omega) = w_1^n(\omega)$ (for every $\omega \in \Omega$, and every $n \in \{1, \cdots, N\}$). Thus, the PN obtains the minimum TTFT of every patient along every sample path. Therefore, PN also minimizes the average TTFT of patients with probability one (i.e., in the almost sure sense). $\qquad \square$

Online Appendix B: Computations Under Imperfect Classification

Assume $I \in \{A, D\}$ represents the true identity of the patient (Admit or Discharge) and $\omega \in \{A, D\}$ is the signaled/identified class. Let $\gamma_A = Pr(\omega = D | I = A)$ and $\gamma_D = Pr(\omega = A | I = D)$. Next, if $\tilde{\gamma}_A = Pr(I = A | \omega = D)$ and $\tilde{\gamma}_D = Pr(I = D | \omega = A)$ represent the misclassification probabilities, with $\alpha = Pr(I = A)$, using Bayes rule we have:

$$\tilde{\gamma}_A = Pr(I = A | \omega = D) = \frac{\alpha\,\gamma_A}{\alpha\gamma_A + (1 - \alpha)(1 - \gamma_D)},$$
$$\tilde{\gamma}_D = Pr(I = D | \omega = A) = \frac{(1 - \alpha)\,\gamma_D}{\alpha(1 - \gamma_A) + (1 - \alpha)\gamma_D}.$$

To isolate the effect of misclassification errors, we eliminate variability in the treatment times, X_A and X_D so that $Pr(X_A = \mu_A) = 1$ and $Pr(X_D = \mu_D) = 1$. Moreover, for the ease of computations, we consider a collaborative service environment whenever the system is working in the pooling mode (i.e., under pooling or under streaming after one stream runs out of patients). Collaborative assumption means that the two servers work together on one patient at a time with service times of $\mu_A/2$ for admits and $\mu_D/2$ for discharges.

Let n be the total number of patients in the clearing system. Suppose N_A and $N_D = n - N_A$ denote the random variable representing the number of patients that are identified as A and D, respectively. Let $\tilde{N}_A$ and $\tilde{N}_D$ be the random variables representing last patients of type A and D that are seen before the system moves to a pooling scenario, reactively. Next notice that given N_A (and hence $N_D = n - N_A$), $\tilde{N}_A$ and $\tilde{N}_D$, expected TTFT of Admits under Streaming can be computed by:

$$E\big[TTFT_A^S | N_A = n_A, \tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D\big] = \frac{1}{(1 - \tilde{\gamma}_D)n_A + \tilde{\gamma}_A(n - n_A)} \times$$

$$\Big[(1 - \tilde{\gamma}_D)\Big[\sum_{j=1}^{\tilde{n}_A}\sum_{k=0}^{j-1} \binom{j-1}{k}\tilde{\gamma}_D^k\big(1 - \tilde{\gamma}_D\big)^{j-k-1}\big(k\mu_D + (j - k - 1)\mu_A\big)$$
$$+ \sum_{j=\tilde{n}_A+1}^{n_A}\Big[\sum_{k=0}^{\tilde{n}_A}\binom{\tilde{n}_A}{k}\tilde{\gamma}_D^k\big(1 - \tilde{\gamma}_D\big)^{\tilde{n}_A - k}\big(k\mu_D + (\tilde{n}_A - k)\mu_A\big)$$
$$+ \sum_{k=0}^{j-\tilde{n}_A-1}\binom{j - \tilde{n}_A - 1}{k}\tilde{\gamma}_D^k\big(1 - \tilde{\gamma}_D\big)^{j-\tilde{n}_A-1-k}\big(k\frac{\mu_D}{2} + (j - \tilde{n}_A - 1 - k)\frac{\mu_A}{2}\big)\Big]\Big]$$

$$+\tilde{\gamma}_A\Big[\sum_{j=1}^{\tilde{n}_D}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\mu_A+(j-k-1)\mu_D\big)$$

$$+\sum_{j=\tilde{n}_D+1}^{n-n_A}\Big[\sum_{k=0}^{\tilde{n}_D}\binom{\tilde{n}_D}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{\tilde{n}_D-k}\big(k\mu_A+(\tilde{n}_D-k)\mu_D\big)$$

$$+\sum_{k=0}^{j-\tilde{n}_D-1}\binom{j-\tilde{n}_D-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-\tilde{n}_D-1-k}\big(k\frac{\mu_A}{2}+(j-\tilde{n}_D-1-k)\frac{\mu_D}{2}\big)\Big]\Big]\Big].$$

$$(EC.16)$$

The first line in the above equation is the reciprocal of the number of A patients (either classified as A or D). The second line considers the jth patient in the stream of the patients classified as A and seen before the system moves to a pooling scenario (i.e., up to $\tilde{n}_A$) and computes its TTFT by conditioning on the number of D patients in front him. Similarly, the third and fourth line consider the jth patient in the stream of the patients classified/signaled as A and seen after the system moves to a pooling scenario (i.e., after $\tilde{n}_A$). The second, third, and fourth lines are multiplied by $(1-\tilde{\gamma}_D)$ (i.e., the probability that a patient classified as A is truly A type) to give the total sum of TTFT of A patients who are also classified as A. Similarly, the fifth, sixth, and seventh lines compute the sum of TTFT of A patients who are classified as D.

Now if $g(n_A,\tilde{n}_A,\tilde{n}_D)$ represents the joint pdf of random variables $N_A,\tilde{N}_A,\tilde{N}_D$ then we have:

$$\overline{TTFT}_A^S=E\big[E[TTFT_A^S|N_A,\tilde{N}_A,\tilde{N}_D]\big]=\sum_{n_A=0}^{n}\sum_{\tilde{n}_A=0}^{n}\sum_{\tilde{n}_D=0}^{n}E[TTFT_A^S|N_A,\tilde{N}_A,\tilde{N}_D]g(n_A,\tilde{n}_A,\tilde{n}_D),$$

$$(EC.17)$$

where $E[TTFT_A^S|N_A,\tilde{N}_A,\tilde{N}_D]$ is computed in (EC.16). To compute $\overline{TTFT}_A^S$ using the above equation, it remains to derive $g(n_A,\tilde{n}_A,\tilde{n}_D)$. To derive $g(n_A,\tilde{n}_A,\tilde{n}_D)$ notice that:

$$g(n_A,\tilde{n}_A,\tilde{n}_D)=$$

$$Pr(N_A=n_A,\tilde{N}_A=\tilde{n}_A,\tilde{N}_D=\tilde{n}_D)=$$

$$Pr(\tilde{N}_A=\tilde{n}_A,\tilde{N}_D=\tilde{n}_D|N_A=n_A)\times Pr(N_A=n_A)=$$

$$Pr(N_A=n_A)\big[Pr(\tilde{N}_A=\tilde{n}_A=N_A=n_A,\tilde{N}_D=\tilde{n}_D)\mathbb{1}\{\tilde{n}_D<n-n_A=n_D\} \tag{EC.18}$$

$$+Pr(\tilde{N}_A=\tilde{n}_A,\tilde{N}_D=\tilde{n}_D=n-N_A=n-n_A)\mathbb{1}\{\tilde{n}_A<n_A\} \tag{EC.19}$$

$$+Pr(\tilde{N}_A=\tilde{n}_A=N_A=n_A,\tilde{N}_D=\tilde{n}_D=n-N_A=n-n_A=n_D)\mathbb{1}\{\tilde{n}_A=n_A,\tilde{n}_D=n-n_A\}\big]$$

$$(EC.20)$$

In above, Eq's (EC.18), (EC.19), and (EC.20) correspond to the cases where the D stream is finished first, the A stream is finished first, and the case where one stream is done when the system is working on the last patient of the other stream, respectively. Next notice that with $p = (1 - \gamma_A)\alpha + \gamma_D(1 - \alpha)$ denoting the probability that a patient is identified as A:

$$Pr(N_A = n_A) = \binom{n}{n_A} p^{n_A}(1 - p)^{n - n_A}. \tag{EC.21}$$

Let K_j^A and K_j^D be the random variables denoting the number of D type patients up to (and including) the jth patient in A and D streams, respectively. Then to compute (EC.18), we need to compute the probability that the time required to see n_A patients in the A stream is between the time required to see $\tilde{n}_D$ and $\tilde{n}_D + 1$ patients in the D stream (so that $\tilde{n}_D$ is the last patient seen in the D stream before the system moves to the pooling scenario). we have:

$$\begin{aligned}
&Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D) \\
&= Pr\big((\tilde{n}_D - K_{\tilde{n}_D}^D)\mu_A + K_{\tilde{n}_D}^D \mu_D \leq (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A \mu_D\big) \\
&\quad - Pr\big((n_A - K_{n_A}^A)\mu_A + K_{n_A}^A \mu_D \geq (\tilde{n}_D + 1 - K_{\tilde{n}_D+1}^D)\mu_A + K_{\tilde{n}_D+1}^D \mu_D\big) \\
&= Pr\big(K_{n_A}^A - K_{\tilde{n}_D}^D \leq \mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}\big) \\
&\quad - Pr\big(K_{n_A}^A - K_{\tilde{n}_D+1}^D \leq \mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}\big) \\
&= F_1\big(\mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}\big) - F_2\big(\mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}\big)
\end{aligned} \tag{EC.22}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the CDF of the random variables $Z_1 = K_{n_A}^A - K_{\tilde{n}_D}^D$ and $Z_2 = K_{n_A}^A - K_{\tilde{n}_D+1}^D$, respectively. Similarly, to compute (EC.19), we have:

$$\begin{aligned}
&Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\
&= Pr\big((\tilde{n}_A - K_{\tilde{n}_A}^A)\mu_A + K_{\tilde{n}_A}^A \mu_D \leq (n_D - K_{n_D}^D)\mu_A + K_{n_D}^D \mu_D\big) \\
&\quad - Pr\big((n_D - K_{n_D}^D)\mu_A + K_{n_D}^D \mu_D \geq (\tilde{n}_A + 1 - K_{\tilde{n}_A+1}^A)\mu_A + K_{\tilde{n}_A+1}^A \mu_D\big) \\
&= Pr\big(K_{n_D}^D - K_{\tilde{n}_A}^A \leq \mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}\big) \\
&\quad - Pr\big(K_{n_D}^D - K_{\tilde{n}_A+1}^D \leq \mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}\big) \\
&= F_3\big(\mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}\big) - F_4\big(\mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}\big)
\end{aligned} \tag{EC.23}$$

where $F_3(\cdot)$ and $F_4(\cdot)$ are the CDF of the random variables $Z_3 = K^D_{n_D} - K^A_{\tilde{n}_A}$ and $Z_4 = K^D_{n_D} - K^D_{\tilde{n}_A+1}$, respectively.

Next, to compute (EC.20), we need to compute the probability that one stream finishes when the system is working on the last patient of the other stream:

$$Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D)$$

$$= Pr(T^A_{n_A-1} < T^D_{n_D} \leq T^A_{n_A}) + Pr(T^D_{n_D-1} < T^A_{n_A} < T^D_{n_D}) \tag{EC.24}$$

$$= Pr(((n_A-1) - K^A_{n_A-1})\mu_A + K^A_{n_A-1}\mu_D < (n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D \leq (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D)$$

$$+ Pr(((n_D-1) - K^D_{n_D-1})\mu_A + K^D_{n_D-1}\mu_D < (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D < (n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D)$$

$$= Pr\left(K^D_{n_D} - K^A_{n_A-1} < \mu_A \frac{n_D - (n_A-1)}{\mu_A - \mu_D}\right) + Pr\left(K^A_{n_A} - K^D_{n_D-1} < \mu_A \frac{n_A - (n_D-1)}{\mu_A - \mu_D}\right)$$

$$- Pr((n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D > (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D)$$

$$- Pr((n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D \leq (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D)$$

$$= Pr\left(K^D_{n_D} - K^A_{n_A-1} < \mu_A \frac{n_D - (n_A-1)}{\mu_A - \mu_D}\right) + Pr\left(K^A_{n_A} - K^D_{n_D-1} < \mu_A \frac{n_A - (n_D-1)}{\mu_A - \mu_D}\right) - 1$$

$$= F_5\left(\mu_A \frac{n_D - (n_A-1)}{\mu_A - \mu_D}\right) + F_6\left(\mu_A \frac{n_A - (n_D-1)}{\mu_A - \mu_D}\right) - 1$$

$$- Pr\left(K^D_{n_D} - K^A_{n_A-1} = \mu_A \frac{n_D - (n_A-1)}{\mu_A - \mu_D}\right) - Pr\left(K^A_{n_A} - K^D_{n_D-1} = \mu_A \frac{n_A - (n_D-1)}{\mu_A - \mu_D}\right)$$

where T in (EC.24) is used to show the finish times of corresponding jobs, and $F_5(\cdot)$ and $F_5(\cdot)$ are CDFs of random variables $Z_5 = K^D_{n_D} - K^A_{n_A-1}$ and $Z_6 = K^A_{n_A} - K^D_{n_D-1}$. Now notice that random variables $Z_1,...,Z_6$ are each the difference between two independent binomial random variables with known parameters. Thus, CDFs $F_1,...,F_6$ are known. Therefore, $g(n_A, \tilde{n}_A, \tilde{n}_D)$ can be computed. As a result, the metric $\overline{TTFT}^S_A$ is completely computed.

Next, in a similar way, we compute the metric $\overline{LOS}^S_D$ (i.e., Expected Length of Stay of D patients under Streaming):

$$\overline{LOS}^S_D = E\left[E[LOS^S_D | N_A, \tilde{N}_A, \tilde{N}_D]\right] = \sum_{n_A=0}^{n} \sum_{\tilde{n}_A=0}^{N} \sum_{\tilde{n}_D=0}^{N} E[LOS^S_D | N_A, \tilde{N}_A, \tilde{N}_D] g(n_A, \tilde{n}_A, \tilde{n}_D) \tag{EC.25}$$

where:

$$[LOS^S_D | N_A = n_A, \tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D] = \frac{1}{\tilde{\gamma}_D n_A + (1 - \tilde{\gamma}_A)(n - n_A)} \times$$
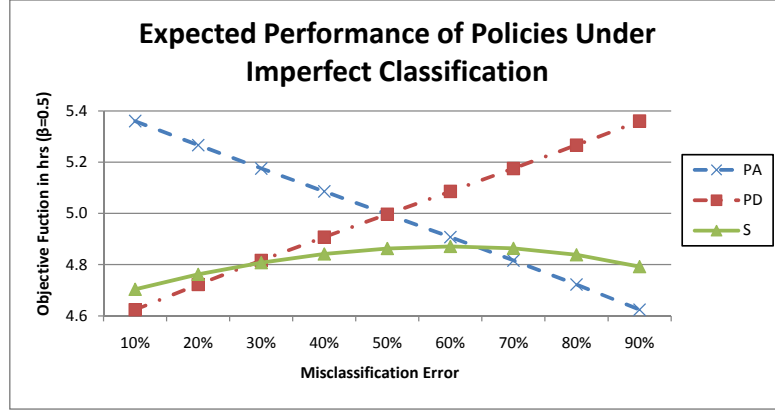
Figure EC.1 Expected performance of policies for a clearing system with $n = 20$, $\mu_A = 80(mins)$, $\mu_D = 45(mins)$, and symmetric misclassification error between A and D patients. Streaming is more robust to misclassification errors than pooling.

$$
\Big[\tilde{\gamma}_D\Big[\sum_{j=1}^{\tilde{n}_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big((k+1)\mu_D+(j-k-1)\mu_A\big)
$$
$$
+\sum_{j=\tilde{n}_A+1}^{n_A}\Big[\sum_{k=0}^{\tilde{n}_A}\binom{\tilde{n}_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{\tilde{n}_A-k}\big((k+1)\mu_D+(\tilde{n}_A-k)\mu_A\big)
$$
$$
+\sum_{k=0}^{j-\tilde{n}_A-1}\binom{j-\tilde{n}_A-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-\tilde{n}_A-k-1}\big((k+1)\frac{\mu_D}{2}+(j-\tilde{n}_A-k-1)\frac{\mu_A}{2}\big)\Big]\Big]
$$
$$
+(1-\tilde{\gamma}_A)\Big[\sum_{j=1}^{\tilde{n}_D}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\mu_A+(j-k)\mu_D\big)
$$
$$
+\sum_{j=\tilde{n}_D+1}^{n_D}\Big[\sum_{k=0}^{\tilde{n}_D}\binom{\tilde{n}_D}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{\tilde{n}_D-k}\big(k\mu_A+(\tilde{n}_D-k)\mu_D\big)
$$
$$
+\sum_{k=0}^{j-\tilde{n}_D-1}\binom{j-\tilde{n}_D-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-\tilde{n}_D-k-1}\big(k\frac{\mu_A}{2}+(j-\tilde{n}_D-k)\frac{\mu_D}{2}\big)\Big]\Big]\Big].
$$
$$(EC.26)$$

Next we need to compute same metrics but under $\pi = PA$ and $\pi = PD$:

$$
E\big[TTFT_A^{PA}|N_A=n_A\big] =
$$

$$
\frac{1}{(1-\tilde{\gamma}_D)n_A+\tilde{\gamma}_A(n-n_A)}\times\Big[(1-\tilde{\gamma}_D)\sum_{j=1}^{n_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big(k\frac{\mu_D}{2}+(j-k-1)\frac{\mu_A}{2}\big)
$$
$$
+\tilde{\gamma}_A\sum_{j=1}^{n-n_A}\Big[\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\frac{\mu_A}{2}+(j-k-1)\frac{\mu_D}{2}\big)
$$
$$
+\sum_{k=0}^{n_A}\binom{n_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{n_A-k}\big(k\frac{\mu_D}{2}+(n_A-k)\frac{\mu_A}{2}\big)\Big]\Big].
$$

Moreover, we have:

$$
\overline{TTFT}_A^{PA} = \sum_{n_A=0}^{n} E\big[TTFT_A^{PA}|N_A=n_A\big]\times Pr(N_A=n_A),
$$

where $Pr(N_A = n_A)$ is given in (EC.21).

Similarly we can compute $\overline{LOS}_D^{PA}$:

$$E\left[\overline{LOS}_D^{PA} | N_A = n_A\right] = \frac{1}{\tilde{\gamma}_D n_A + (1 - \tilde{\gamma}_A)(n - n_A)} \times$$

$$\left[\tilde{\gamma}_D \sum_{j=1}^{n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-k-1} \left((k+1)\frac{\mu_D}{2} + (j-k-1)\frac{\mu_A}{2}\right)\right.$$

$$+ (1 - \tilde{\gamma}_A) \sum_{j=1}^{n-n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-k-1} \left(k\frac{\mu_A}{2} + (j-k)\frac{\mu_D}{2}\right)\right.$$

$$\left.\left.+ \sum_{k=0}^{n_A} \binom{n_A}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{n_A-k} \left(k\frac{\mu_D}{2} + (n_A - k)\frac{\mu_A}{2}\right)\right]\right],$$

$$\text{(EC.27)}$$

and:

$$\overline{LOS}_D^{PA} = \sum_{n_A=0}^{n} E\left[LOS_D^{PA} | N_A = n_A\right] \times Pr(N_A = n_A).$$

It remains to compute the metrics under $\pi = PD$:

$$E\left[TTFT_A^{PD} | N_A = n_A\right] = \frac{1}{(1 - \tilde{\gamma}_D) n_A + \tilde{\gamma}_A (n - n_A)} \times$$

$$\left[\tilde{\gamma}_A \sum_{j=1}^{n-n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{j-k-1} \left(k\frac{\mu_D}{2} + (j-k-1)\frac{\mu_A}{2}\right)\right.$$

$$+ (1 - \tilde{\gamma}_D) \sum_{j=1}^{n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \gamma_D)^k \gamma_D^{j-k-1} \left(k\frac{\mu_A}{2} + (j-k-1)\frac{\mu_D}{2}\right)\right.$$

$$\left.\left.+ \sum_{k=0}^{n-n_A} \binom{n-n_A}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{n-n_A-k} \left(k\frac{\mu_D}{2} + (n - n_A - k)\frac{\mu_A}{2}\right)\right]\right],$$

and:

$$\overline{TTFT}_A^{PD} = \sum_{n_A=0}^{n} E\left[TTFT_A^{PD} | N_A = n_A\right] \times Pr(N_A = n_A).$$

Similarly, we have:

$$E\left[LOS_D^{PD} | N_A = n_A\right] = \frac{1}{\tilde{\gamma}_D n_A + (1 - \tilde{\gamma}_A)(n - n_A)} \times$$

$$\left[(1 - \tilde{\gamma}_A) \sum_{j=1}^{n-n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{j-k-1} \left((k+1)\frac{\mu_D}{2} + (j-k-1)\frac{\mu_A}{2}\right)\right.$$

$$+\tilde{\gamma}_D \sum_{j=1}^{n_A} \Big[\sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{(1-\tilde{\gamma}_D)}^k \gamma_D^{j-k-1} \big(k\frac{\mu_A}{2} + (j-k)\frac{\mu_D}{2}\big)$$

$$+ \sum_{k=0}^{n-n_A} \binom{n-n_A}{k} (1-\tilde{\gamma}_A)^k \tilde{\gamma}_A^{n-n_A-k} \big(k\frac{\mu_D}{2} + (n-n_A-k)\frac{\mu_A}{2}\big)\Big]\Big],$$

and:

$$\overline{LOS}_D^{PD} = \sum_{n_A=0}^{n} E\big[LOS_D^{PD}|N_A=n_A\big] \times Pr(N_A=n_A).$$

Therefore, we have computed expected values of all metrics under different possible policies. Using these computation, Figure EC.1 depicts the performances for a typical numerical example. An important observation is that streaming is much more robust to misclassification errors than the pooling policies.

Online Appendix C: Further Descriptions of the Simulation Framework and Assumptions.

In this section we describe the patient flow and assumptions of our simulation framework in more details. Many assumptions are made to be as close as possible to the practice observed in University of Michigan Emergency Department (UMED). A year of data from UMED is gathered to calibrate the simulation. The simulation was developed in a C++ framework. Our model can be described as a cycle-stationary model with a period of one week. Each data point is obtained for 5000 replications of simulating a week, where each replication is preceded by a warm up period of one week (which was observed to be a sufficient warm up period because correlations in the ED flow are small for spans of two or more days). The number of replications (5000) is chosen so that the confidence intervals are tight enough that (1) the sample averages are reliable, and (2) our data presentation need not to visualize these very tight intervals.

Arrival Process. Arrivals for patient classes are modeled using non-stationary Poisson processes. The arrival rates for different classes (obtained from a year of UMED data) are depicted in Figure 6. The general pattern is similar to those found in other studies (e.g., Green et al. (2006)). A "thinning" mechanism (see Lewis and Shedler (1979a) and Lewis and Shedler (1979b)) is used to simulate the non-stationary Poisson process arrivals for each class of patients (with rates depicted in Figure 6).

Service Process. The service process in the ED is depicted in Figure 5. Each patient goes through several phases of patient-physician interactions/treatment followed by tests and preparations. The duration of each interaction is stochastic and depends on the class of the patient and the number of previous interactions. For instance, the first and last interactions are usually longer than intermediate ones. Also, the duration of "wait" states is stochastic and depend on the class of the patient, based on the information at the UMED. For instance, the last "wait" state, i.e., where the patient is given final directions and is waiting to be disposed is much longer for admits since they have to be boarded until a bed becomes available in the hospital (the so-called hospital bed-block effect). The number of interactions with a physician per patient ranges from 1 to 7 and depends on the
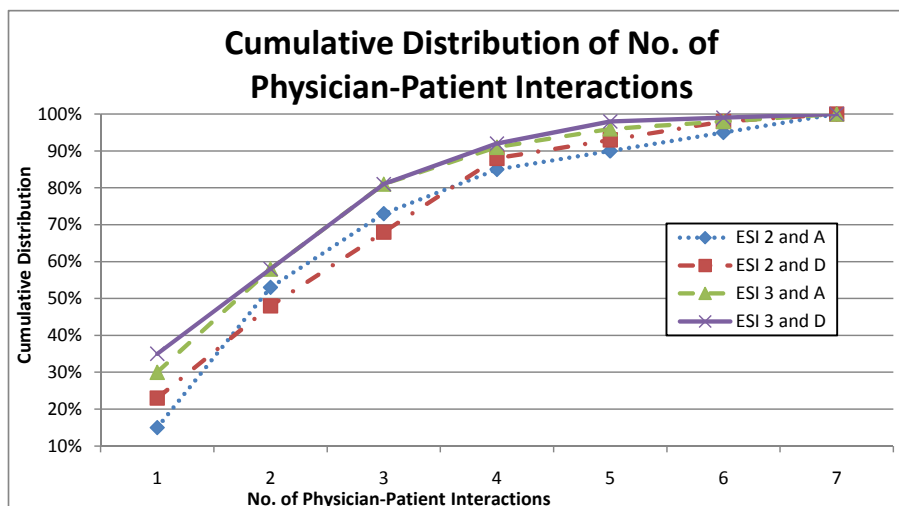
Figure EC.2   Cumulative number of class based physician-patient interactions

class of the patient, as well as several other factors. Based on the class of the patient, we draw the number of such interactions from a distribution constructed from a detailed time study published in Graff et al. (1993) (see Table 3 there) after modifying the data to represent our four patient classes. These class based distributions are depicted in Figure EC.2. The simulated service process is non-collaborative (an ED physician rarely transfers his/her patients to another physician) and non-preemptive (an ED physician rarely moves to another patient in the middle of his/her current interaction).

Phase 1: Assigning Patients to Rooms and Physicians. Whenever a room/bed becomes available, the nurse who is in charge of bed assignment transfers a triaged patient from the waiting area to that room. S/he uses a Phase 1 sequencing rule to decide which patient to bring in to an exam room from the main waiting area (see the body of the paper for different Phase 1 rules implemented). In the VS designs, if an A(D) bed becomes available, the nurse in charge brings an A(D) patient (with priority to patients of ESI 2) from the waiting area in to one of the rooms. If however, an A(D) patient is not waiting in the waiting area, the nurse brings in a D(A) patient (with priority to patients of ESI 2). Also, after an A(D) patient is triaged, s/he is directly guided to one of the A(D) beds if one such bed is available, and if not, to one of the D(A) beds (i.e., bed sharing is allowed, since beds are only virtually separated). If, however, no bed is available, the patient has to wait in the waiting area. Once a room/bed is assigned to a patient, the bed cannot be occupied by another patient until s/he leaves the ED; the bed assigned to a patient cannot be

assigned to any other one, even if the patient is sent to another facility for a test. After the patient is brought into the room, s/he goes through the first "waiting" state (i.e., initial preparation by a nurses) which takes some stochastic amount of time. The average duration of this stage depends on the class of the patient. After this stage the patient is assigned to a physician (if a physician is available) where his/her first treatment starts. The rule to choose a physician is generally to assign the patient to the physician who is handling the lowest number of patients (among those available at that time). However, the rules to choose a physician is different between the virtual streaming (VS) and the pooling patient flow designs, since in a VS design the physicians are divided to two groups one for A patients and one for D patients. Under a VS design, if the patient is assessed to be of A(D) type, the priority is given to physicians devoted to A(D). In other words, an available A(D) type physician is allowed to cross to the other stream only if a physician of D(A) type is needed but is not available (due to being busy with a patient or being currently assigned to the maximum number of patients that a physician is willing to handle). Under pooling designs, physicians do not have labels and therefore a physician who is handling the lowest number of patients (among those available at that time) becomes responsible for the newly arrived patient. Once a physician is assigned to a patient s/he is the only physician who can work on that patient. If no physician is available at the time the patient is ready for his/her first interaction with the physician, the patient has to wait in the exam room.

Phase 2: Which patient to choose next? Whenever a physician finishes a treatment stage (including direct and indirect interactions), s/he is available to visit another patient. The physician chooses the next patient based on the instructions s/he is given according to the Phase 2 sequencing rule. If the physician has less than the upper bound on the number of patients that a physician is willing to handle (7 was used based on the UMED data), s/he can also choose to initialize a new journey by taking a new patient: visiting a patient who has been taken to a room but has been waiting for a physician to become available. Under the VS designs, physicians with A(D) label first use the Phase 2 priority rule on the patients of A(D) type and are allowed to handle D(A) patients only to avoid starvation.

References (for the Online Appendix)

Green, L. V., J. Soares, J. F. Giglio, R.A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1) 61-68.

Lewis, P. A.W., G. S. Shedler. 1979a. Simulation of nonhomogenous poisson processes by thinning. *Naval Research Logistics Quarterly* 26(3) 403-413.

Lewis, P. A.W., G. S. Shedler. 1979b. Simulation of nonhomogenous poisson processes with degree-two exponential polynomial rate function. *Oper. Res.* 27(5) 1026-1039.