

Patient Streaming as a Mechanism for Improving
Responsiveness in Emergency Departments

Wallace J. Hopp
Stephen M. Ross School of Business
at the University of Michigan

Jeffrey S. Desmond
Emergency Department
University of Michigan Hospital

Soroush Saghafian
Dept. of Industrial & Operational Engineering
University of Michigan

Mark P. Van Oyen
Dept. of Industrial & Operational Engineering
University of Michigan

Steven L. Kronick
Emergency Department
University of Michigan Hospital

Ross School of Business Working Paper
Working Paper No. 1162
November 2010

This work cannot be used without the author's permission.
This paper can be downloaded without charge from the
Social Sciences Research Network Electronic Paper Collection:
<http://ssrn.com/abstract=1911856>

Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments

Sorouh Saghafian¹, Wallace J. Hopp², Mark P. Van Oyen¹, Jeffrey S. Desmond³ (M.D.),
Steven L. Kronick³ (M.D.)

¹ Dept. of Industrial & Operations Eng., Univ. of Michigan, Ann Arbor, MI

² Ross School of Business, Univ. of Michigan, Ann Arbor, MI

³Emergency Department, Univ. of Michigan Hospital, Ann Arbor, MI

Crisis level overcrowding conditions in Emergency Departments (ED's) 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 predications 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 virtually (rather than physically) separated. This leads us to propose a new "virtual-streaming" patient flow design for ED's. Second, this type of streaming is most effective in ED's with (1) high percentage of admitted patients, (2) longer care times for admitted patients than discharged patients, (3) low relative weight assigned to TTFT (time to first treatment) compared to LOS (length of stay) of patients, (4) high day-to-day variation in the percentage of admitted patients, (5) long patient boarding time (e.g., caused by hospital "bed-block"), and (6) high average physician utilization. Finally, to take full advantage of streaming, physicians assigned to admit patients should prioritize upstream (new) patients, while physicians assigned to discharge patients should prioritize downstream (old) patients.

Key words: Health Care Operations Management; Emergency Department; Patient Flow; Patient Sequencing.

History: Last Revision: November 11, 2010

1. Introduction

Between 1996 and 2006, annual visits to Emergency Departments (ED's) in the U.S. increased by 32% (from 90.3 million to 119.2 million), while the number of hospital ED's decreased from 4,019 to 3,833 (NHAMCS by Pitts et al. (2008)). This trend has elevated ED overcrowding to crisis levels in many hospitals in the U.S. and 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, IL, complaining of nausea, shortness of breath, and chest pain. Triageed 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 “elope-ment” (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 U.S, 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. ED’s (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) non-urgent 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. But, since this is also the most expensive approach, it is generally not an attractive option. Recognizing this, Richardson (2003) 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 particular, Flinders Medical Center in Australia has implemented a system in which ED patients and resources are divided into two streams: one for those likely to be discharged (hereafter “D patients”) and one for those likely to be admitted to the hospital (hereafter “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. Of course, since Flinders represented a single uncontrolled experiment in a specific environment in which other changes (e.g., lean initiatives) were implemented along with their streaming system, it is impossible to infer that these results would translate to other ED’s. Nevertheless, motivated by positive reports from Flinders, other hospitals such as Bendigo Health ED (Kinsman et al. (2008)) have begun implementing similar strategies.

While streaming patients based on the likelihood of being admitted to the hospital is new,

patient streaming is not. By the 1980s most ED's (although not Flinders) had adopted separate "fast tracks" for patients with minor injuries (Welch (2008)). In the 1990s, many ED's initiated separate "observation units" for patients requiring lengthy diagnosis. But, as Welch (2008) 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 convenience, we will refer to the new policy that separates patients (and resources) into different streams according to anticipated disposition (A or D) as "streaming". 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. But since 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. While this suggests that sequencing alone is not

enough to overcome the anti-pooling 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 Section 5, we consider another analytic clearing model, with perfect patient classification but with multi-stage service processes, in order to understand the impact of patient sequencing within the exam rooms (i.e., the order in which physicians visit patients assigned to them) on the streaming versus pooling comparison. We find that prioritizing downstream D patients and upstream A type patients enhances the advantage of streaming over pooling. In Section 6, we use a simulation model of a realistic scenario with dynamic patient arrivals and multi-stage service processes, in addition to patient misclassification error, to test the conjectures made from our analytic models. 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 Section 7 with a summary of our overall insights about whether, where and how streaming can be a potentially attractive strategy for improving ED responsiveness.

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”*. 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 stream 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 have used queueing models to study various aspects of the ED (see, for instance, Kochran and Roche (2009), Green et al. (2006), and Allon et al. (2010)). However, 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 concerning pooling in queueing systems is that when two classes of customers in a queueing system become sufficiently different, pooling becomes ineffective and may be even 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 anti-pooling disadvantage. But verifying this requires some new analysis 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 (see e.g., Rothkopf and Rech (1987), Whitt (1999), Hu and Benjafar (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 Benjafar (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 ED's) suggest that these insights may not hold when classification is imperfect (see, e.g., Argon and Ziya (2009)).

Another line of research is relevant to sequencing within the examination rooms. 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), while related studies of parallel queueing systems can be found, for instance, in Andradóttir et al. (2003) and the references therein. In particular, Van Oyen et al. (2001) proposed a “pick-and-run” policy for servers in a serial system which 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 so as to reduce 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 non-stationary stochastic manner. Upon arrival, patients first go to the triage stage 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

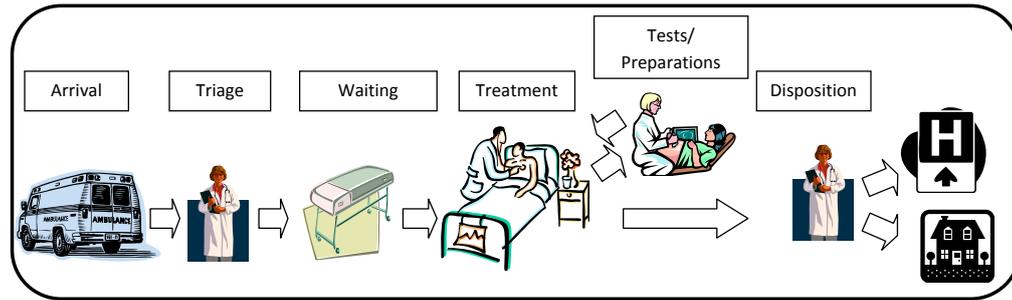


Figure 1 The general flow of patients in an ED.

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 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 a physician, as shown in Figure 1. (While caregivers may not be physicians (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 preparation/processing activities by nurse (e.g., wound cleaning) 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’s task 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 for each treatment phase. Also, in most ED’s, 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). Since 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. This is usually done by a nurse in consideration of ESI levels and patient arrival orders. In theory, it could also make use of A/D predictions. *Phase 2 sequencing* determines the order in which patients are seen within the examination rooms. 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 physicians. Furthermore, physicians 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 above-mentioned 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 (since a low LOS for D patients also guarantees a low TTFT for them). But for A patients, LOS in the ED is usually a very 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. Since safety is enhanced

by starting treatment as soon as possible (recall the case mentioned in the Introduction, in which a patient died from lack of attention while waiting), TTFT is the most appropriate 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 $\pi \in \mathbf{\Pi}$, respectively, where $\mathbf{\Pi} = \{PA(\text{Pooling with priority to A's}), PD(\text{Pooling with priority to D's}), S(\text{Streaming})\}$ represents the set of admissible policies. To construct a single objective function, we let β represent the relative weight placed on 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 $\pi \in \mathbf{\Pi}$. 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 TTTF 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 TTTF of A patients and LOS of D patients by one minute is equally costly. We also note that while other metrics are used to evaluate the performance of an ED, most of them are highly correlated with our objective function. For example, the percentage of patients who leave without being seen (LWBS) is commonly tracked in ED's, but studies such as Fernandes et al. (1994) conclude 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.

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 tradeoff.

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 is 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 simple *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, this clearing model will help us gain insights into the mechanics of the streaming versus pooling comparison. We summarize our main findings in the propositions and observations below.

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). *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 , $\operatorname{argmin}_{\pi \in \Pi} T_A^\pi(\alpha, \omega) = PA$. That is, if only TTFT of A's matters, then pooling with priority to A's is the optimal policy (in the almost sure sense).*
- (ii) *For every $\alpha \in [0, 1]$ and every sample path ω , $\operatorname{argmin}_{\pi \in \Pi} L_D^\pi(\alpha, \omega) = PD$. That is, if only LOS of D's matters, then pooling with priority to D's is the optimal policy (in the almost sure sense).*

This proposition suggests that streaming is not attractive unless we care about both TTFT and for A's and LOS for D's. Thus, 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 $\pi : [0, 1] \times [0, 1] \rightarrow \mathbf{\Pi}$ 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 $\pi \in \mathbf{\Pi}$, the π region, denoted by \mathcal{A}^π , is the collection of (α, β) for which policy π is optimal. That is, the π region is $\mathcal{A}^\pi = \{(\alpha, \beta) \in [0, 1] \times [0, 1] : \pi^*(\alpha, \beta) = \pi\}$.

DEFINITION 3 (β -CONVEXITY). The optimal strategy $\pi^* : [0, 1] \times [0, 1] \rightarrow \mathbf{\Pi}$ is said to be β -convex if all the π regions (i.e, sets \mathcal{A}^π ($\forall \pi \in \mathbf{\Pi}$)) 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 (**Threshold Policy**). *For every fixed $\alpha \in [0, 1]$, there exist double thresholds $\underline{\beta}(\alpha), \bar{\beta}(\alpha)$ such that streaming is optimal if, and only if, $\beta \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. If $\beta < \underline{\beta}(\alpha)$ then pooling with priority to D's is optimal. If $\beta > \bar{\beta}(\alpha)$ then pooling with priority to A's is optimal.*

Since ED's vary in their percentage of A's (α) and relative weight of TTFT of D's (β), the breadth of 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. Indeed, Figure 2 illustrates an example with deterministic service times in which there is no region of optimality for streaming. There are two reasons for this: (1) we allow prioritization of D patients, and (2) we assume we know the exact proportion of A's (α). Regarding the first issue, it may be optimal in theory to prioritize D's for given values of α and β , but many ED's do not consider this as a practical option, because A's are more likely to need to be stabilized quickly. This was the case at Flinders Medical Center (King et al. (2006) and Ben-Tovim et al. (2008)). If we rule out the $\pi = PD$ option (and assume deterministic service times to enable us to compute the exact performance metrics as a function of service times and patient mix), then we can state the following conditions under which streaming is preferred to pooling with priority given to A's.

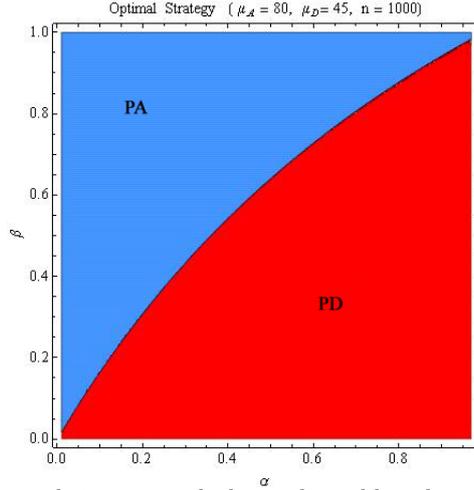


Figure 2 An example of the optimal strategy with three admissible policies (PA, PD, S) and deterministic service times for which streaming is almost never optimal.

PROPOSITION 3 (Attractiveness of Streaming). *If $\Pi = \{PA, S\}$ and $Pr(X_A = \mu_A) = Pr(X_D = \mu_D) = 1$:*

(i) *Streaming is optimal if, and only if, $\beta \leq \beta^*(\alpha)$ for some increasing threshold function $\beta^*(\alpha) : [0, 1] \rightarrow [0, 1]$.*

(ii) *The S region, A^S , is increasing (in the set sense) in μ_A , and decreasing in μ_D .*

This structure is illustrated graphically for a numerical example in Figure 3. This leads us to a conjecture regarding the *where* question: streaming should be more appealing to an ED with (1) a higher percentage of A's, (2) a lower β , and (3) a larger gap between the service times for A's and D's. We will test this conjecture via simulation in Section 6. We also note that the percentage of A patients at Flinders is relatively high ($\alpha = 43\%$) compared to the average rate of admission in the U.S. ED's, which was $\alpha = 12.8\%$ in 2006 (NHAMCS by Pitts et al. (2008)). This may be one reason that streaming was considered a success at Flinders.

These results may give the impression that streaming is only effective when we rule out the PD policy. However, as we noted above, the behavior shown in Figure 2 requires the assumption that the fraction of A patients (α) is known and constant. If the percentage of A patients varies from day to day, then a pooling policy that prioritizes either A or D patients may not be effective. 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 occurs, performance could be very poor. We illustrate this in Figure 4, which plots the optimality gap (i.e., difference between the

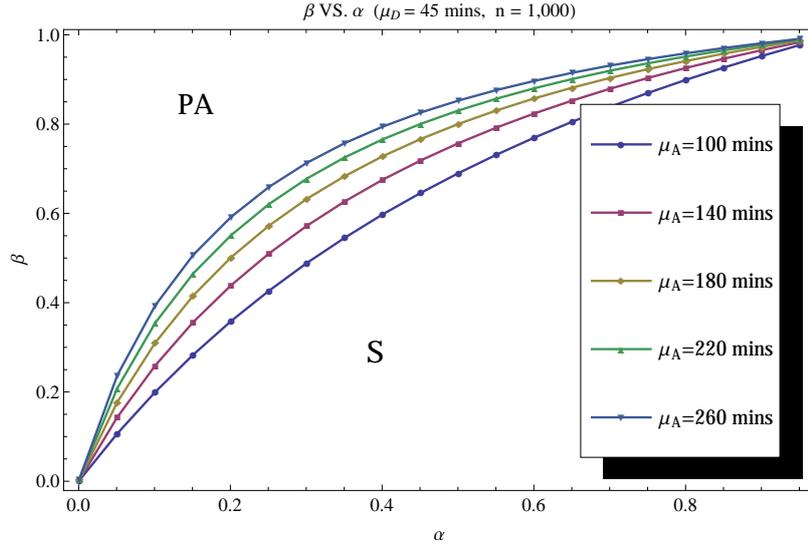


Figure 3 The effect of increasing care time of A patients on the optimal strategy. The attractiveness of streaming increases in the size of the gap between care times of A and D patients.

objective function of a given policy and that of the optimal policy) for the S, PA and PD policies. These results show that while 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, streaming is much more *robust* to changes in α than is pooling.

To examine the impact of uncertainty in α , we assume α is chosen from a family of Beta distributions given by $\text{Beta}(f(x), 2f(x))$, where $f(x) = (2 - 9x)/(27x)$, $x \in (0, 2/9)$. This results in $\mu_\alpha = 1/3$, which approximates the fraction of A's 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 . Figure 5 illustrates the impact of varying α 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 4 and 5, we can make an important conjecture (that we will test in Section 6): streaming is more robust than pooling to variation in patient mix.

The intuition is that a pooling system that completely prioritizes one type of patients sequences patients in the wrong order (e.g., putting D patients at the end of the line in a system where they should be at the beginning of the line). In contrast, a streaming system gives some priority to both types of patients by “reserving” some capacity for each type. So altering the mix of patient types

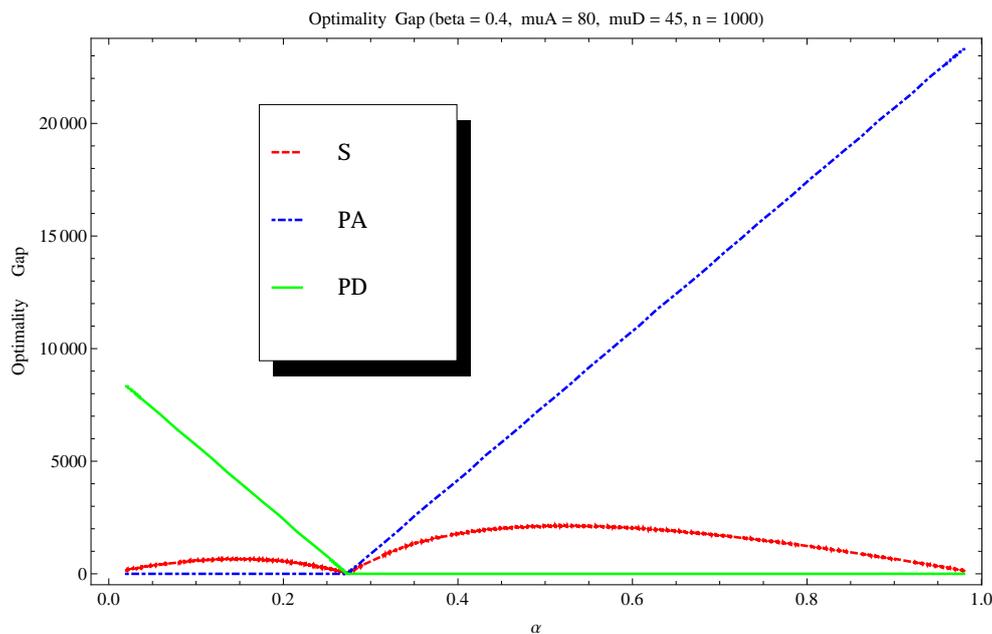


Figure 4 Sensitivity of policies to changes in α to the changes in α . Streaming is more robust to changes in patient mix than are the pooling policies.

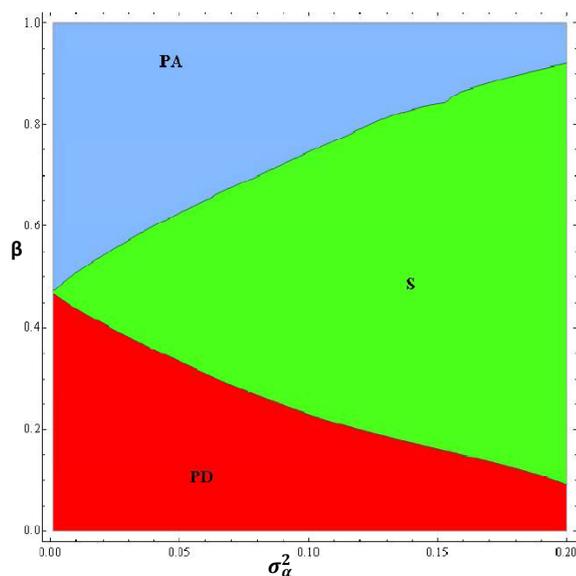


Figure 5 When the level of uncertainty in the percentage of A patients (measured by σ_α^2) increases, streaming becomes the optimal policy for an increasingly wide range of β values.

has a much more modest impact on performance in the streaming system. While we confirmed the above behavior with another analytic clearing model, we relegate discussion of that model to Online Appendix B for the sake of brevity. Online Appendix B and the above discussion leads to another important conjecture that streaming is more robust than pooling to misclassification errors. We will test this conjecture in Section 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 by the physician (designated as “treatment” states), interspersed by tests, waiting for test results and intermediate preparations/processings (designated as “wait” states), during which the patient is not available for interaction with the physician. To examine the Phase 2 sequencing decisions of what patient to see next whenever a physician completes a treatment stage, we now relax the single-stage service assumption and consider a multi-stage treatment process. Note that we still face the Phase 1 sequencing decision of what order 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 competition.

To explore the Phase 2 sequencing problem and its impact on the streaming vs. 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 multi-stage model in Figure 6. 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. After that, 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 the patient. After the final treatment by the physician, the patient experiences the final wait state (labeled as FW), which involves final preparations 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) after which 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, 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 (because of 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 the patients. Third, we assume that times in wait states are i.i.d. 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 queuing for test equipment or nurses that could serve to correlate the times in the wait states. But since there are many different activities that take place in these states, such correlation should not be large. The exponential assumption reflects the unpredictability of the activities between physician interactions. Finally, we assume that preemption is allowed. 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. 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 down stream job. In the ED, the equivalent policy would be for physicians to work on the patient closest to competition 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

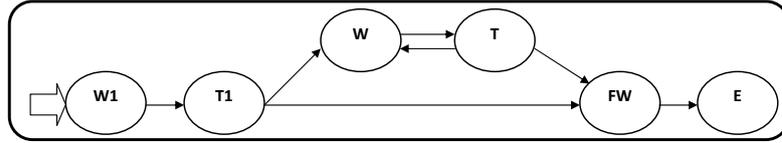


Figure 6 Multi-stage ED service: (W(1): (initial) wait, T(1):(initial) treatment, FW: final wait, E: exit).

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 *multi-armed restless bandit* model).

PROPOSITION 4 (Who to See Next?). *In the clearing model of a streaming ED flow with one physician assigned to each stream and multi-stage exponential treatment and wait stages modeled as in Figure 6:*

- (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 Section 1 by suggesting a policy simple enough to be implemented in ED's. 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." 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.

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.

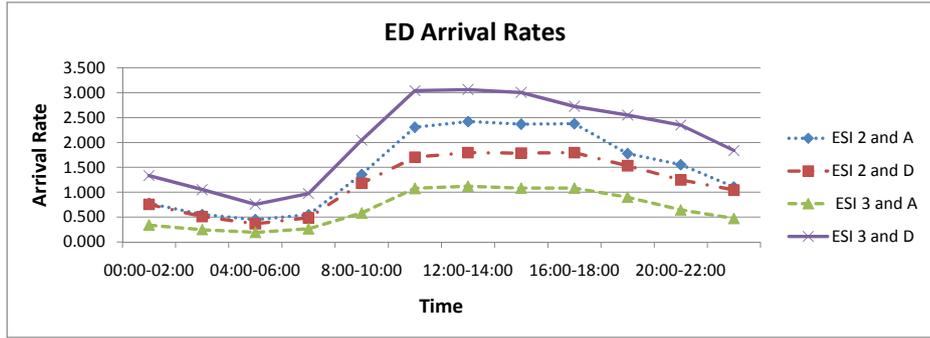


Figure 7 Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED).

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 non-stationary arrivals, multi-stage service, multiple physicians and exam rooms, inaccuracy in disposition prediction, bed-block by the hospital, among others. Our base case model was based on 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 2A, 2D, 3A, and 3D. However, the A/D prediction at the 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 non-stationary Poisson processes (which closely approximate the data) with arrival rates by class (obtained from a year of UMED data) depicted in Figure 7. The general pattern is similar to those found in other studies (e.g., Green et al. (2006)).

Service Process. The ED service process is depicted in Figure 6. 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 more details). 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).

Physician-Patient Assignments. As noted earlier, the process of connecting patients with physicians involves two phases. In Phase 1, patients are assigned to exam rooms (whenever a room becomes available) based on a Phase 1 sequencing priority. Phase one is usually performed by the charge nurse. 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. In Phase 1, prioritizing ESI 2 patients over ESI 3 patients 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 the A/D information depending on the scenario under consideration. If A/D information is not available, then Phase 1 sequencing only considers ESI levels by prioritizing ESI 2 over ESI 3 with a FCFS rule to break ties. If A/D information is available under pooling, then Phase 1 sequencing is based on the following order: 2A, 2D, 3A, 3D, with FCFS to break ties within a class.

To keeping with practice in UMED and elsewhere, we assume physicians do not take on more than seven patients at any time. To sequence these patients, 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 a 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 model represents current

practice in many ED's where physicians are not specifically encouraged to follow a specific rule, and hence, exogenous factors (changes in patient urgency level, patient discomfort, physician preference, anticipation of interactions with testing facilities, etc.) override systematic sequencing of patients.

(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 physicians 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 down-stream patient assigned to them, in order 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 down-stream 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. While new to ED's, PNPO is also an implementable policy that we argued previously is well-suited to streaming.

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 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, extra resources/capacity assigned to one stream cannot be used by the patients of the other stream. In contrast, in VS, they are only logically segregated and thus extra capacity can be shared. 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, and we use "AD+EDI"

to denote the case where, in addition to ESI levels, A/D info is used to 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 cannot be implemented in a P protocol because physicians and patient classes are not segregated under P.

By comparing the pooling and streaming policies (S and VS) under the basic SIRO Phase 2 sequencing rules, we can address the question of *whether* streaming can improve ED performance. By performing sensitivity analysis on the model parameters, we address the question of *where* streaming is most effective. And by matching streaming and pooling with various Phase 1 and Phase 2 sequencing rules, we gain insight into the question of *how* to implement streaming for maximum effectiveness.

In the next subsections, we present our main findings from the simulation experiments. For each patient flow design described above, the objective function ($\beta TTF(T(A)) + (1 - \beta)LOS(D)$) is computed as an average over 5000 replications of a week, where the result for each replication is obtained after a warmup 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 ED's and physical streaming (where, unlike virtual streaming and our analytical clearing model, capacity sharing is not possible).

OBSERVATION 1. 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 times, where 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 "anti-pooling effect," which occurs because physical separation in either physicians or beds prevents capacity sharing. To place the observed magnitude of the anti-pooling effect of physical streaming (77%) in context, we make use of Kingman's formula for a $G/G/s$ queueing system

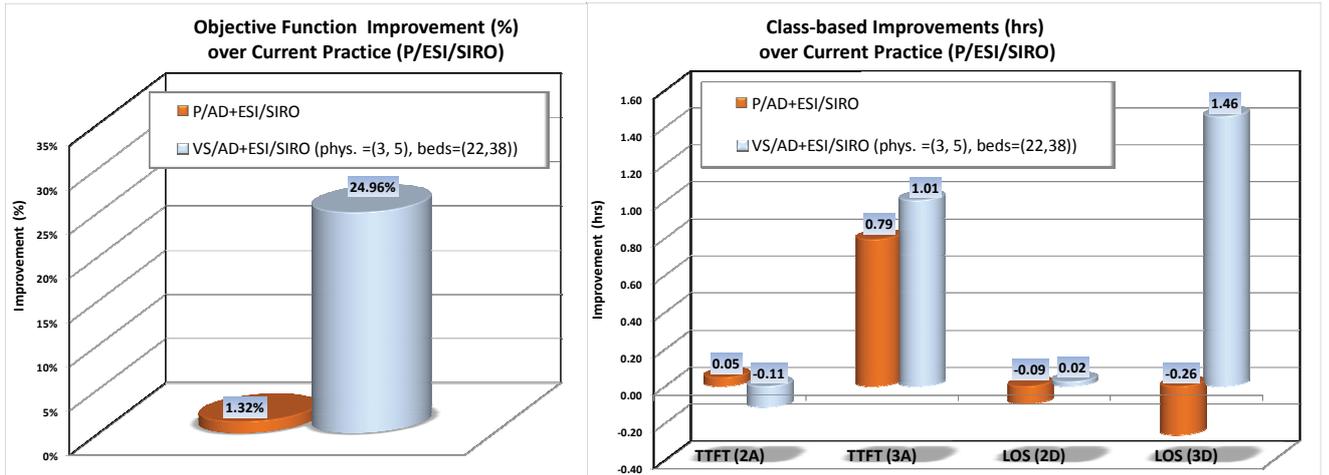


Figure 8 Virtual streaming significantly outperforms pooling and improved pooling. The reason is that VS dramatically decreases LOS for 3D patients with only a minor increase in TTFT of 2A patients. (Results for an ED with 8 physicians and 60 beds, a 20% misclassification error rate, and a weight for TTFT of A patients of $\beta = 0.50$).

with $s = 8$ physicians and two parallel $G/G/s$ queueing systems with $s = 4$ physicians each, with a server utilization close to 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. Since our objective function is a weighted average of TTFT (queue time) and LOS (system time), we would expect the anti-pooling penalty to 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 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 (unlike the $G/G/s$ example) means patients can be held in the waiting room even when physicians are idle, an effect that is more pronounced in the smaller systems under streaming (indeed, if we make the number of beds in the simulation infinite, the anti-pooling penalty falls to 17%). As a result, physical streaming is decidedly worse for performance than is a conventional pooling protocol.

Since physical streaming is so unattractive, we do not consider it further and instead we examine whether virtual streaming (VS) can improve ED performance. To this end, we consider the SIRO Phase 2 sequencing rule (as it is consistent with the current practice in most ED's) and

compare VS/AD+ESI/SIRO (basic virtual streaming) and P/AD+ESI/SIRO (improved pooling) with P/ESI/SIRO (current practice of ED's). Figure 8 depicts the simulation results. The graph on the left depicts the percentage improvement in the objective function (with $\beta = 0.5$). The graph on the right illustrates the improvement (in hrs) achieved for each class of patients separately.

This leads to the following observation about *whether* virtual streaming can outperform pooling:

OBSERVATION 2. Virtual streaming significantly outperforms both pooling and improved pooling by striking a better balance between TTFT of A's and LOS of D's.

As Figure 8 (left) shows, implementing VS can improve the overall objective function substantially (e.g., 25% in our base case model). This is achieved because VS dramatically decreases LOS of 3D's while only slightly increasing TTFT of 2A's (see Figure 8 (right)). Since VS does not require any physical reconfiguration of the ED, this finding strongly suggests that virtual streaming can be an attractive and practical option for improving ED responsiveness.

Having answered the *whether* question, we now seek to answer *how* VS should be implemented for maximum benefit. Proposition 4 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. This is due to the fact that ED physicians often do not have many patients to choose among, due to the amount of time patients are unavailable while waiting for test results. In ED's 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 value of the PNPO policy over a wide range of environments, we considered an ED with 50% shorter test times than UMED, as well as higher case loads (12 vs. 7) and 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). But physicians also engage in many indirect activities (e.g., staff management,

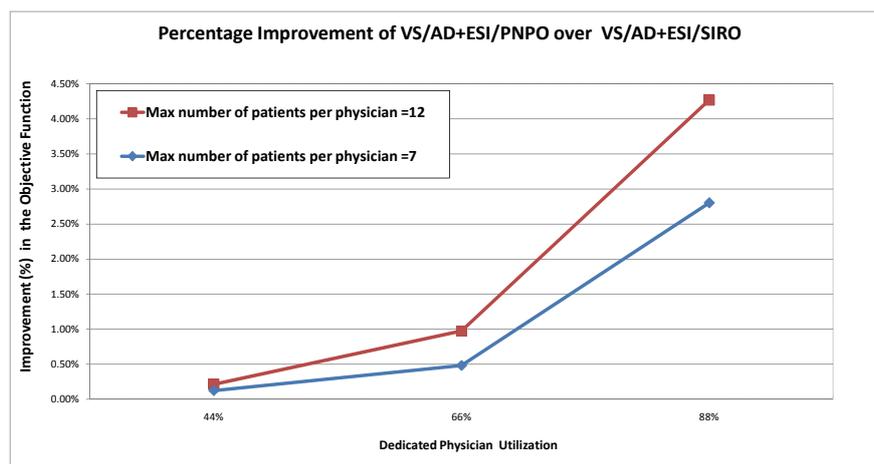


Figure 9 The benefit of implementing PNPO sequencing rule. ED's with a higher physician utilization or with a higher maximum number of patients allowed per physician benefit more from PNPO.

paper work, discussions with colleagues) that are preemptible and hence do not contribute to patient queueing. Studies report that direct care activities represent only a 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. But here we are only concerned with dedicated utilization, since this is what drives congestion.

The percentage improvement due to implementing the PNPO policy is shown in Figure 9 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 on the amount of improvement possible via better Phase 2 sequencing of an additional 4%, which when combined with the benefit of virtual streaming, results in a 29% improvement in the overall objective function when 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 Section 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 in order to identify key factors that amplify the advantage of implementing virtual streaming over pooling.

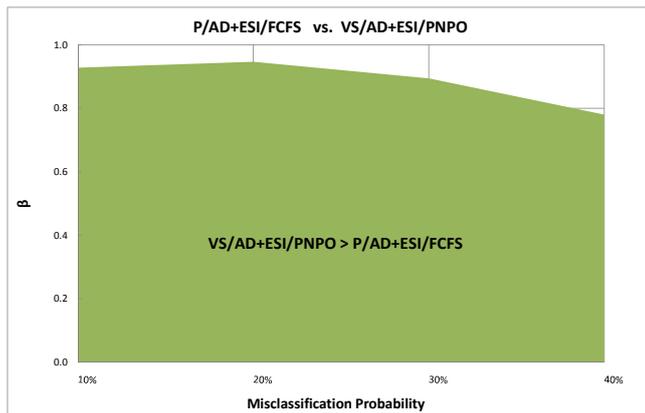


Figure 10 Sensitivity of virtual streaming and pooling designs. Lower weight of TTFT of A patients (β) or misclassification probability favors virtual streaming over pooling.

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, which was used in 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 phase 2 sequencing rules.

We start by examining the role of misclassification errors and β (the relative weight given to TTFT of A patients) on the relative benefit of virtual streaming over pooling. Based on our earlier clearing model, we conjectured that higher β should favor pooling. Common sense suggest that A/D information is less valuable if it is inaccurate, so we also expect a higher misclassification probability to also favor pooling. Figure 10 confirms these conjectures and shows that unless an ED gives a very 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 Section 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 11 (left) shows simulation results indicating that virtual streaming is indeed more attractive in ED's with a higher percentage of A patients. Figure 11 (right) shows the effect of increasing day-to-day variation in the mix of patients by drawing α from a family of beta distributions, $\text{Beta}(f(x), 2f(x))$ 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 variance to vary as $\sigma_\alpha^2 = x$, $x \in (0, 2/9)$.) This indicates that higher variability in α also makes virtual streaming more attractive, as our analytic models predicted.

We can summarize our observations concerning the effect of patient mix on the attractiveness of virtual streaming as follows.

OBSERVATION 3. 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.

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 Section 4, based on our examination of a clearing model, we conjectured that streaming becomes more attractive as the difference between service times of A's and D's increases (see Proposition 3). In Figure 12, 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 13, we similarly consider the sensitivity of these two configurations to increases in the treatment times of A (left) and D (right) patients. These lead to the following:

OBSERVATION 4. 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, ED's with congested or slow test facilities (which are used more frequently by

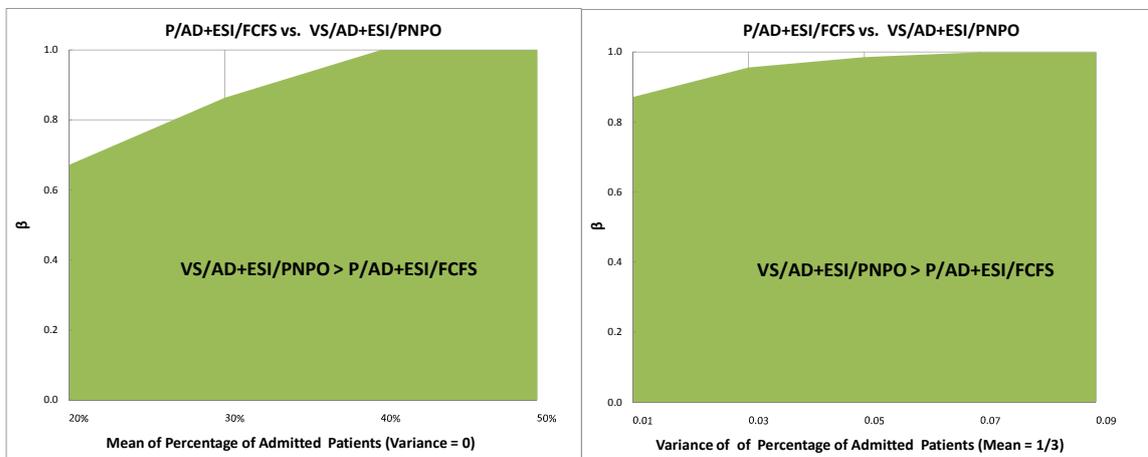


Figure 11 Sensitivity of virtual streaming and pooling designs with respect to mean μ_α (left) and variance σ_α^2 (right) of the percentage of A patients.

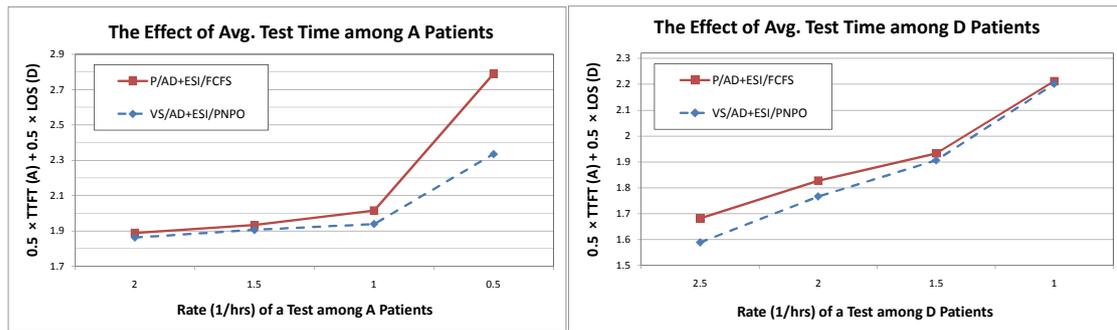


Figure 12 The effect of average patient test time (MRI, CT Scan, etc.) on the relative performance of two virtual streaming and pooling configurations. As test time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling.

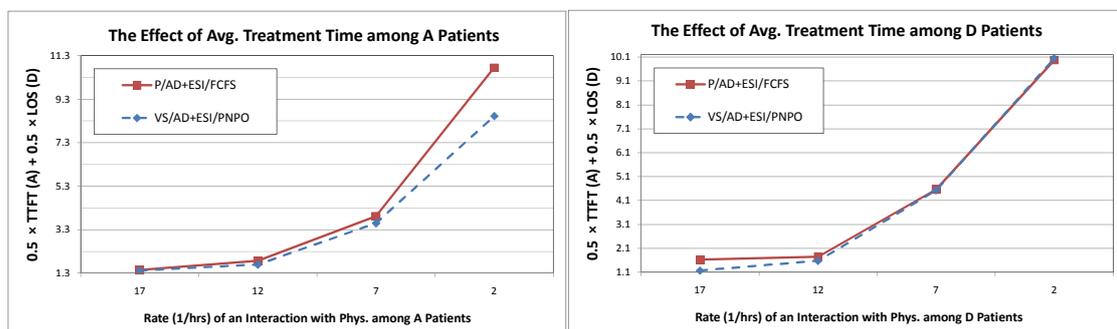


Figure 13 The effect of treatment times on the relative performance of two virtual streaming and pooling configurations. As treatment time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling.

A's than by D's) are likely to benefit more from virtual streaming than ED's with fast or ample test facilities. Second, ED's 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 ED's 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 ED's, 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 14 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. This leads to the following:

OBSERVATION 5. The relative attractiveness of virtual streaming over pooling increases with the

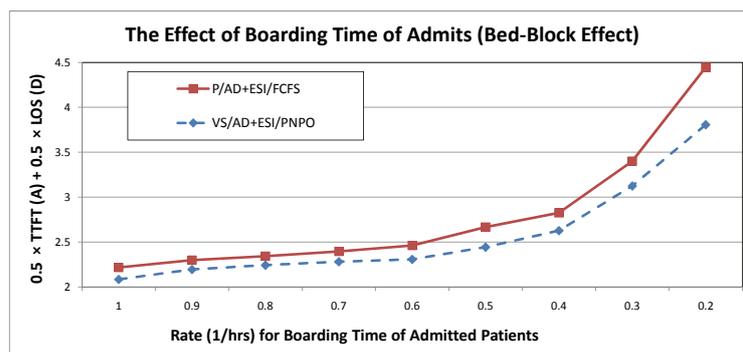


Figure 14 The effect of the average boarding time on the performance of two virtual streaming and pooling configurations. ED's with longer boarding of A's benefit more from virtual streaming.

average boarding time of A patients.

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

Finally, we consider the effect of average dedicated utilization of physician on the attractiveness of virtual streaming. Figure 15 (left) depicts the objective function for policies VS/AD+ESI/PNPO and P/AD+ESI/FCFS, while Figure 15 (right) shows the improvement in the objective function from implementing VS/AD+ESI/PNPO instead of P/AD+ESI/FCFS. These lead to the following:

OBSERVATION 6. The relative attractiveness of virtual streaming over pooling increases with average dedicated utilization of physicians.

The implication is that congested ED's with high arrival rates or low number of physicians can benefit more from virtual streaming. Furthermore, because we did not explicitly account for physician interruptions, such as treating ESI-1 patients or dealing with other non-patient issues, which would add to physician's non-preemptible activities (and hence dedicated utilization), our estimates of the benefits of virtual streaming should be conservative.

7. Conclusion

This paper describes our investigation of a new approach to managing patient flows in ED's, called streaming, which separates patients based on an up-front prediction on their final 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 professional, they are

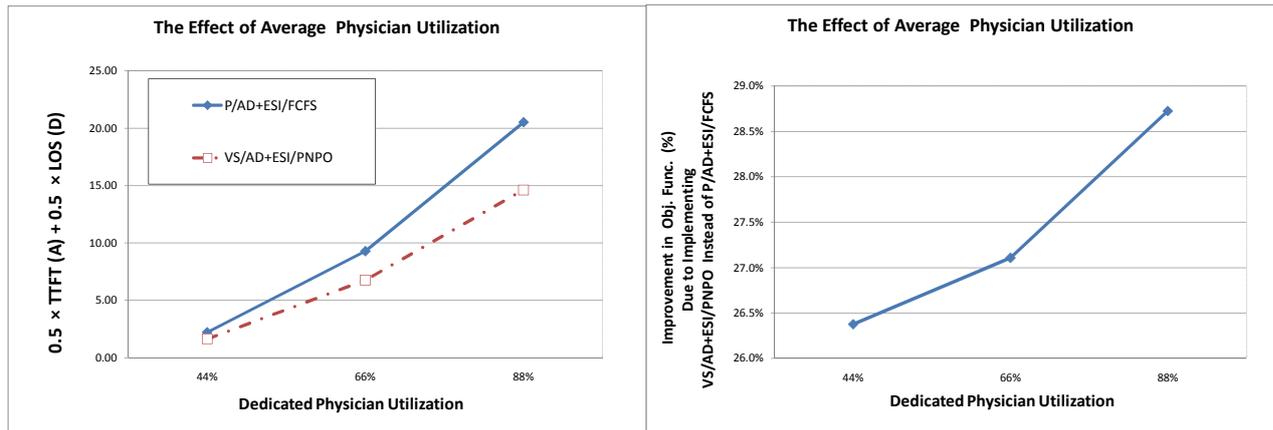


Figure 15 The effect of average physician utilization on the attractiveness of virtual streaming. ED’s with higher average physician utilization benefit more.

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 ED’s (e.g., the fractions of A patients at Flinders is substantially above the norm). Indeed, our results suggest that the physical streaming approach described by the Flinders should actually degrade ED performance through an “anti-pooling” effect caused by separating patients into streams. Hence, we suspect that the Flinders success is partly due to informal capacity sharing to overcome the anti-pooling effect and partly due to other process improvements.

To avoid the anti-pooling 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 LOS of D patients by devoting some capacity to each patient type, rather than giving full priority to either. 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). Since implementing virtual streaming does not require a physical layout redesign in the ED, it provides a practical as well as high-performance option that can be easily implemented to improve the responsiveness in ED’s.

We also found that the information used to stream patients (i.e., A or D classification) can be

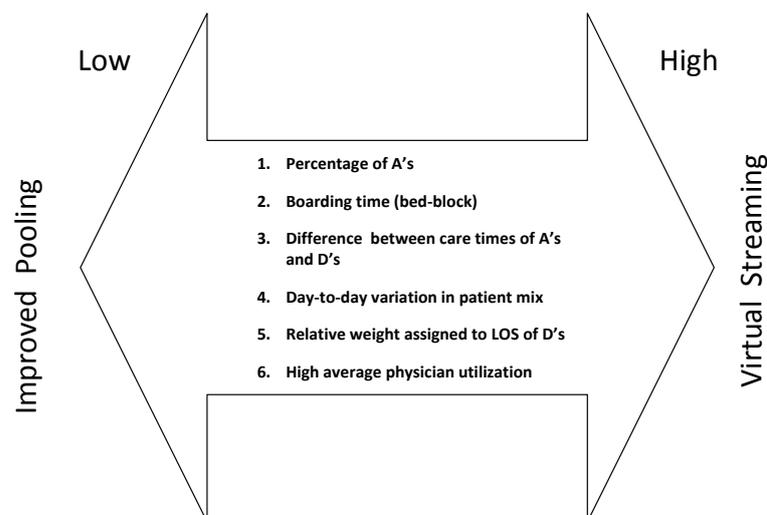


Figure 16 ED patient flow design strategy based on key environmental characteristics of the ED.

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 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 ED's. Figure 16 summarizes the results of our sensitivity analyses, which suggest that virtual streaming is best suited for ED's with (1) a high percentage of A patients, (2) longer service times for A's than D's, (3) a high relative weight assigned to LOS of D's compared to TTFT of A's, (4) long patient boarding times due to bed-block, (5) high day-to-day variation in patient mix, and (6) high average physician utilization. Using a PNPO Phase 2 sequencing rule is more effective in ED's 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 up front 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.

References

- Allon, G., S. Deo, W. Lin. 2010. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. Working Paper, Kellogg School of Business.
- American College of Emergency Physicians. 2006. American college of emergency physicians national report card on the state of emergency medicine. Available at <http://www.acep.org/assets/0/16/648/1994/00FA9DFA-9B89-4DA8-A3D8-5FBD37DD858D.pdf>.
- Andradóttir, S., H. Ayhan, D.G. Down. 2003. Dynamic server allocation for queueing networks with flexible servers. *Oper. Res.* **51**(6) 952–968.
- Argon, N.T., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management* **11**(4) 674–693.
- Ben-Tovim, D. I., J. E. Bassham, D. M. Bennett, M. L. Dougherty, M. A. Martin, S. J. O'Neill, J. L. Sincock, M. G. Szwarcbord. 2008. Redesigning care at the Flinders Medical Centre: clinical process redesign using lean thinking. *Medical Journal of Australia* **188**(6) 27–31.
- Duenyas, I., D. Gupta, T. Olsen. 1998. Control of a single-server tandem queueing system with setups. *Oper. Res.* **46** 218–230.
- Fernandes, C.M., M.R. Daya, S. Barry, N. Palmer. 1994. Emergency department patients who leave without seeing a physician: The Toronto hospital experience. *Annals of Emergency Medicine* **24**(6) 1092–1096.
- Graff, L. G., S. Wolf, R. Dinwoodie, D. Buono, D. Mucci. 1993. Emergency physician workload: A time study. *Annals of Emergency Medicine* **22**(7) 1156–1163.
- Green, L. V., J. Soares, J. F. Giglio, R.A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.
- Holdgate, A., J. Morris, M. Fry, M. Zecevic. 2007. Accuracy of triage nurses in predicting patient disposition. *Emergency Medicine Australasia* **19** 341–345.
- Hollingsworth, J.C., C.D. Chisholm, B.K. Giles, W.H. Cordell, D.R. Nelson. 1998. How do physicians and nurses spend their time in the emergency department. *Annals of Emergency Medicine* **31**(1) 87–91.
- Hoot, N. R., D. Aronsky. 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2) 126–136.
- Hopp, W.J., S.M.R. Iravani, B. Shou. 2005. Serial agile production systems with automation. *Oper. Res.* **53**(5) 852–866.
- Howell, E.E., E.D. Bessman, H.R. Rubin. 2004. Hospitals and innovative emergency department admission process. *Journal of General Internal Medicine* **19** 266–2686.
- Hu, B., S. Benjafar. 2009. Partitioning of servers in queueing systems during rush hour. *Manufacturing Service Oper. Management* **11**(3) 416–428.
- Khare, R.K., E.S. Powell, G. Reinhardt, M. Lucenti. 2008. Adding more beds to the emergency department or reducing admitted patient boarding times: Which has a more significant influence on emergency department congestion? *Annals of Emergency Medicine* **53**(5) 575–585.

- King, D. L., D. I. Ben-Tovim, J. Bassham. 2006. Redesigning emergency department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia* **18** 391–397.
- Kinsman, L., R. Champion, G. Lee, M. Martin, K. Masman, E. May, T. Mills, M.D. Taylor, P. Thomas, R.J. Williams, S. Zalstein. 2008. Assessing the impact of streaming in a regional emergency department. *Emergency Medicine Australasia* **20** 221–227.
- Kochran, J., K.T. Roche. 2009. A multi-class queueing network analysis methodology for improving hospital emergency department performance. *Comput. and Oper. Res.* **36**(1) 1497–1512.
- Kronick, S.L., J.S. Desmond. 2009. Blink: Accuracy of physician estimates of patient disposition at the time of ED triage. SAEM Midwest Regional Meeting.
- Liew, D., D. Liew, M.P. Kennedy. 2003. Emergency department length of stay independently predicts excess inpatient length of stay. *Medical Journal of Australia* **179**(17) 524–526.
- Mandelbaum, A., M.I. Reiman. 1997. On pooling in queueing networks. *Management Sci.* **44**(7) 971–981.
- Miro, O., M. Sanchez, G. Espinosa, B. Coll-Veient, E. Bragulat, J. Milla. 2003. Analysis of patient flow in the emergency department and the effect of an extensive reorganisation. *Emergency Medicine Journal* **20** 143–148.
- Pitts, S. R., R. W. Niska, J. Xu, C. W. Burt. 2008. National hospital ambulatory medical care survey: 2006 emergency department summary. *National Health Statistics Report* **7** 1–39.
- Richardson, D. 2003. Reducing patient times in emergency departments. *Medical Journal of Australia* **179**(17) 516–517.
- Rothkopf, M.H., P. Rech. 1987. Perspective on queueing: Combining queues is not always beneficial. *Oper. Res.* **35** 906–909.
- Schull, M.J., A. Kiss, J.-P. Szali. 2007. The effect of low complexity patients on emergency department waiting times. *Annals of Emergency Medicine* **49**(3) 257–264.
- Solberg, L. I., B. R. Asplin, D. J. Magid R. M. Weinick. 2003. Emergency department crowding: Consensus development of potential measures. *Annals of Emergency Medicine* **42**(6) 824–834.
- SoRelle, R. 2006. Homicide charges against Illinois ED stun EM. *Emergency Medicine News* **28**(12) 1,25.
- SoRelle, R. 2010. Breaking news: Health reform and the ED: Prepare for the surge. *Emergency Medicine News* **32**(5) 1,20.
- Tekin, E., W.J. Hopp, , M.P. Van Oyen. 2009. Pooling strategies for service center agent cross-training. *IIE Transactions* **41** 546–561.
- Van Dijk, N.M., E. Van Der Sluis. 2008. To pool or not to pool in call centers. *Prod. Oper. Man.* **17**(3) 296–305.
- Van Oyen, M.P., E.G.S. Gel, W. J. Hopp. 2001. Opportunity for workforce agility in collaborative and non-collaborative work systems. *IIE Transactions* **33**(9) 761–777.
- Welch, S. J. 2008. Patient segmentation: Redesigning flow. *Emergency Medicine News* **31**(8).
- Whitt, W. 1999. Partitioning customers into service groups. *Management Sci.* **45** 1579–1592.

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 k th Admit and k th 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) $T_A^{PA}(\alpha, \omega) \leq T_A^S(\alpha, \omega)$, and (b) $T_A^{PA}(\alpha, \omega) \leq T_A^{PD}(\alpha, \omega)$. 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 k th Admit patient starts its treatment at $CA_{k-1}^S(\omega)$ but under $\pi = PA$, the k th Admit patient starts its treatment at $\min\{CA_{k-1}^{PA}(\omega), CA_{k-2}^{PA}(\omega)\} \leq \min\{CA_{k-1}^S(\omega), CA_{k-2}^S(\omega)\} \leq CA_{k-1}^S(\omega)$, 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 $t(\omega) - CA_{n_t(\omega)-1}^S(\omega) \geq 0$ 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 $CD_{n_D}^{PD}(\omega)$ 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) $L_D^{PD}(\alpha, \omega) \leq L_D^S(\alpha, \omega)$, and (2) $L_D^{PD}(\alpha, \omega) \leq L_D^{PA}(\alpha, \omega)$. To show (1), fixing α , we show

that $CD_k^{PD}(\omega) \leq CD_k^S(\omega)$ ($\forall k \in 1, 2, \dots, n_D$). 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 k th Discharge patient under sample path ω , then $TD_k^{PD}(\omega) \leq TD_k^S(\omega)$ ($\forall k \in 1, 2, \dots, n_D$). Next, if $SD_k(\omega)$ is the service time of k th 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 $CD_k^{PD}(\omega) \leq CD_k^S(\omega)$ ($\forall k \in 1, 2, \dots, n_D$), 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. \square

Proof of Lemma 1. To prove this lemma, using the definition of β -convexity, we need to show that sets \mathcal{A}^π ($\forall \pi \in \mathbf{\Pi}$) 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:

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

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

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

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:

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

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 . \square

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

$$\begin{aligned} \beta_1(\alpha) &= \inf\{\beta : f^S(\alpha, \beta) \leq f^{PD}(\alpha, \beta)\}, \\ \beta_2(\alpha) &= \sup\{\beta : f^S(\alpha, \beta) \leq f^{PA}(\alpha, \beta)\}. \end{aligned}$$

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) $0 \leq \beta(\alpha) \leq \underline{\beta}(\alpha) - \epsilon$, or (b) $\bar{\beta}(\alpha) + \epsilon \leq \beta(\alpha) \leq 1$. 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, $\underline{\beta}(\alpha) - \epsilon < \underline{\beta}(\alpha) \leq \beta_1(\alpha)$. 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 $\tilde{\beta}(\alpha) = 1$ (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 $\bar{\beta}(\alpha) + \epsilon$ and $\tilde{\beta}(\alpha) = 1$. Then, similar to the discussion of part (a), notice that by definition of $\beta_2(\alpha)$, $\pi = PA$ is optimal at $\bar{\beta}(\alpha) + \epsilon$. Moreover, by Proposition 1, $\pi = PA$ is also optimal at $\tilde{\beta}(\alpha) = 1$. Thus, from Lemma 1 we see that $\pi = PA$ should be also optimal at $\beta(\alpha)$. This completes the proof. \square

Proof of Proposition 3. To prove part (i), first notice that using a proof similar to that of Lemma 1, the optimal strategy with $\mathbf{\Pi} = \{PA, S\}$ is β -convex. Thus, defining $\beta^*(\alpha) = \sup\{\beta : f^S(\alpha, \beta) \leq f^{PA}(\alpha, \beta)\}$, $\mathcal{A}^S = \{(\alpha, \beta) : \beta \leq \beta^*(\alpha)\}$. Next, we show that $\beta^*(\alpha)$ is nondecreasing in α . (Thus, likewise, for every β , there would be a threshold $\alpha^*(\beta)$ such that $\mathcal{A}^S = \{(\alpha, \beta) : \alpha \geq \alpha^*(\beta)\}$). To this end, choose $(\alpha, \beta) \in \mathcal{A}^S$. We show then that $(\alpha', \beta) \in \mathcal{A}^S$ for every $\alpha' > \alpha$. To show this, it is sufficient to show that if $(\alpha, \beta) \in \mathcal{A}^S$ then by decreasing the number of D patients, n_D , by one, and adding one patient to A patients, n_A , thus reaching some $\alpha' > \alpha$, we would still have $f^S(\alpha', \beta) \leq f^{PA}(\alpha', \beta)$. That is, fixing β , we show that if $\pi = S$ is better than $\pi = PA$ with (n_A, n_D) patients, then it is also better with $(n_A + 1, n_D - 1)$ patients. Suppose $f^S(n_A, n_D) \leq f^{PA}(n_A, n_D)$ and define $\Delta f^\pi = f^\pi(n_A + 1, n_D - 1) - f^\pi(n_A, n_D)$. Next notice that since $pr(X_A = \mu_A) = pr(X_D = \mu_D) = 1$, it is straightforward to show that $\Delta f^S \leq \Delta f^{PA}$. Thus, adding $\Delta f^S \leq \Delta f^{PA}$ and $f^S(n_A, n_D) \leq$

$f^{PA}(n_A, n_D)$, we would get $f^S(n_A + 1, n_D - 1) \leq f^{PA}(n_A + 1, n_D - 1)$.

To prove part (ii), we show that for every $(\alpha, \beta) \in \mathcal{A}^S$, we have $(\alpha, \beta) \in \mathcal{A}^{tS}$, where \mathcal{A}^{tS} is the Streaming region with $\mu'_A \geq \mu_A$ (or $\mu'_D \leq \mu_D$). To show this, notice that it is sufficient to show that the threshold curve β^* is increasing (in the weak sense) in μ_A (decreasing in μ_D). Fix α and β and let $\beta^*(\mu_A)$ denote the threshold as a function of μ_A . Define $f_1(\mu_A) = T_A^S(\mu_A) - T_A^{PA}(\mu_A)$, $f_2(\mu_A) = L_D^{PA}(\mu_A) - L_D^S(\mu_A)$, and $g(\mu_A) = f_1(\mu_A)/f_2(\mu_A)$. Since the threshold $\beta^*(\mu_A)$ is such that the policies $\pi = PA$ and $\pi = S$ are indifferent, we have $\beta^*(\mu_A) = [1 + g(\mu_A)]^{-1}$. Thus, to show $\beta^*(\mu_A)$ is weakly increasing in μ_A , it is sufficient to show that $g(\mu_A) = f_1(\mu_A)/f_2(\mu_A)$ is weakly decreasing in μ_A . This is equivalent to $(df_1/d\mu_A)f_2 - (df_2/d\mu_A)f_1 \leq 0$ and equivalent to

$$\frac{df_1/d\mu_A}{df_2/d\mu_A} \leq \frac{f_1}{f_2}. \quad (\text{EC.3})$$

We now consider the case where, under the streaming policy, the D stream finishes before the A stream (i.e., when $n_A\mu_A \geq n_D\mu_D$) and show inequality EC.3. Other cases will follow a similar line of proof (with only slight modifications). First notice that an upper bound for $T_A^S(\mu_A + \epsilon) - T_A^S(\mu_A)$ can be obtained by assuming that all A patients are seen before moving to a Pooling scenario. In that case, the increase in TTFT for the second patient is ϵ , for the third patient is 2ϵ , and so on. Thus, since $\sum_{k=1}^{n-1} k = n(n-1)/2$, we have: $T_A^S(\mu_A + \epsilon) - T_A^S(\mu_A) \leq \epsilon n_A(n_A - 1)/(2n_A) = \epsilon(n_A - 1)/2$. Therefore, $dT_A^S/d\mu_A = \lim_{\epsilon \rightarrow 0} [T_A^S(\mu_A + \epsilon) - T_A^S(\mu_A)]/\epsilon \leq (n_A - 1)/2$. Similarly, considering that under $\pi = PA$, A patients are paired into two parallel lines we have $dL_D^S/d\mu_A = 2[(n_A/2)(n_A/2 - 1)]/(2n_A) = (n_A - 2)/4$. Hence, $df_1/d\mu_A \leq (n_A - 2)/4$. Next notice that under $\pi = PA$ every D patient is delayed at least equal to the delay of the last A patient. Therefore, $dL_D^{PA}/d\mu_A \geq n_A/2$. Moreover, since under $\pi = S$ the stream of D patients is finished before the stream of A patients (i.e., all D patients are seen before moving to a Pooling scenario), we have $dL_D^S/d\mu_A = 0$. Hence, $df_2/d\mu_A \geq n_A/2$. Thus we have:

$$\frac{df_1/d\mu_A}{df_2/d\mu_A} \leq \frac{(n_A - 2)/4}{n_A/2} = \frac{1}{2} - \frac{1}{n_A}. \quad (\text{EC.4})$$

Next, notice that since $\sum_{k=1}^{n-1} k = n(n-1)/2$ we have:

$$f_1(\mu_A) = T_A^S(\mu_A) - T_A^{PA}(\mu_A) \geq (n_A - 1)\mu_A/2 - 2(n_A/2)(n_A/2 - 1)\mu_A/(2n_A) = n_A\mu_A/4.$$

Additionally, $L_D^{PA} \leq (n_A/2)\mu_A + 2(n_D/2)(n_D/2 - 1)\mu_D/(2n_D) = n_A\mu_A/2 + (n_D + 2)\mu_D/4$. Also, $L_D^S = n_D(n_D + 1)\mu_D/(2n_D) = (n_D + 1)\mu_D/2$. Hence, we have:

$$f_2(\mu_A) = L_D^{PA}(\mu_A) - L_D^S(\mu_A) \leq n_A\mu_A/2 - n_D\mu_D/4 \leq n_A\mu_A/2.$$

From above we have:

$$\frac{f_1}{f_2} \geq \frac{n_A\mu_A/4}{n_A\mu_A/2} = 1/2. \quad (\text{EC.5})$$

Now combining inequalities (EC.4) and (EC.5) we have:

$$\frac{f_1}{f_2} \geq \frac{1}{2} > \frac{1}{2} - \frac{1}{n_A} \geq \frac{df_1/d\mu_A}{df_2/d\mu_A}.$$

Hence, (EC.3) holds and therefore the threshold $\beta^*(\mu_A)$ is weakly increasing in μ_A as was required.

The analyses to show $\beta^*(\mu_D)$ is weakly decreasing in μ_D would also follow the similar line of proof after swapping the A/D labels of patients. \square

Proof of Proposition 4 - Part (i). 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 6. The i th 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 6). 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 $\mathbf{X} = (x_1, x_2, \dots, x_{\bar{k}+1})$ and $\mathbf{Y} = (y_1, y_2, \dots, y_{\bar{k}+1})$, where x_i and y_i denote the number of patients in i th 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 $(N, 0, \dots, 0)$ to state $(0, 0, \dots, N)$ with the minimum expected average LOS or equivalently with the minimum sum of patient competition 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 $J(0, 0, \dots, N) = 0$):

$$\begin{aligned}
J(\mathbf{X}, \mathbf{Y}) = & \frac{1}{\psi} \left[\sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \right. \\
& + \mu \min_{l \in \mathcal{L}(\mathbf{X})} \left\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l = k\} [p_k J(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1})] \right\} \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J(\mathbf{X} + e_i, \mathbf{Y} - e_i) \\
& \left. + (\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\{\sum_{ki=1}^{\bar{k}} x_i \geq 1\}) J(\mathbf{X}, \mathbf{Y}) \right], \tag{EC.6}
\end{aligned}$$

where e_k is a row vector of size $\bar{k} + 1$ with a one in k th 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 $n \in \mathbb{N}$:

$$J_{n+1}(\mathbf{X}, \mathbf{Y}) = \frac{1}{\psi} \left[\sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \right]$$

$$\begin{aligned}
& + \mu \min_{l \in \mathcal{L}(\mathbf{x})} \left\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l=k\} [p_k J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1-p_k) J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1})] \right\} \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_n(\mathbf{X} + e_i, \mathbf{Y} - e_i) \\
& + \left(\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\left\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \right\} \right) J_n(\mathbf{X}, \mathbf{Y}), \tag{EC.7}
\end{aligned}$$

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 $J_n(\mathbf{X}, \mathbf{Y}) \rightarrow J(\mathbf{X}, \mathbf{Y})$ as $n \rightarrow \infty$ 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.7), 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.7), to show that serving stage k^* in period $n + 1$ is optimal, it is sufficient to show that for any such k , we have:

$$\begin{aligned}
\underline{\text{Property i:}} \quad & p_{k^*} J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*}) J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \\
& \leq p_k J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}). \tag{EC.8}
\end{aligned}$$

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

$$\begin{aligned}
\underline{\text{Property ii:}} \quad & p_k^* J_n(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_k^*) J_n(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \\
& p_k J_n(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_n(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}). \tag{EC.9}
\end{aligned}$$

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.8) 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:

$$\begin{aligned}
J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) &\leq \frac{1}{\psi} \left[\left(\sum_{i=1}^{\bar{k}} x_i - 1 \right) + \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) \right. \\
&\quad + \mu \left[p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1}) \right. \\
&\quad \quad \left. + (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \right] \\
&\quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{\bar{k}+1}) \\
&\quad + \gamma J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) \\
&\quad \left. + \left(\psi - \gamma \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) - \mu \mathbb{1} \left\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \right\} \right) J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \right].
\end{aligned} \tag{EC.10}$$

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

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

Now multiplying both sides of (EC.10) by p_{k^*} , both sides of (EC.11) by $(1 - p_{k^*})$, and summing up the results we have:

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

$$\begin{aligned}
& + (1 - p_{k^*}) (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{k+1})] \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i [p_{k^*} J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{\bar{k}+1}) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} + e_i - e_{k^*}, \mathbf{Y} - e_i + e_{k^*+1})] \\
& + \gamma [p_{k^*} J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y})] \\
& + \bar{\psi} (p_{k^*} J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}))], \tag{EC.12}
\end{aligned}$$

where, for the ease of notation, we let $\bar{\psi}$ denote the self-loop rate, i.e., $\bar{\psi} = (\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\{\sum_{i=1}^{\bar{k}} x_i \geq 1\})$. 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:

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

Thus, we have obtained an upper bound for the LHS of (EC.8). Now consider the RHS of (EC.8) and first for state $(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1})$ use (EC.7) 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.7) we have:

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

$$\begin{aligned}
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{\bar{k}+1}) \\
& + \gamma J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) \\
& + \left(\psi - \gamma \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) - \mu \mathbb{1} \left\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \right\} \right) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}).
\end{aligned} \tag{EC.14}$$

Similarly, using (EC.7) to obtain $J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1})$ we have:

$$\begin{aligned}
J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) &= \frac{1}{\psi} \left[\left(\sum_{i=1}^{\bar{k}} x_i - 1 \right) + \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) \right. \\
& + \mu \left[p_{k^*} J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \right. \\
& \quad \left. + (1 - p_{k^*}) J_{n-1}(\mathbf{X} - e_{k^*} - e_k, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \right] \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{k+1}) \\
& + \gamma J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \\
& + \left(\psi - \gamma \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) - \mu \mathbb{1} \left\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \right\} \right) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \left. \right].
\end{aligned} \tag{EC.15}$$

Now multiplying both sides of (EC.14) by p_k , both sides of (EC.15) by $(1 - p_k)$, and summing up the results we have:

$$\begin{aligned}
& p_k J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) = \\
& \frac{1}{\psi} \left[\left(\sum_{i=1}^{\bar{k}} x_i - 1 \right) + \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) + \mu \left[p_{k^*} p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1} + e_{\bar{k}+1}) \right. \right. \\
& \quad + p_{k^*} (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1} + e_{\bar{k}+1}) \\
& \quad + (1 - p_{k^*}) p_k J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{\bar{k}+1}) \\
& \quad \left. \left. + (1 - p_{k^*}) (1 - p_k) J_{n-1}(\mathbf{X} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1} + e_{k+1}) \right] \right. \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i \left[p_k J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{\bar{k}+1}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_i - e_k, \mathbf{Y} - e_i + e_{k+1}) \right] \\
& + \gamma \left[p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \right] \\
& \left. + \bar{\psi} \left(p_k J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k) J_{n-1}(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \right) \right],
\end{aligned} \tag{EC.16}$$

where, for the ease of notation, we again let $\bar{\psi} = \left(\psi - \gamma \left(\sum_{i=1}^{\bar{k}+1} y_i + 1 \right) - \mu \mathbb{1} \left\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \right\} \right)$. Notice that RHS of (EC.16) is equal to the upper bound of the LHS of (EC.8) derived in (EC.13). 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.9) 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.7) 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.8):

$$\begin{aligned}
& p_{k^*} J_n(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_{k^*}) J_n(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \leq \\
& \frac{1}{\psi} \left[-p_{k^*} + \left(\sum_{i=1}^{\bar{k}} x_i \right) + \left(\sum_{i=1}^{\bar{k}+1} y_i \right) \right. \\
& \quad + \mu \left[p_{k^*} p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \right. \\
& \quad \quad + p_{k^*} (1 - p_k) J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1}) \\
& \quad \quad + (1 - p_{k^*}) p_k J_{n-1}(\mathbf{X} + e_{k^*+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \\
& \quad \quad \left. \left. + (1 - p_{k^*}) (1 - p_k) J_{n-1}(\mathbf{X} + e_{k^*+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k+1}) \right] \right. \\
& \quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i \left[p_{k^*} J_{n-1}(\mathbf{X} + e_{\bar{k}+1} + e_i - e_{k^*}, \mathbf{Y} - e_i) + (1 - p_{k^*}) J_{n-1}(\mathbf{X} + e_{k^*+1} + e_i - e_{k^*}, \mathbf{Y} - e_i) \right] \\
& \quad \left. + \psi \left(p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k^*+1} - e_{k^*}, \mathbf{Y}) \right) \right]. \tag{EC.17}
\end{aligned}$$

Now, using the optimality equation (EC.7) 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.8). (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 $J_n(\mathbf{X} + e_{k+1} - \mathbf{e}_k, \mathbf{Y})$.)

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

$$\begin{aligned}
& + (1 - p_{k^*}) (1 - p_k) J_{n-1}(\mathbf{X} + e_{k+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1})] \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i \left[p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} + e_i - e_k, \mathbf{Y} - e_i) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k+1} + e_i - e_k, \mathbf{Y} - e_i) \right] \\
& + \psi \left(p_k J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k) J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \right) \Big]. \tag{EC.18}
\end{aligned}$$

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.17) is less than or equal to (EC.18), which establishes Property ii for n and completes the proof. \square

Proof of Proposition 4 - 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 6. Let $w_1^n(\omega)$ be the realized duration of the initial waiting stage, W_1 , for patient $n \in \{1, \dots, N\}$ 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 $n \in \{1, \dots, N\}$, since a patient cannot be 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 $n \in \{1, \dots, N\}$). 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). \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:

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

To isolate the effect of misclassification errors, we eliminate variability in the treatment times, X_A and X_D so that $Pr(X_A = \mu_A)$ and $Pr(X_D = \mu_D)$. 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:

$$\begin{aligned}E[TTFT_A^S | N_A = n_A, \tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D] &= \frac{1}{(1 - \tilde{\gamma}_D)n_A + \tilde{\gamma}_A(n - n_A)} \times \\ & \left[(1 - \tilde{\gamma}_D) \left[\sum_{j=1}^{\tilde{n}_A} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-k-1} (k\mu_D + (j-k-1)\mu_A) \right. \right. \\ & + \sum_{j=\tilde{n}_A+1}^{n_A} \left[\sum_{k=0}^{\tilde{n}_A} \binom{\tilde{n}_A}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{\tilde{n}_A-k} (k\mu_D + (\tilde{n}_A - k)\mu_A) \right. \\ & \left. \left. + \sum_{k=0}^{j-\tilde{n}_A-1} \binom{j-\tilde{n}_A-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-\tilde{n}_A-1-k} \left(k \frac{\mu_D}{2} + (j - \tilde{n}_A - 1 - k) \frac{\mu_A}{2} \right) \right] \right]\end{aligned}$$

$$\begin{aligned}
& + \tilde{\gamma}_A \left[\sum_{j=1}^{\tilde{n}_D} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-k-1} (k\mu_A + (j-k-1)\mu_D) \right. \\
& + \sum_{j=\tilde{n}_D+1}^{n-n_A} \left[\sum_{k=0}^{\tilde{n}_D} \binom{\tilde{n}_D}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{\tilde{n}_D-k} (k\mu_A + (\tilde{n}_D-k)\mu_D) \right. \\
& \quad \left. \left. + \sum_{k=0}^{j-\tilde{n}_D-1} \binom{j-\tilde{n}_D-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-\tilde{n}_D-1-k} \left(k\frac{\mu_A}{2} + (j-\tilde{n}_D-1-k)\frac{\mu_D}{2} \right) \right] \right].
\end{aligned} \tag{EC.19}$$

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 j th 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 j th 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:

$$\overline{TTFT}_A^S = E[E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D]] = \sum_{n_A=0}^n \sum_{\tilde{n}_A=0}^n \sum_{\tilde{n}_D=0}^n E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D] g(n_A, \tilde{n}_A, \tilde{n}_D), \tag{EC.20}$$

where $E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D]$ is computed in (EC.19). 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:

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

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

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

$$Pr(N_A = n_A) [Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D) \mathbb{1}\{\tilde{n}_D < n - n_A = n_D\} \tag{EC.21}$$

$$+ Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A) \mathbb{1}\{\tilde{n}_A < n_A\} \tag{EC.22}$$

$$+ Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \mathbb{1}\{\tilde{n}_A = n_A, \tilde{n}_D = n - n_A\}] \tag{EC.23}$$

In above, Eq's (EC.21), (EC.22), and (EC.23) 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:

$$Pr(N_A = n_A) = \binom{n}{n_A} p^{n_A} (1 - p)^{n - n_A}. \quad (\text{EC.24})$$

Let K_j^A and K_j^D be the random variables denoting the number of D type patients up to (and including) the j th patient in A and D streams, respectively. Then to compute (EC.21), 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:

$$\begin{aligned} & Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D) \\ &= Pr((\tilde{n}_D - K_{\tilde{n}_D}^D)\mu_A + K_{\tilde{n}_D}^D\mu_D \leq (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D) \\ &- Pr((n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D \geq (\tilde{n}_D + 1 - K_{\tilde{n}_D+1}^D)\mu_A + K_{\tilde{n}_D+1}^D\mu_D) \\ &= Pr(K_{n_A}^A - K_{\tilde{n}_D}^D \leq \mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}) \\ &- Pr(K_{n_A}^A - K_{\tilde{n}_D+1}^D \leq \mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}) \\ &= F_1\left(\mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}\right) - F_2\left(\mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}\right) \end{aligned} \quad (\text{EC.25})$$

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.22), we have:

$$\begin{aligned} & Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\ &= Pr((\tilde{n}_A - K_{\tilde{n}_A}^A)\mu_A + K_{\tilde{n}_A}^A\mu_D \leq (n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D) \\ &- Pr((n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D \geq (\tilde{n}_A + 1 - K_{\tilde{n}_A+1}^A)\mu_A + K_{\tilde{n}_A+1}^A\mu_D) \\ &= Pr(K_{n_D}^D - K_{\tilde{n}_A}^A \leq \mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}) \\ &- Pr(K_{n_D}^D - K_{\tilde{n}_A+1}^A \leq \mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}) \\ &= F_3\left(\mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}\right) - F_4\left(\mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}\right) \end{aligned} \quad (\text{EC.26})$$

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

Next, to compute (EC.23), we need to compute the probability that one stream finishes when the system is working on the last patient of the other stream:

$$\begin{aligned}
& Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\
&= Pr(T_{n_A-1}^A < T_{n_D}^D \leq T_{n_A}^A) + Pr(T_{n_D-1}^D < T_{n_A}^A < T_{n_D}^D) \tag{EC.27} \\
&= Pr(((n_A - 1) - K_{n_A-1}^A)\mu_A + K_{n_A-1}^A\mu_D < (n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D \leq (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D) \\
&+ Pr(((n_D - 1) - K_{n_D-1}^D)\mu_A + K_{n_D-1}^D\mu_D < (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D < (n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D) \\
&= Pr(K_{n_D}^D - K_{n_A-1}^A < \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) + Pr(K_{n_A}^A - K_{n_D-1}^D < \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}) \\
&- Pr((n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D > (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D) \\
&- Pr((n_D - K_{n_D}^D)\mu_A + K_{n_D}^D\mu_D \leq (n_A - K_{n_A}^A)\mu_A + K_{n_A}^A\mu_D) \\
&= Pr(K_{n_D}^D - K_{n_A-1}^A < \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) + Pr(K_{n_A}^A - K_{n_D-1}^D < \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}) - 1 \\
&= F_5(\mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) + F_6(\mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}) - 1 \\
&- Pr(K_{n_D}^D - K_{n_A-1}^A = \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) - Pr(K_{n_A}^A - K_{n_D-1}^D = \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D})
\end{aligned}$$

where T in (EC.27) is used to show the finish times of corresponding jobs, and $F_5(\cdot)$ and $F_6(\cdot)$ are CDFs of random variables $Z_5 = K_{n_D}^D - K_{n_A-1}^A$ and $Z_6 = K_{n_A}^A - K_{n_D-1}^D$. Now notice that random variables Z_1, \dots, Z_6 are each the difference between two independent binomial random variables with known parameters. Thus, CDFs F_1, \dots, F_6 are known. Therefore, $g(n_A, \tilde{n}_A, \tilde{n}_D)$ can be computed.

As a result, the metric \overline{TFT}_A^S is completely computed.

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

$$\overline{LOS}_D^S = E[E[LOS_D^S | N_A, \tilde{N}_A, \tilde{N}_D]] = \sum_{n_A=0}^n \sum_{\tilde{n}_A=0}^N \sum_{\tilde{n}_D=0}^N E[LOS_D^S | N_A, \tilde{N}_A, \tilde{N}_D] g(n_A, \tilde{n}_A, \tilde{n}_D) \tag{EC.28}$$

where:

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

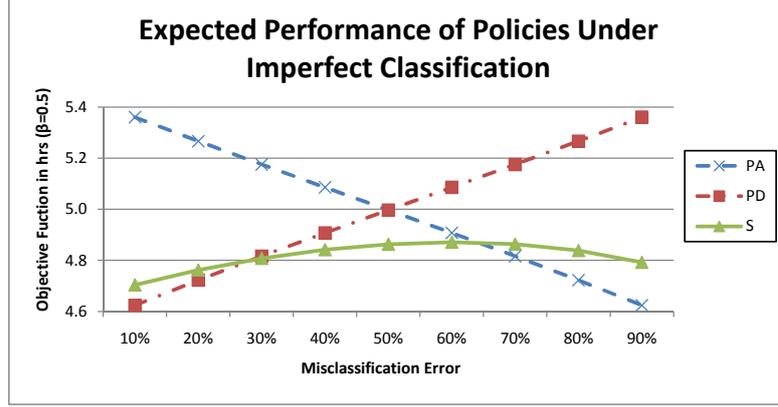


Figure EC.1 Expected performance of policies for a clearing system with $n = 20$, $\mu_A = 80(\text{mins})$, $\mu_D = 45(\text{mins})$, and symmetric misclassification error between A and D patients. Streaming is more robust to misclassification errors than pooling.

$$\begin{aligned}
& \left[\tilde{\gamma}_D \left[\sum_{j=1}^{\tilde{n}_A} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-k-1} ((k+1)\mu_D + (j-k-1)\mu_A) \right. \right. \\
& \quad + \sum_{j=\tilde{n}_A+1}^{n_A} \left[\sum_{k=0}^{\tilde{n}_A} \binom{\tilde{n}_A}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{\tilde{n}_A-k} ((k+1)\mu_D + (\tilde{n}_A - k)\mu_A) \right. \\
& \quad \quad \left. \left. + \sum_{k=0}^{j-\tilde{n}_A-1} \binom{j-\tilde{n}_A-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-\tilde{n}_A-k-1} \left((k+1)\frac{\mu_D}{2} + (j-\tilde{n}_A-k-1)\frac{\mu_A}{2} \right) \right] \right] \\
& + (1 - \tilde{\gamma}_A) \left[\sum_{j=1}^{\tilde{n}_D} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-k-1} (k\mu_A + (j-k)\mu_D) \right. \\
& \quad + \sum_{j=\tilde{n}_D+1}^{n_D} \left[\sum_{k=0}^{\tilde{n}_D} \binom{\tilde{n}_D}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{\tilde{n}_D-k} (k\mu_A + (\tilde{n}_D - k)\mu_D) \right. \\
& \quad \quad \left. \left. + \sum_{k=0}^{j-\tilde{n}_D-1} \binom{j-\tilde{n}_D-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-\tilde{n}_D-k-1} \left(k\frac{\mu_A}{2} + (j-\tilde{n}_D-k)\frac{\mu_D}{2} \right) \right] \right]. \tag{EC.29}
\end{aligned}$$

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

$$\begin{aligned}
& E[TTFT_A^{PA} | N_A = n_A] = \\
& \frac{1}{(1 - \tilde{\gamma}_D)n_A + \tilde{\gamma}_A(n - n_A)} \times \left[(1 - \tilde{\gamma}_D) \sum_{j=1}^{n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-k-1} \left(k\frac{\mu_D}{2} + (j-k-1)\frac{\mu_A}{2} \right) \right. \\
& \quad + \tilde{\gamma}_A \sum_{j=1}^{n-n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-k-1} \left(k\frac{\mu_A}{2} + (j-k-1)\frac{\mu_D}{2} \right) \right. \\
& \quad \quad \left. \left. + \sum_{k=0}^{n_A} \binom{n_A}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{n_A-k} \left(k\frac{\mu_D}{2} + (n_A - k)\frac{\mu_A}{2} \right) \right] \right].
\end{aligned}$$

Moreover, we have:

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

where $Pr(N_A = n_A)$ is given in (EC.24).

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

$$\begin{aligned}
E[\overline{LOS}_D^{PA} | N_A = n_A] &= \frac{1}{\tilde{\gamma}_D n_A + (1 - \tilde{\gamma}_A)(n - n_A)} \times \\
&\left[\tilde{\gamma}_D \sum_{j=1}^{n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{j-k-1} \left((k+1) \frac{\mu_D}{2} + (j-k-1) \frac{\mu_A}{2} \right) \right. \\
&+ (1 - \tilde{\gamma}_A) \sum_{j=1}^{n-n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_A^k (1 - \tilde{\gamma}_A)^{j-k-1} \left(k \frac{\mu_A}{2} + (j-k) \frac{\mu_D}{2} \right) \right. \\
&\quad \left. \left. + \sum_{k=0}^{n_A} \binom{n_A}{k} \tilde{\gamma}_D^k (1 - \tilde{\gamma}_D)^{n_A-k} \left(k \frac{\mu_D}{2} + (n_A - k) \frac{\mu_A}{2} \right) \right] \right], \tag{EC.30}
\end{aligned}$$

and:

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

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

$$\begin{aligned}
E[TTFT_A^{PD} | N_A = n_A] &= \frac{1}{(1 - \tilde{\gamma}_D)n_A + \tilde{\gamma}_A(n - n_A)} \times \\
&\left[\tilde{\gamma}_A \sum_{j=1}^{n-n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{j-k-1} \left(k \frac{\mu_D}{2} + (j-k-1) \frac{\mu_A}{2} \right) \right. \\
&+ (1 - \tilde{\gamma}_D) \sum_{j=1}^{n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \tilde{\gamma}_D)^k \tilde{\gamma}_D^{j-k-1} \left(k \frac{\mu_A}{2} + (j-k-1) \frac{\mu_D}{2} \right) \right. \\
&\quad \left. \left. + \sum_{k=0}^{n-n_A} \binom{n-n_A}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{n-n_A-k} \left(k \frac{\mu_D}{2} + (n-n_A-k) \frac{\mu_A}{2} \right) \right] \right],
\end{aligned}$$

and:

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

Similarly, we have:

$$\begin{aligned}
E[\overline{LOS}_D^{PD} | N_A = n_A] &= \frac{1}{\tilde{\gamma}_D n_A + (1 - \tilde{\gamma}_A)(n - n_A)} \times \\
&\left[(1 - \tilde{\gamma}_A) \sum_{j=1}^{n-n_A} \sum_{k=0}^{j-1} \binom{j-1}{k} (1 - \tilde{\gamma}_A)^k \tilde{\gamma}_A^{j-k-1} \left((k+1) \frac{\mu_D}{2} + (j-k-1) \frac{\mu_A}{2} \right) \right.
\end{aligned}$$

$$\begin{aligned}
& + \tilde{\gamma}_D \sum_{j=1}^{n_A} \left[\sum_{k=0}^{j-1} \binom{j-1}{k} \tilde{\gamma}_D^k \gamma_D^{j-k-1} \left(k \frac{\mu_A}{2} + (j-k) \frac{\mu_D}{2} \right) \right. \\
& \quad \left. + \sum_{k=0}^{n-n_A} \binom{n-n_A}{k} (1-\tilde{\gamma}_A)^k \tilde{\gamma}_A^{n-n_A-k} \left(k \frac{\mu_D}{2} + (n-n_A-k) \frac{\mu_A}{2} \right) \right],
\end{aligned}$$

and:

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

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 7. 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 7).

Service Process. The service process in the ED is depicted in Figure 6. 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

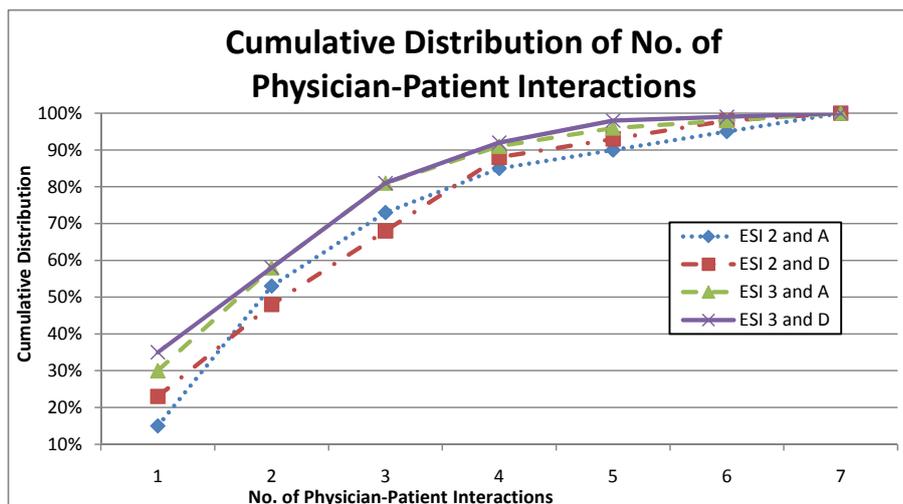


Figure EC.2 Cumulative number of class based physician-patient interactions

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

- 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.