


An Operational Framework for the Adoption and Integration of New Diagnostic Tests

Pengyi Shi* 

Krannert School of Management, Purdue University, West Lafayette, Indiana 47907, USA, shi178@purdue.edu

Jonathan E. Helm 

Kelley School of Business, Indiana University, Bloomington, Indiana 47405-1701, USA, helmj@indiana.edu

H. Sebastian Heese 

Poole College of Management, North Carolina State University, Raleigh, North Carolina 27695, USA, hsheese@ncsu.edu

Alice M. Mitchell

Indiana University School of Medicine, Indianapolis, Indiana 46202, USA, alimitch@iu.edu

The gap between medical research on diagnostic testing and clinical workflow can lead to rejection of valuable medical research in a busy clinical environment due to increased workloads, or rejection of medical research in the laboratory that may be valuable in practice due to a misunderstanding of the system-level benefits of the new test. This has implications for research organizations, diagnostic test manufacturers, and hospital managers among others. To bridge this gap, we develop a Markov decision process (MDP) from which we create “adoption regions” that specify the combination of test characteristics medical research must achieve for the test to be feasible for adoption in practice. To address the curse of dimensionality from patient risk stratification, we develop a decomposition algorithm along with structural properties that shed light on which patients and when a new diagnostic test should be used. In a case study of a partner Emergency Department, we show that the conventional myopic medical criterion can lead to poor decision making in both research development and clinical practice. In particular, we find that specificity—long a secondary consideration and often overlooked in the research process—is, in fact, the key to effective implementation of new tests into clinical environments. This myopic approach can lead to overvaluing or undervaluing new medical research. This mismatch is accentuated when a simple (current) policy is used to integrate research into the clinical environment compared with our MDP’s policy—poor implementation of a new test can also lead to unnecessary rejection. Our framework provides easily interpretable guidelines for medical research development and clinical adoption decisions that can guide medical research as to which test characteristics to focus on to improve the chances of adoption.

Key words: diagnostic test adoption; emergency department; system congestion; Markov decision process; decomposition algorithm

History: Received: August 2019; Accepted: May 2020 by Sergei Savin, after 2 revisions.

1. Introduction

An important area of medical research focuses on the development of new diagnostic tests for prevalent diseases. The development of a new test usually takes place in a clinically controlled environment and focuses primarily on the sensitivity and specificity of the test. However, the interactions with existing workflows when introducing a new test into clinical practice are largely ignored during the research phase. This can lead to resistance toward adopting new tests into already hectic hospital environments such as the Emergency Department (ED). For example, using a new test with high sensitivity but low specificity will produce many false positive results,

which leads to an excessive amount of follow-up testing, increasing patient length-of-stay, staff workload and ED congestion. In section 1.1, we discuss testing for pulmonary embolism (PE) as an example of this pitfall, which is brought to us by our clinical collaborator (an ED physician) and motivates this research. In that section, we also extend our discussion to the broader implications to the healthcare industry. This research aims to bridge the gap between medical research and workflow management by considering (i) the decision of adopting a new test and (ii) how to integrate a new test into existing clinical workflows.

To answer these two research objectives, we develop an analytical framework for evaluating the impact of introducing a new diagnostic test into a

busy hospital environment. This framework accounts for both the clinical and operational implications of the new test by capturing several key tradeoffs at the patient level (e.g., harmful tests avoided vs. misdiagnosis rate) as well as the system level (e.g., reduced workload at the original test vs. increased chance of double testing). We define two adoption criteria—medical and operational—to study the impact of test characteristics on whether a new test will be adopted in practice. Using these criteria, we define adoption regions to demonstrate how different test characteristics affect test adoption. Our framework for evaluating the adoption regions and analyzing the manner of integration of a new test is a queueing-based Markov decision process (MDP) from which we obtain optimal routing decisions regarding *which patients* should receive the new test and *when* the new test should be used. Through a case study with real data from our collaborating ED, we demonstrate how ignoring the operational aspects can lead to erroneous estimates of the value of new tests in the clinical environment outside the laboratory. More broadly, we identify how each of the characteristics of the new test and the manner of integration (use of the new test) contribute to adoption. This analysis provides general guidance for (1) the medical research process in terms of which test characteristics medical researchers should focus on to improve chances of adoption and (2) how new tests should be used in practice.

1.1. Implications for the Healthcare Industry

The adoption of new diagnostic tests, or lack thereof, has implications for numerous players in the healthcare industry. Rejection of new tests leads to wasted efforts in the medical research space in terms of researcher time and money for laboratory space, equipment, and clinical testing. Beyond the research efforts, diagnostics manufacturers have a strong vested interest in having their tests widely used and adopted. Better understanding what factors impact this adoption can help them better design their new products and demonstrate the broader benefits to potential customers. Hospital management also strives to improve both the quality of care as well as patient throughput. Identifying the right set of tests to adopt and how to integrate them into clinical workflow can help improve both of these metrics. Next, we use a concrete example of diagnosing PE in an ED to illustrate the gap between medical research on diagnostic testing and the adoption of a new test in the actual environment in which it will be used. In particular, we show that the system-level congestion effect is often ignored in medical research, but plays a critical role in the adoption decision. Gaining insights into this problem necessitates the development of a

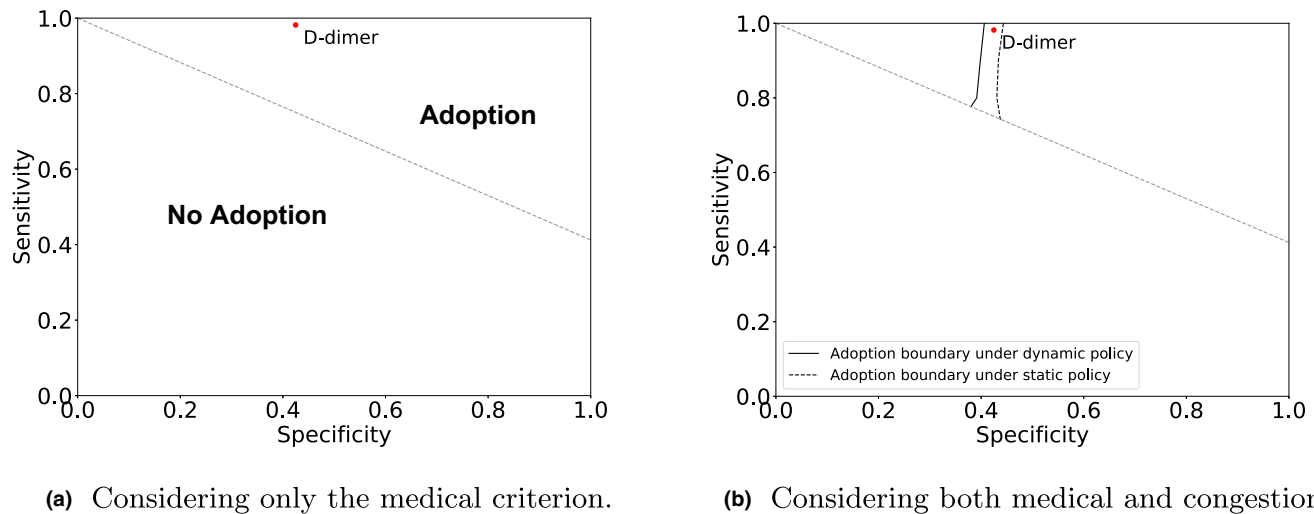
sophisticated analytical framework considering both the medical and the operational implications.

Diagnosis of pulmonary embolism (PE). When patients present to the ED exhibiting symptoms such as chest pain and shortness of breath, they are suspected of having PE. In current practice, a pretest survey asks a series of questions, which are subsequently scored. If the score, called the pretest probability, is above a predefined threshold, patients are sent to have a CT scan to confirm whether they do indeed have PE. Unfortunately, CT imaging is associated with an increased lifetime risk of cancer from ionizing radiation and, albeit rare, severe reactions that can result in kidney failure or even cardiopulmonary collapse and death. Within this context, our clinical research collaborator is interested in introducing a new test, called D-dimer, to avoid unnecessary CT scans. This test is a fast, simple blood test and has a high sensitivity. It can be used to *rule out* patients who do not have PE and thus mitigate the risk of adverse effects from the CT dye and reduce the workload burden on the CT, which is one of the ED's most heavily utilized and most expensive diagnostic resources.

The CT test is highly accurate, and, for simplicity, we assume it is a perfect test in diagnosing PE in the rest of the study. The D-dimer test, in contrast, is an imperfect test with high sensitivity but moderate specificity. Thus, a patient with a positive D-dimer test result will still be sent to the CT for confirmation. According to our medical collaborator, the moderate specificity of the D-dimer raises concerns about the adoption of this new test by ED physicians, in spite of the obvious benefits of ruling-out healthy patients and reducing unnecessary CT scans. The higher false positive rate causes many patients to be unnecessarily routed to the CT for confirmation, resulting in *double testing*. Consequently, introducing this test may increase delays to patient discharge and increase the workload and size of each doctor's panel of patients. This is a strong barrier to adoption, since physicians are reluctant to use additional testing if they view it as a workload burden. We provide an example of this adoption tradeoff below with numerical results for PE diagnosis in our partner ED.

Medical criterion: the current practice. In the medical literature, a widely accepted criterion for adopting a new diagnostic test focuses on the post-test probability, defined as the probability that a patient with a negative test result actually has the disease. For example, Kohn et al. (2017) suggests an upper bound of 3% post-test probability for PE diagnostic tests, which translates into <3% of PE cases being "missed" by the test. Figure 1a shows the test *adoption region* based on this criterion in terms of which values of sensitivity and specificity that a new test for PE must achieve to satisfy this medical criterion;

Figure 1 Criterion-based Test Adoption Decision. The dotted lines in both plots represent the boundary to achieve <3% misdiagnosis for (low-risk) PE patients. In plot (b), the solid and dashed vertical lines represent the boundaries to achieve at least 7.5% reduction in the average throughput time using the dynamic and static routing policies, respectively [Color figure can be viewed at wileyonlinelibrary.com]



parameters are estimated from our collaborating ED (details are provided in section 6). The region above the dotted line indicates clinically acceptable levels of sensitivity and specificity (i.e., adoption region), and the region below the line indicates no adoption due to a post-test probability above the required clinical threshold.

Operational criterion: accounting for congestion and health outcomes. Despite its obvious clinical value as observed in Figure 1a, the D-dimer is not universally implemented in EDs. A possible reason is the aforementioned issue of increased workloads due to double testing. We therefore consider a criterion that focuses solely on ED throughput time, which reflects patient delays and doctor workload, should be reduced after adopting the new test. Figure 1b superimposes this criterion as vertical lines on top of the medical criterion, where points to the right of these lines (adoption boundaries) indicate values of sensitivity and specificity that the new test must achieve to avoid rejection of the new test due to increased patient delays. The vertical dashed line corresponds to the adoption boundary under the current practice of routing all low-risk patients to D-dimer and routing medium- and high-risk patients directly to the CT. The D-dimer dot indicates that there may be resistance to adoption due to increased doctor workloads, which may explain why some hospitals are hesitant to adopt the D-dimer. The solid line shows the adoption boundary under the *optimal routing* based on the MDP developed in this study. Under the optimal routing, the D-dimer *would* meet the requirement of throughput time reduction for adoption. This highlights the importance not only of characteristics of the test, but also of the manner of integration into the

clinical workflow for the adoption decision. The effect of routing is further exacerbated if the new test were slower as we will discuss later in this study.

This example shows that, without a sophisticated analytical framework that considers both the medical and congestion effects, medical researchers may think their new test is sufficiently accurate (satisfying the accepted medical criterion) only to find physicians reluctant to use it in practice due to the associated workload burden. Meanwhile, physicians may reject some tests that could be valuable in practice due to poor integration of the test into the hospital workflow. In this study, we present an operational framework that takes a holistic perspective to account for both the medical and congestion effects. Within this framework, we define an *operational criterion* that unifies the ED throughput time and congestion concerns with multiple health outcomes (misdiagnosis and unnecessary CT).

1.2. Contributions and Overview of the Paper

In this research, we make the following technical and practical contributions.

Analytical framework. We develop an analytical framework that captures critical tradeoffs not previously considered in the adoption of medical research in practice, which can lead to poor decision making in both research development and clinical practice. Our framework specifically accounts for tradeoffs at: (1) the patient level: misdiagnosis vs. adverse effects of the current test; and (2) the system level: longer throughput time caused by overcrowding at the current test vs. double testing caused by introducing the new test. Using this framework, we develop novel, easily interpretable adoption region plots that demarcate the boundaries for adoption based on test characteristics,

which can be used to guide both research development and clinical adoption decisions.

MDP and algorithm. We formulate a queueing-based MDP that captures these tradeoffs through different costs. Importantly, we show that the different adoption criteria are unified via the MDP cost objective function, allowing us to compute adoption regions and to capture different methods of integrating the new test into the existing clinical environment.

Our model captures important features in the ED patient flow, incorporating patient heterogeneity into a queueing model that covers a number of commonly used queues to reflect different types of testing environments. Including these features leads to the curse of dimensionality in the MDP. Thus, we develop a decomposition algorithm based on a multi-step policy improvement approach that leverages a class of static routing policies to significantly reduce the state space size. Via extensive numerical studies, we show this algorithm is near optimal; among all the numerical experiments, the largest optimality gap is <1.50%. In addition, we derive useful structural properties that lead to managerial insights about the dominance among different patient classes: if the negative (positive) result is more accurate, it is preferable to use the test on low-risk (high-risk) patients to rule out (confirm) the disease.

Test adoption and integration. The purpose of the MDP is to provide an analytical framework for evaluating the impact of introducing a new test into a clinical environment. The main research output from this analytical framework is the test adoption region under the various adoption criteria as illustrated in Figure 1. This model-based evaluation mechanism is critical because a main finding of this study is that a test should not be evaluated in an isolated laboratory environment, but instead should be considered in the actual *system* context in which it will be used. This requires an analytical model capturing patient flow in the full system. Via a case study of diagnosing PE in EDs using data from our partner hospital, we show that the traditional way of measuring test value in the isolated clinical laboratory can in fact overlook both congestion effects and health outcomes.

- *Operational performance:* Focusing solely on the medical criterion can either overvalue (introducing a test that may not be adopted) or undervalue (rejecting a test that should be adopted) a new test. The medical criterion overvalues a test with high sensitivity and low specificity (common in practice) by ignoring the operational inefficiency of double-testing, which can lead to lack of adoption in practice. With high specificity and low sensitivity, the medical criterion may undervalue a new test, causing the clinical

community to reject a test which may have significant operational benefits that could outweigh a slightly higher misdiagnosis rate.

- *Medical effects:* Another failure of the medical criterion is that it only considers the individual misdiagnosis rate while overlooking the benefits of avoiding the current test (e.g., CT), which may have serious health implications (e.g., exposure to radiation and adverse health events). Furthermore, the current definition of misdiagnosis rate ignores the effect of sequential testing, which can lead to unnecessary rejection of valuable research.
- *Test integration:* We show that the manner in which the new test is integrated into the clinical workflow can have a significant effect on the test's value and feasibility for adoption. When the new test is slow or the specificity is low, the test may be rejected under the current (static) practice, while the dynamic MDP policy can take advantage of fluctuations in ED workload to extract medical and operational benefits that would otherwise be lost due to poor implementation.
- *Driver of adoption and integration:* Surprisingly, we find that specificity, which has long been a secondary consideration in medical research due to a weaker impact on the current medical measure of a test's value, is in fact the key to unifying the medical and operational value of new tests. Specificity has a major impact on how effectively a new test can be integrated into an existing clinical environment, which cannot be measured in an isolated laboratory environment. Hence, we find that specificity has significant implications for adoption into practice that are currently overlooked.

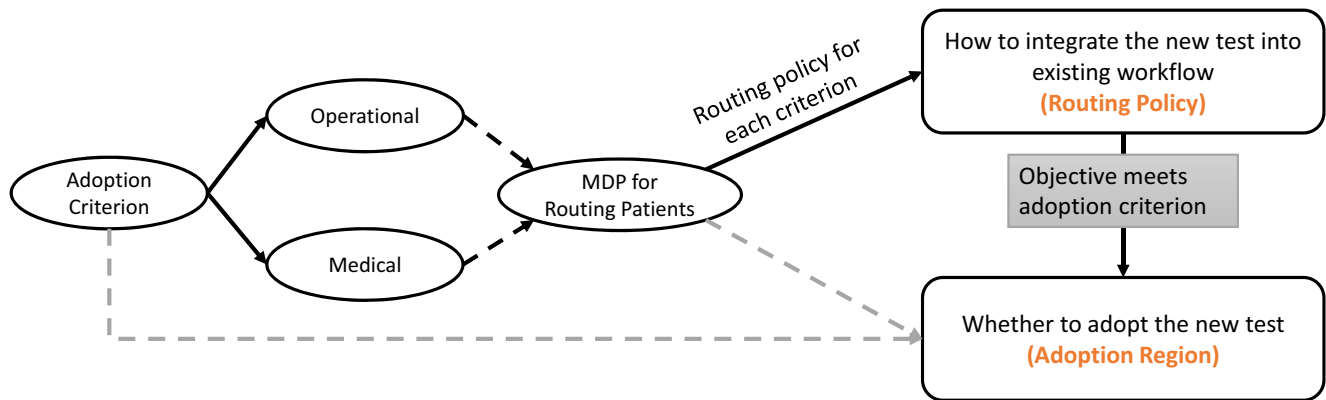
Figure 2 summarizes the main components of the research framework developed in this study.

2. Literature Review

We give a brief review of the related literature.

Dynamic control of queueing systems. Our modeling and routing decision connects with the literature on dynamic control of queueing systems (e.g., Stidham 1985, Stidham Jr and Weber 1993). There is a rich body of research on admission control via MDP techniques, for example, Miller (1969), Yoon and Lewis (2004), Zhang and Ayhan (2013), to name a few. Koole (2007) provides a comprehensive survey and the event-based operator technique for proving structural properties of MDPs for control of queueing systems. The work most relevant to ours is that of Hajek (1984), who studies the dynamic routing decision in a two-

Figure 2 Overview of the Paper [Color figure can be viewed at wileyonlinelibrary.com]



station queueing network, that is, which station should an arriving customer be sent to. Hajek (1984) shows that the structure of the optimal routing decision is of threshold form. A key difference to our model is that we consider a general queueing setup with a state-dependent service-rate function for each station, which is more general than their single-server setting with constant service rate. In general, it is challenging to characterize the optimal routing policy in settings beyond the single-server one, so we focus on establishing that there exists a dominance among different classes in the optimal policy. In addition, we also consider routing based on the outcome of the test in the first station which Hajek (1984) did not study. MDP is also widely used for optimizing patient screening/triage and routing in resource-limited healthcare environments; for example, Argon and Ziya (2009), Örmeci et al. (2015), Sun et al. (2017), Hu et al. (2019). Also see Chapter 14 in Dai and Tayur (2018) for a detailed review of MDP applications in healthcare operations.

Diagnostic decision making in OM. Wang et al. (2010) study the staffing and service depth decisions of a nurse triage line in which the patients decide whether to call the line or not based on their expectations of diagnostic accuracy and congestion. They focus on the equilibrium analysis between the provider and patients. Alizamir et al. (2013) study the tradeoff between test accuracy and system congestion in a diagnostic process, in which the service provider conducts multiple tests to determine the customer's type. They consider the dynamic decision making of whether to run more tests or to stop the process and identify the customer's type. Our paper focuses on the routing between multiple tests while the sequence of tests to conduct is predetermined in their paper. More broadly speaking, our paper relates to the well-known speed-quality tradeoff in service operations, that is, the service quality (diagnostic accuracy) differs with the speed of the service (time to finish a test).

Shumsky and Pinker (2003) model a gatekeeper who makes an initial diagnosis of the customer's problem and decides whether to refer the customer to a specialist. They focus on the information asymmetry between the gatekeeper and the specialist and use a principal-agent framework. Hasija et al. (2005) extend their model to include queueing at both the gatekeeper and the expert. They solve for optimal staffing levels and referral rates between gatekeeper and expert from a centralized perspective.

Diagnostic decision making in clinical literature. Diagnostic decision making has been studied extensively in the medical literature (e.g., Barrows and Pickell 1991, Bordage 1994, Guyatt et al. 2002, Kassirer et al. 1991). Guyatt et al. (1986) provide a general framework for clinical evaluation of diagnostic technologies. Most of these studies are developed in a clinically controlled environment, without considering the interactions with existing workflows. In this study, we explicitly consider the operational effects relevant to diagnostic decision making. Regarding PE, the motivating case brought to us from our partner ED, Fesmire et al. (2011) provide the clinical policy from the American College of Emergency Physicians on the diagnosis of patients presenting with suspected PE. In particular, they discuss the role of the D-dimer test in the exclusion of PE. Given the fact that the sensitivity of the D-dimer test is very high and its specificity is moderate, the D-dimer test is suggested to be used on low-risk patients (see Fesmire et al. 2011, Qaseem et al. 2007, Segal et al. 2007).

3. Analytical Framework for Evaluating Test Adoption and Integration

In this section, we introduce the analytical framework for answering our research questions. In section 3.1, we present the general queueing setup for modeling

the patient flow with one perfect, existing test (e.g., CT scan), and an imperfect, potential new test. In section 3.2, we introduce the adoption criteria and define the test adoption regions to determine when a test would be adopted. In section 3.3, we present an MDP framework to derive the optimal routing policy under a cost-based objective that unifies these adoption criteria, and we address how a new test should be integrated into clinical workflows.

3.1. Patient Flow Model

We consider J classes of patients, capturing different levels of patient risk of having a suspected disease (e.g., PE). Without loss of generality, we assume that the pretest probability of having the disease is lowest for class 1 patients, and highest for class J patients: $p_0^1 < p_0^2 < \dots < p_0^J$, where p_0^j is the pretest probability for class j patients. We also assume that the lowest risk class passes the minimum threshold for undergoing testing, such that all patients we consider here need to receive at least one test.

Figure 3 shows the patient flow when there are two tests available—denoted as two stations in our model. The first station represents an “imperfect test,” and the second station represents a “perfect test.” Here, an imperfect test (e.g., the D-dimer test) means the test may give a false positive or a false negative result, while a perfect test (e.g., the CT scan) does not. Each patient may go through station 1 only, station 2 only, or station 1 and then station 2, before leaving the system. Note that leaving the system (departure) does not necessarily mean that the patient is discharged home; it could be sending the patient for treatment if the patient is diagnosed to have the suspected disease. We assume the test at the second station is perfect for two reasons. First, it allows us to gain insights

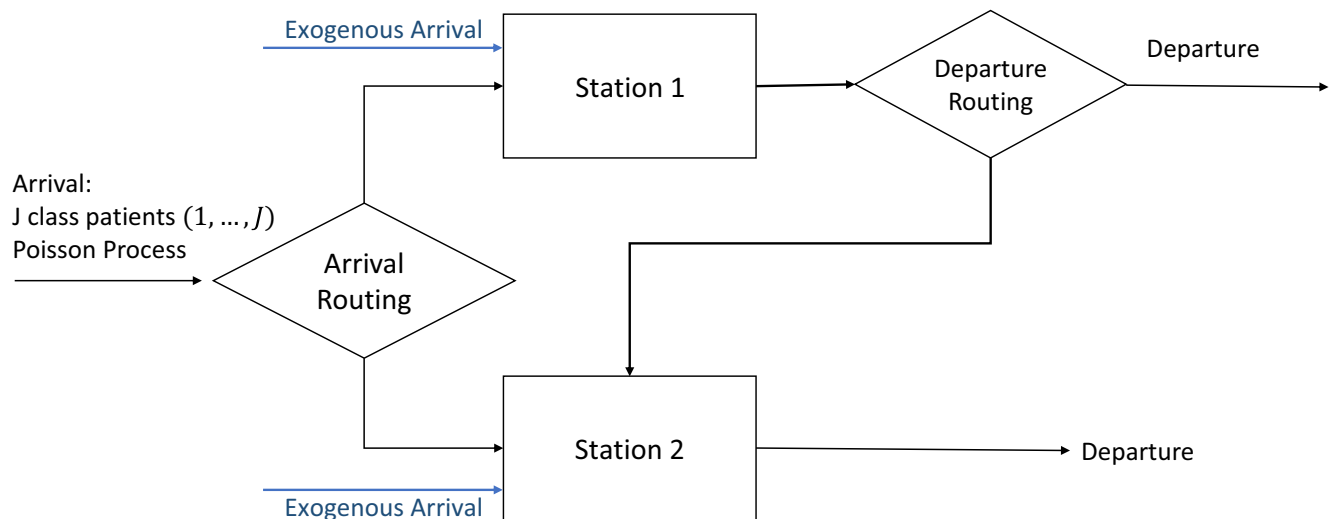
under the largest differentiation between the two types of tests. Second, according to our clinical collaborator, assuming a perfect follow-up test is a good approximation of reality. Making the test “close to perfect” would complicate the analysis, but add little additional insight.

We model the patient flow through the two tests as a queueing network with two stations. We adopt a queueing model for each station that covers a number of commonly used queueing structures, which can reasonably capture different types of testing environments. For example, our framework encompasses $M/M/k$ for the case where there are multiple diagnostic machines at a station, $M/M/1$ for the case where there is only one diagnostic machine, and processor-sharing queues for more complex settings such as batch testing, preemptive STAT (urgent priority) testing, and time-varying availability of testing equipment, all of which can occur in ED settings. Next, we detail the arrival and service processes for our queueing framework.

Arrival process. Arrivals for each of the J classes form independent Poisson processes. We assume the arrival rate of class j is λ^j , $j = 1, \dots, J$. The modeling framework can be extended to incorporate time-varying arrival rates using the pointwise stationary approximation; see details in section 7.2. For ease of exposition, we focus on the time-homogeneous case when introducing the framework. Let $\lambda = \sum_j \lambda^j$ denote the total arrival rate of the J classes. In addition to these patients with the suspected disease, we consider patients who may use each station for other reasons. The exogenous arrivals to station i also follow a Poisson process with rate λ_i^{ex} , $i = 1, 2$.

Service process. A patient starts receiving service once being admitted to the service station (upon

Figure 3 Basic Patient Flow [Color figure can be viewed at wileyonlinelibrary.com]



arrival or after waiting). The service process has the following properties. The service completion process during the busy period follows a time-nonhomogeneous Poisson process with a general, state-dependent service rate function $\mu_i(s_i)$, where s_i is the total number of patients *in service* at station i . When a service completion event occurs, a patient in service is randomly chosen to depart, that is, the service rate of each patient in service is $\mu_i(s_i)/s_i$. The service rate function satisfies the following: the total service rate $\mu_i(s_i)$ is non-decreasing in s_i ; an individual patient's service rate $\mu_i(s_i)/s_i$ is non-increasing in s_i ; and the total service rate $\mu_i(s_i)$ is bounded by $\bar{\mu}_i$. This service setting covers commonly used queues for modeling service stations in healthcare research, for example,

- **M/M/1 queue**, where $\mu_i(s_i) = \mu$ if $s_i = 1$ and 0 if $s_i = 0$ (μ is the service rate of the single server),
- **M/M/k queue**, where $\mu_i(s_i) = \mu \cdot s_i$ ($0 \leq s_i \leq k$), μ is the service rate of each of the k servers, and $\mu_i(s_i)$ is thus bounded by $k\mu$,
- **Processor-sharing queue**, where each patient is admitted to service upon arrival and gets a uniform share of the total service capacity, that is, each patient's service rate is $\mu_i(s_i)/s_i$.

Service discipline. In scenarios where patients may need to wait, we adopt the random-processing service discipline when choosing a waiting patient to admit upon a service completion. In multi-class queues, it is shown that the random-processing service discipline leads to the same steady-state distribution as the first-in first-out (FIFO) discipline under Poisson arrivals and the service process setting described above (Buitenhok et al. 1997). This leads to the same performance measures such as the mean number of patients in service or in the queue. Since these are the main operational performance measures used in our decision model, the two disciplines are equivalent, particularly from the system perspective since we focus on system-level operational performance measures as opposed to tracking each individual patient. We adopt the random-processing service discipline since it allows a more parsimonious state space (we do not need to track the order of each patient in the queue).

We present our decision framework and solution algorithm using the general queueing setting described above for the rest of this section and in sections 4 and 5. In the case study in section 6, however, we focus on the processor-sharing queue setting and parameterize the model accordingly. The processor-sharing queue can be viewed as an infinite-server queue with state-dependent service rate function $\mu_i(s_i)/s_i$ for each patient, which has been shown to

have the flexibility to model complicated healthcare systems while approximating the actual system performance reasonably well (Armony et al. 2015, Whitt and Zhang 2017). We choose this approach for our case study because it allows us to closely replicate the empirical distribution of test occupancy levels, which cannot be achieved with the M/M/1 queue (e.g., see Figure 4 later in section 6.1).

3.2. Adoption Criteria and Adoption Regions

Diagnostic tests possess a set of characteristics that affect their medical and operational value. We define these characteristics using the tuple

$$C := (b^+, b^-, \mu_1(x)), \quad (1)$$

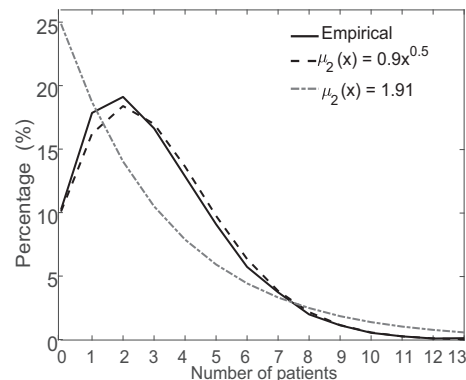
where b^+ and b^- , respectively, are the sensitivity and specificity of the new test that could potentially be class-dependent, and $\mu_1(x)$ is the service rate of the new test if there are x patients at the test. We reserve the notation $\mu_2(x)$ for the existing test (at station 2). Let $\mathcal{U} = \{C : b^+, b^- \in [0, 1]; \mu_1(x) \in [0, \bar{\mu}_1] \forall x\}$ be the set of all feasible test characteristics.

We define the adoption criterion $f^\pi : \mathcal{U} \rightarrow [0, 1]$ as a function that maps test characteristics to a quantitative measure of efficacy between 0 and 1 under a routing policy π . We say that a test with characteristics C is adopted at a level ε if $f^\pi(C) \geq \varepsilon$. That is, ε is the minimum level (lower bound) of efficacy that must be achieved for the test to be adopted under criterion f^π . We now formally define what we call an *adoption region*, which determines the set of all test characteristics that satisfy the criterion.

DEFINITION 1. (ADOPTION REGION).

$$\mathcal{A} = \{C \in \mathcal{U} : f^\pi(C) \geq \varepsilon\} \quad (2)$$

Figure 4 Histogram of Patient Occupancy in CT. The x-axis numerates the possible states, that is, the number of patients waiting or receiving CT, and the y-axis corresponds to the frequency of the state



Next, we define three different criteria to specify the functional form of $f^\pi(C)$.

In the medical literature, the criterion for an acceptable diagnostic test follows the “rule-out” approach: the post-test probability of being sick, given a negative test result, should be $< \alpha$. That is, a doctor should be confident with at least $1 - \alpha$ chance that any patient who is routed to the test and receives a negative test result does not have the disease. Mathematically, this criterion is defined as follows.

ADOPTION CRITERION 1. (MEDICAL CRITERION). For a given $\alpha \in [0,1]$, the medical criterion is defined as

$$f_M^\pi(C) := 1 - \frac{p_0(1 - b^+)}{p_0(1 - b^+) + (1 - p_0)b^-} \geq 1 - \alpha. \quad (3)$$

We use $1 - \alpha$ to be consistent with the general form in Equation (2). Here, p_0 is the highest pretest probability among all patients that policy π chooses to route to the new (imperfect) test. For example, for the D-dimer test, we consider in the case study, the medical practice suggests it should be used on low-risk patients, thus, $p_0 = p_0^1$. Criterion (1) is equivalent to

$$\frac{1 - b^+}{b^-} < \frac{\alpha(1 - p_0)}{(1 - \alpha)p_0}, \quad (4)$$

where $(1 - b^+)/b^-$ is called the negative likelihood ratio in the medical literature; see Penalzoza et al. (2012) for an example. Kohn et al. (2017) suggest 3% ($\alpha = 0.03$) as an appropriate threshold for the medical efficacy of the D-dimer test in our case study. For a given α , the boundary of the adoption region can be simply obtained by solving (3), which has the solution

$$b^+ + \frac{\alpha(1 - p_0)}{(1 - \alpha)p_0} b^- = 1. \quad (5)$$

From (5), we see that sensitivity is a more powerful lever than specificity in satisfying the medical criterion. This may explain why the current medical research often focuses on improving test sensitivity. For example, in discussing the efficacy of artemisinin-based combination treatments such as Davis et al. (2005), sensitivity is the major consideration. However, if one considers the doctor workload and system congestion, specificity becomes important to avoid excessive double testing from false positives. To capture this effect, we next introduce the congestion criterion.

Analogous to the medical criterion, the congestion criterion only considers one aspect of the testing

system: ED throughput time. This *congestion criterion* demarcates the largest region in which the test would be considered acceptable solely from reducing doctor workload and patient delays.

ADOPTION CRITERION 2. (CONGESTION CRITERION). Let $\mathbb{E}[T_0]$ and $\mathbb{E}[T^\pi(C)]$, respectively, denote the average throughput time for all ED patients before and after adopting a test with characteristics C , using routing policy π . For a given $\beta \in [0,1]$, the congestion criterion is defined as

$$f_W^\pi(C) := \frac{\mathbb{E}[T_0] - \mathbb{E}[T^\pi(C)]}{\mathbb{E}[T_0]} \geq \beta. \quad (6)$$

This criterion requires that the average patient throughput time is reduced (relatively) by at least β after adopting the new test. By Little’s Law, throughput time linearly maps to physician workload. The average throughput time depends on how the new test is integrated into the existing workflow, that is, the routing policy π .

Lastly, we define the *holistic operational criterion*, which considers both congestion and the health outcomes such as misdiagnosis and health costs of unnecessary CT testing. This criterion optimizes the long-run average cost from the MDP framework to be introduced in section 3.3.

ADOPTION CRITERION 3. (OPERATIONAL CRITERION). Let V_{CT} and $V_{\text{adopt}}^\pi(C)$, respectively, denote the long-run average cost before and after adopting a test with characteristics C , using routing policy π . For a given $\gamma \in [0,1]$, the cost-based operational criterion is defined as

$$f^\pi(C) := \frac{V_{CT} - V_{\text{adopt}}^\pi(C)}{V_{CT}} \geq \gamma. \quad (7)$$

This criterion requires that the long-run average cost be reduced (relatively) by more than γ after adopting the new test. As we will formally define in section 3.3, the cost captures misdiagnosis rate, health costs of unnecessary testing, and patient delays. Next, we introduce the MDP formulation to compute the routing policy.

3.3. MDP Formulation for Test Routing Policy

Assume two tests are available for clinicians to use, as shown in Figure 3. We formulate the test routing decision problem as an infinite-horizon, long-run average cost, continuous-time MDP. We introduce the state space, action space, cost structure, and objective function below.

3.3.1. State Space. The system state is captured with a $2(J + 3)$ -dimensional vector

$$\underline{x} = (x_1^0, x_1^1, \dots, x_1^J, x_2^0, x_2^1, w_1^0, w_1^1, \dots, w_1^J, w_2^0, w_2^1).$$

- x_1^0 and x_2^0 denote the number of patients from exogenous arrivals that are receiving service at stations 1 and 2, respectively.
- x_1^j denotes the number of class j patients with the suspected disease that are receiving service in station 1, $j = 1, \dots, J$.
- x_2^j denotes the total number of patients with the suspected disease (i.e., excluding external arrivals) that are receiving service in station 2. We do not need to differentiate patient classes in station 2 since it is a perfect test. See further explanation in section 3.3.3 where we introduce the cost structure.
- w_1^j denotes the number of class j patients ($j = 0$ for exogenous) who are currently waiting to receive service from station i .
- w_2^0 and w_2^1 denote the number of exogenous and suspected patients that are waiting to receive service from station 2. Similarly to the total count, we do not differentiate by patient classes.

For notational convenience, we define

$$\underline{x}_1 = (x_1^0, x_1^1, \dots, x_1^J, w_1^0, w_1^1, \dots, w_1^J), \quad (8)$$

$$\underline{x}_2 = (x_2^0, x_2^1, w_2^0, w_2^1), \quad (9)$$

$$s_1 = \sum_{j=0}^J x_1^j, \quad s_2 = x_2^0 + x_2^1, \quad (10)$$

$$n_i^j = x_i^j + w_i^j, \quad n_1 = \sum_{j=0}^J n_1^j, \quad n_2 = n_2^0 + n_2^1, \quad (11)$$

Here, \underline{x}_i is the state vector that records patients in service or waiting from each class in station i , s_i counts all patients who are in service at station i , and n_i is the total patient count in station i , including those who are in service and are waiting. We can recover the queue-length count as $\sum_{j=0}^J w_1^j = n_1 - s_1$ for station 1 and similarly for station 2.

3.3.2. Action for Routing Policy. An action is triggered upon a new patient arrival or a patient departure from station 1 (the imperfect test). At an arrival event, we make an *arrival routing* decision, that

depends on which class j the arriving patient belongs to. The arrival action is given by

$$a_{\text{arr}}^i(\underline{x}) = i \quad \text{if the arriving patient is routed to station } i, \quad i = 1, 2.$$

Recall that $i = 1$ corresponds to the new, imperfect test (e.g., D-dimer) while $i = 2$ corresponds to the existing, perfect test (e.g., CT). No routing decisions are made for exogenous arrivals.

When a class j patient departs from station 1, we make a *departure routing* decision of whether or not to send the patient for additional testing at station 2, which depends on the test result at station 1. We denote the decision as $a_{\text{dep}}^{i,+}(\underline{x})$ or $a_{\text{dep}}^{i,-}(\underline{x})$ for a patient receiving a positive or a negative test result at station 1, respectively. For $r = +, -$,

$$a_{\text{dep}}^{i,r}(\underline{x}) = \begin{cases} 1 & \text{if the patient is directly discharged from the system,} \\ 2 & \text{if the patient is routed to station 2.} \end{cases}$$

Exogenous patients at station 1 and all patients at station 2 directly depart from the system (for follow-up treatment, testing for other diseases, discharged home, etc.).

3.3.3. Cost Structure. We consider two types of costs: individual-level diagnostic cost and system-level congestion cost.

Diagnostic cost. Diagnostic cost is measured by the accuracy of the diagnostic result. We define $c_i^{\text{FN}}, c_i^{\text{TN}}, c_i^{\text{FP}}$, and c_i^{TP} as the cost associated with a false negative, true negative, false positive, and true positive result, if the patient directly departs from the system after getting the test result from station i ($i = 1, 2$). To calculate the expected diagnostic costs, we specify the post-test probability as follows.

The test result in station i is positive with probability $b_i^{j,+}$ given a class j patient has the suspected disease (i.e., true positive rate), and is negative with probability $b_i^{j,-}$ given the patient does not have the suspected disease (i.e., true negative rate). Mathematically, sensitivity and specificity are independent of the patient's pretest probability, since sensitivity and specificity are conditional on the patient having or not having the disease. However, for each of the risk class we consider, beyond the pretest probability, there may be unobservable patient characteristics in each class that could affect the sensitivity and specificity of the test in that subpopulation. The clinical literature (Leeflang et al. 2013) indicates that sensitivity and specificity may or may not depend on disease

prevalence, likely due to these unobservable factors. This is why, in the MDP framework, we choose to present the more general formulation where $b^{+,-}$ depends on patient class, which can account for this spectrum bias. However, it typically requires large variations in prevalence and severity in the target population to see measurable impact on sensitivity or specificity (Linden 2006, Mulherin and Miller 2002). According to our clinical collaborator, for PE and many diseases that are diagnosed in the ED, the variance in the population that is actually tested is low such that individual patient characteristics would not significantly impact sensitivity or specificity, hence, we assume the test sensitivity and specificity are class-independent in our case study.

If the test at station i is positive, the post-test probability for a patient having the disease is

$$\tilde{p}_i^{j,+} = \frac{p_0^j b_i^{j,+}}{p_0^j b_i^{j,+} + (1 - p_0^j)(1 - b_i^{j,-})}, \quad i = 1, 2; j = 1, \dots, J. \quad (12)$$

If the test result is negative,

$$\tilde{p}_i^{j,-} = \frac{p_0^j(1 - b_i^{j,+})}{p_0^j(1 - b_i^{j,+}) + (1 - p_0^j)b_i^{j,-}}, \quad i = 1, 2; j = 1, \dots, J. \quad (13)$$

Thus, the expected diagnostic cost for a class j patient getting a test at station i (and directly departing)

$$c_i^{j,+} = \tilde{p}_i^{j,+} c_i^{\text{TP}} + (1 - \tilde{p}_i^{j,+}) c_i^{\text{FP}}, \quad i = 1, 2, \quad (14)$$

$$c_i^{j,-} = \tilde{p}_i^{j,-} c_i^{\text{FN}} + (1 - \tilde{p}_i^{j,-}) c_i^{\text{TN}}, \quad i = 1, 2. \quad (15)$$

Since we assume the test at station $i = 2$ is perfect, we have $b_2^{j,+} = b_2^{j,-} = 1$ and $\tilde{p}_2^{j,+} = 1$ ($\tilde{p}_2^{j,-} = 0$) if the test result is positive (negative). Thus, the expected diagnostic cost after a test at station 2 can be simplified as $c_2^{j,+} = c_2^{\text{TP}}$ and $c_2^{j,-} = c_2^{\text{TN}}$. Without loss of generality, we also assume that $c_2^{j,+} = c_2^{j,-} = c_2$, because the test at station 2 is perfect and will diagnose the patient correctly. This assumption simplifies the state space since it is only necessary to account for the total number of patients in station 2, rather than differentiating by class. Moreover, for the imperfect test at station 1, consistent with the medical literature and unless specified otherwise, we assume $c_1^{\text{FN}} \geq c_1^{\text{TN}}$ and $c_1^{\text{FP}} \geq c_1^{\text{TP}}$ in the remainder of the study. That is, the cost associate with a false-negative (false-positive) is

higher than that associated with a true-negative (true-positive).

Congestion cost. To capture the workload impact of routing patients into each station, we consider the unit-time holding cost $h_i(\underline{x}_i)$. We assume $h_i(\underline{x}_i)$ is increasing in each coordinate of \underline{x}_i . In the rest of the study, unless specified otherwise, we use the linear holding cost $h_i(\underline{x}_i) = h \cdot n_i$ in terms of the total patient count n_i . We refer to holding cost and congestion cost interchangeably. We use

$$h(\underline{x}) = h_1(\underline{x}) + h_2(\underline{x})$$

to denote the total congestion cost from both stations.

Long-run average cost formulation. We formulate an infinite-horizon, long-run average cost MDP. For policy π , we define the average cost V^π as

$$V^\pi = \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T h(\underline{X}^\pi(t)) dt + \sum_{j=1}^J \left(c_1^{j,+} Y_1^{j,+}(T) + c_1^{j,-} Y_1^{j,-}(T) \right) + c_2 Y_2(T) \right],$$

where $\underline{X}^\pi(t)$ denotes the (stochastic) state under policy π at time t . Within the expectation, the first integral captures the cumulative holding costs at both stations, while the other terms capture the cumulative diagnostic costs: $Y_1^{j,+}(t)$ and $Y_1^{j,-}(t)$ are counting processes to track the total number of class j patients exiting the system after getting a positive or negative result, respectively, from station 1 up to time t , under policy π ; $Y_2(t)$ is the counting process to track the total number patients exiting the system after finishing the test at station 2 up to time t , under policy π .

Our objective is to minimize the long-run average cost, defined as

$$\omega^* = \inf_{\pi} V^\pi.$$

3.3.4. Routing Policy for Computing Adoption Regions. In this section, we connect each adoption criterion from section 3.2 with the cost structure in the MDP, from which we can compute the corresponding adoption regions (whether to adopt a test or not). In other words, we show that the medical and congestion criteria can be unified under the cost-based objective of our MDP.

Adoption under the Medical Criterion: The following proposition provides a one-to-one mapping between the routing in the MDP and the medical criterion.

PROPOSITION 1. Let $h = 0$ and assume patients with positive results from station 1 will be sent to station 2 for follow-up check. Assume that

$$\frac{c_2 - c_1^{TN}}{c_1^{FN} - c_1^{TN}} = \alpha. \quad (16)$$

Then the new test will be adopted (i.e., a patient would be routed to the new test in the MDP) if and only if the medical criterion, $f_M^\pi(C) \geq 1 - \alpha$, is satisfied.

The proof of Proposition 1 is in Appendix B.1. We call the set of test characteristics $\mathcal{C}_M = \{C : f_M^\pi(C) \geq 1 - \alpha\}$ the medical adoption region.

Adoption under the Congestion Criterion: Setting all $c_1^{FN}, c_1^{TN}, c_1^{FP}, c_1^{TP}, c_2$ to be 0 and $h > 0$, then solving policy π from the MDP for a test with characteristics C , the congestion criterion is met with $f_W^\pi(C) \geq \beta$, if and only if $f^\pi(C) \geq \beta$. This follows from Little's Law, as the relative reduction in throughput time is equal to the relative reduction in queue length, which is captured by the holding cost. The congestion adoption region, which is the largest set of test characteristics that would be considered acceptable solely from a workload and patient delay perspective, is the set $\mathcal{C}_W = \{C : f_W^\pi(C) \geq \beta\}$.

Adoption under the Operational Criterion: The operational criterion models a combination of both the congestion effects and health outcomes. We call the set of test characteristics $\mathcal{C}_O = \{C : f^\pi(C) \geq \gamma\}$ the operational adoption region, where policy π is the optimal routing policy solved from the MDP with non-zero holding and diagnostic costs.

REMARK 1. The medical effects considered under the operational criterion are different from those under the medical criterion. First, the operational criterion considers the misdiagnosis rate of the entire testing system that has the CT station for follow-ups. The medical criterion only considers the misdiagnosis rate of the individual test, which can overestimate the true misdiagnosis rate. That is, when a new test is adopted in a clinical environment, it is not implemented in a vacuum (as the medical criterion often assumes), but in a system context. The system-level misdiagnosis rate provides a more accurate representation of the misdiagnosis rate of a new test than the medical criterion. Second, the health outcomes considered under the operational criterion are broader than those considered under the myopic medical criterion. The operational criterion considers the impact of using the new test on reducing the unnecessary usage of the potentially harmful and heavily utilized CT. Note that the

health outcome measures captured in the operational criterion depend on the usage of the new test in the system setting and are operational in natural. Thus, we subsume health outcomes and throughput within the single, all-encompassing term “operational criterion,” which mirrors the defect detecting problems studied in the manufacturing systems.

4. Bellman Equation and Structural Properties

In this section, we specify the Bellman equation for the MDP introduced in section 3.3. We then derive structural properties that provide insights into the dominance among patient classes.

4.1. Bellman Equation

We first need to uniformize the continuous time MDP. Recall that $\lambda = \sum_j \lambda^j$ is the total arrival rate of the disease suspects (λ^j is the arrival from class j disease suspects), λ_i^{ex} is the exogenous arrival rate to station i , and $\bar{\mu}_i$ is the upper bound for the service rate function $\mu_i(s_i)$ for $i = 1, 2$. Without loss of generality, we assume $\bar{\mu}_1 + \bar{\mu}_2 + \lambda + \lambda_1^{\text{ex}} + \lambda_2^{\text{ex}} = 1$, such that the uniformization rate is 1. Let $v(\underline{x})$ denote the relative value function for state \underline{x} . The optimal Bellman equation then can be written using event-based dynamic programming (Koole 2007):

$$\begin{aligned} \omega^* + v(\underline{x}) &= h_1(\underline{x}_1) + h_2(\underline{x}_2) + \lambda_1^{\text{ex}} T_{A_1^0} v(\underline{x}) \\ &+ \lambda_2^{\text{ex}} T_{A_2^0} v(\underline{x}) + \mu_2(s_2) T_{D_2} v(\underline{x}) \\ &+ \sum_{j=1}^J \lambda^j T_{AR^j} v(\underline{x}) + \mu_1(s_1) T_{D_1} v(\underline{x}) \\ &+ (\bar{\mu}_1 - \mu_1(s_1)) v(\underline{x}) + (\bar{\mu}_2 - \mu_2(s_2)) v(\underline{x}), \end{aligned} \quad (17)$$

where the T_{AR^j} are the arrival routing operators for class j patients with the suspected disease used to determine the arrival routing decision d_{arr}^j given previously, the $T_{A_i^0}$ are the arrival operators for the exogenous patients to station i , and the T_{D_i} are the departure operators from station i used to determine the departure routing decision d_{dep}^i given previously. We specify each operator below.

We focus on scenarios when $0 < s_i \leq k_i$ (k_i is the capacity of station i), delegating the complete specification with the boundary conditions (i.e., $s_i = 0$ or the queue is not empty) to the appendix. For a function $f : \mathbb{Z}_+^{J+3} \rightarrow \mathbb{R}$,

$$\begin{aligned} &[\text{Arrival routing operators}] \\ T_{AR^j} f(\underline{x}) &= \min\{f(\underline{x} + e_1^j), f(\underline{x} + e_2^j)\}; \end{aligned} \quad (18)$$

[Departure routing operators]

$$T_{D_1}f(\underline{x}) = \sum_{j=1}^J x_1^j/s_1 \left(l_1^{j+} T_{DR_1^{j+}}f(\underline{x}) + l_1^{j-} T_{DR_1^{j-}}f(\underline{x}) \right) + x_1^0/s_1 T_{D_1^0}f(\underline{x}), \quad (19)$$

$$T_{D_2}f(\underline{x}) = x_2^0/s_2 T_{D_2^0}f(\underline{x}) + x_2^1/s_2 T_{D_2^1}f(\underline{x}), \quad (20)$$

Here, $e_i^j = (0, \dots, 1, \dots, 0)$ is the unit vector with a 1 in the x_i^j coordinate and 0 elsewhere, indicating the addition (or removal if -1) of a single patient from class j at station i . The departure operator for station 1, T_{D_1} , depends on the test results. If the result is positive,

$$T_{DR_1^{j+}}f(\underline{x}) = \min\{c_1^{j+} + f(\underline{x} - e_1^j), f(\underline{x} - e_1^j + e_2^1)\}, \quad (21)$$

if $x_1^j > 0$,

and the operator upon a negative results, $T_{DR_1^{j-}}$, can be defined similarly; l_i^{j+} and l_i^{j-} in Equation (19) denote the likelihood of the test result from station i being positive or negative, respectively, so

$$l_i^{j+} = p_0^j b_i^{j+} + (1 - p_0^j)(1 - b_i^{j-}), \quad i = 1, 2,$$

and $l_i^{j-} = 1 - l_i^{j+}$. $T_{D_1^0}f(\underline{x}) = f((\underline{x} - e_1^0)^+)$ represents the departure of an exogenous patient from station 1. The departure operator for a disease suspect from station 2 is given by

$$T_{D_2^0}f(\underline{x}) = c_2 \mathbb{1}_{x_2^1 > 0} + f(\max(\underline{x} - e_2^1, 0)), \quad (22)$$

and $T_{D_2^0}$ represents the departure of an exogenous patient from station 2 and is defined similarly as $T_{D_1^0}$. It is straightforward to write the arrival operators for exogenous patients, $T_{A_i^0}$. The last two terms in Equation (17) are dummy transitions.

The following theorem verifies that there exists a pair (ω^*, v) satisfying (17).

THEOREM 1. Assume a stable queue under the existing test $\lambda + \lambda_1^{\text{ex}} + \lambda_2^{\text{ex}} < \bar{\mu}_2$. For the average-cost optimality equation defined by Equation (17):

- (i). There exists an average-cost optimal stationary policy.
- (ii). The optimal average cost can be computed as $\omega^* = \inf_{\pi} V^{\pi} = \lim_{\delta \rightarrow 1^-} \lim_{n \rightarrow \infty} V_{n,\delta}(x)$, where $V_{n,\delta}(x)$ is the n -period discounted value function with discount factor δ .
- (iii). Let $\pi_{n,\delta}$ denote an optimal policy for the n -period discounted cost problem. Then any limit point π_{δ} of the sequence $\{\pi_{\delta,n}\}_{n \geq 1}$ as $n \rightarrow \infty$ is optimal for the infinite-horizon discounted cost problem.

Moreover, any limit point of the sequence $\{\pi_{\delta}\}_{\delta \in (0,1)}$ (as $\delta \rightarrow 1^-$) is average-cost optimal.

The proof of Theorem 1 is in Appendix B.

4.2. Structural properties

In this section, we derive structural properties on the dominance among classes under the optimal actions, which provide useful managerial insights for understanding which patients should be prioritized for routing to the new test. Moreover, these properties facilitate the search over the action space when solving the MDP. For example, under the conditions present in Theorem 2 below, if a lower risk patient is directly routed to station 2 upon arrival, given state \underline{x} , it will be optimal to route a higher risk patient to station 2 seeing the same state without the need to resolve the optimal action from (17).

THEOREM 2. Assume test sensitivity b_1^{j+} and specificity b_1^{j-} are independent of j , and the diagnostic costs satisfy

- (i) $c_1^{j+} \geq c_1^{j-}$, for $j = 1, \dots, J$,
- (ii) $c_1^{\text{FN}}(1 - b_1^+) \geq c_1^{\text{TN}} b_1^-$,
- (iii) $c_1^{\text{FN}}(1 - b_1^+) + c_1^{\text{TP}} b_1^+ \geq c_1^{\text{TN}} b_1^- + c_1^{\text{FP}}(1 - b_1^-)$,
- (iv) $b_1^+ \geq 1 - b_1^- \iff b_1^- \geq 1 - b_1^+$.

Then,

1. $a_{\text{arr}}^{j_1}(\underline{x}) \leq a_{\text{arr}}^{j_2}(\underline{x})$, for $1 \leq j_1 < j_2 \leq J$.
2. $a_{\text{dep}}^{j_1^+}(\underline{x}) \geq a_{\text{dep}}^{j_2^-}(\underline{x})$, for $1 \leq j \leq J$.

Condition (i) says the treatment cost after diagnosing a patient as having the disease is higher than the cost after ruling out the disease (no treatment). Condition (iv) requires that the probability of a true negative (b_1^-) is greater than the probability of a false negative ($1 - b_1^+$), while condition (ii) requires that the expected (pre-test) cost of a false negative is larger than the expected cost of a true negative to further penalize false negatives. Moreover, condition (iii) requires that the combined expected cost of false negative and true positive exceeds the combined expected cost of true negative and false positive. In other words, it requires that the penalty on false negatives (relative to true negatives) is larger than the penalty on false positives (relative to true positives).

Conditions (ii)–(iv) essentially require that the new test has a high sensitivity and that the test needs to be accurate enough for ruling out patients (the cost of false negative is high). These conditions hold for the D-dimer test for PE in our study and also tend to be true in general practice due to the fact that the medical community tends to prioritize sensitivity over specificity (see, e.g., Davis et al. (2005)). Result (1) of

Theorem 2 states that, for a new test satisfying these conditions, it is preferable to use the test on low-risk patients to rule out the disease. Specifically, if a lower risk patient would not use the new test upon arrival, a higher risk patient should also not use the new test. Result (2) provides a realism check on the model, stating that a patient is more likely to be discharged from the system after a negative test result at station 1 than after a positive test result. Analogously, if a test is more accurate in *confirming* a disease, it can be shown that the test should be prioritized to be used on high-risk patients (see Proposition 2 in the appendix). However, this case relates to a different setting than that of our case study and is less likely to occur in medical practice according to our clinical collaborator.

5. Decomposition Algorithm

The MDP introduced in section 3.3 suffers from the curse of dimensionality. Because the model tracks the patient count for each risk class, the state space grows exponentially when applying a finer risk stratification. For example, when $J = 3$, with a state truncation of 10 for each x_i^j and w_i^j , the size of the state space is 10^{12} . Solving $v(\underline{x})$ from the Bellman equation presented in section 4.1 with conventional value or policy iterations becomes computationally challenging, if not completely infeasible.

To overcome the curse of dimensionality, we develop a decomposition algorithm that only requires solving an MDP with a two-dimensional state space. The main idea of this algorithm is to perform an iterative, *multi-step policy improvement* based on a static routing policy that routes class k patients to station 1 with a fixed probability p_r^k and to station 2 with probability $1 - p_r^k$. Leveraging this simpler policy, we can formulate J subproblems, one for each class j . In each subproblem, we solve for the optimal arrival and routing actions for the target class j , using an approximation for the cost-to-go. This approximation is obtained from an iterative algorithm where, in each iteration, we update both the optimal actions for the target class j and the cost-to-go approximation. We initialize the iterative algorithm with the value function solved using the static routing policy for all patients. Then, in each subsequent iteration, we obtain the optimal arrival and departure routing actions for class j using the current approximation of the cost-to-go. We update the cost-to-go using the optimal actions solved in that iteration for class j while employing the static policy for all other classes.

5.1. Algorithm Description

We first specify properties and Bellman equations under a fixed, static routing policy. Then, we introduce the one-step and multi-step policy improvement used in the decomposition algorithm.

Static routing policy. Consider a static routing policy that sends each class k patient to station 1 with probability p_r^k and to station 2 with probability $1 - p_r^k$. We further assume that each patient, upon finishing service at the station she was initially routed to, directly leaves the system. We first show the following properties under this static routing policy.

LEMMA 1. *Under the static routing policy, the two stations run as two independent queueing systems, where the arrival process for station i is Poisson with rate $\lambda_i^{\text{ex}} + \sum_{\ell=1}^J \lambda^\ell p_r^\ell$ and the service completion process is Poisson with rate $\mu_i(\cdot)$ during the busy period. Furthermore, it is sufficient to track the total number of patients in each system, n_i , to recover the number of patients from each class k in service or waiting in the steady state.*

It is straightforward to show the independence between the two stations using the Poisson thinning property and the direct departure assumption. To recover the patient count of each class that are in service or waiting, we note that, under the class-independent service rate and the random-service processing discipline, the following are equivalent in the probabilistic sense: a patient's class is revealed (1) upon arrival or (2) upon admission to service or (3) upon service completion. As a result, given the total patient count n_1 in station 1, the number of patients from each class k that are in station 1, $(n_1^0, n_1^1, \dots, n_1^J)$, follows a multinomial distribution with parameters n_1 and probabilities $(q_1^0, q_1^1, \dots, q_1^J)$, where

$$q_1^0 = \frac{\lambda_1^{\text{ex}}}{\lambda_1^{\text{ex}} + \sum_{\ell=1}^J \lambda^\ell p_r^\ell}, \quad (23)$$

$$q_1^k = \frac{\lambda^k p_r^k}{\lambda_1^{\text{ex}} + \sum_{\ell=1}^J \lambda^\ell p_r^\ell}, \quad k = 1, \dots, J. \quad (24)$$

Similarly, the number of patients from each class k that are *in service* at station 1, $(x_1^0, x_1^1, \dots, x_1^J)$, also follows a multinomial distribution with parameters s_1 and probabilities $(q_1^0, q_1^1, \dots, q_1^J)$. The same results apply to the queue length count. Here, s_1 can be calculated from n_1 depending on which queueing setting we use. For example, in the $M/M/k$ setting, $s_1 = \min(n_1, k)$, and in the processor-sharing queue setting, $s_1 = n_1$. We use the function $s_1 = s_1(n_1)$ to denote the general dependence of s_1 on n_1 , and it is sufficient to just track n_1 .

Using the same argument, we can recover n_2^k and x_2^k for station 2 using binomial distributions with parameters n_2 and $s_2 = s_2(n_2)$, respectively, and probabilities

$$q_2^0 = \frac{\lambda_2^{\text{ex}}}{\lambda_2^{\text{ex}} + \sum_{\ell=1}^J \lambda^\ell (1 - p_r^\ell)}, \quad (25)$$

$$q_2^1 = \frac{\sum_{\ell=1}^J \lambda^\ell (1 - p_r^\ell)}{\lambda_2^{\text{ex}} + \sum_{\ell=1}^J \lambda^\ell (1 - p_r^\ell)} = 1 - q_2^0. \quad (26)$$

Value functions under the static routing policy.

Let $\underline{q}_1 = (q_1^0, \dots, q_1^J)$ and $\underline{q}_2 = (q_2^0, q_2^1)$. Given the properties shown in Lemma 1, we consider the two-dimensional state space (n_1, n_2) . For any given fixed routing probabilities p_r^k ($k = 1, \dots, J$), we can solve the value functions $V^r(n_1, n_2)$ from the following Poisson equation:

$$\begin{aligned} \omega^r + V^r(n_1, n_2) &= \mathbb{E}[h_1(n_1, \underline{q}_1)] + \mathbb{E}[h_2(n_2, \underline{q}_2)] \\ &+ \lambda_1^{\text{ex}} T_{A_1^0} V^r(n_1, n_2) + \lambda_2^{\text{ex}} T_{A_2^0} V^r(n_1, n_2) \\ &+ \sum_{k=1}^J \lambda^k T_{AR^{r,k}} V^r(n_1, n_2) \\ &+ \mu_1(s_1) T_{D_1^r} V^r(n_1, n_2) \\ &+ \mu_2(s_2) T_{D_2^r} V^r(n_1, n_2) \\ &+ (\bar{\mu}_1 - \mu_1(s_1)) V^r(n_1, n_2) \\ &+ (\bar{\mu}_2 - \mu_2(s_2)) V^r(n_1, n_2). \end{aligned} \quad (27)$$

Here, we use the superscript “ r ” over the value function and the arrival/departure operators to emphasize that they depend on the given static routing policy. The holding cost $\mathbb{E}[h_i(n_i, q_i)]$ is the expected version of the original $h_i(x_i)$ with respect to the multinomial distribution. The arrival routing operator $T_{AR^{r,k}}$ follows

$$T_{AR^{r,k}} = p_r^k T_{A_1^0} + (1 - p_r^k) T_{A_2^0}, \quad (28)$$

where $T_{A_1^0} f(n_1, n_2) = f(n_1 + 1, n_2)$ and $T_{A_2^0} f(n_1, n_2) = f(n_1, n_2 + 1)$. Combining terms, we have

$$\begin{aligned} &\sum_{k=1}^J \lambda^k T_{AR^{r,k}} V^r(n_1, n_2) \\ &= \sum_{k=1}^J \lambda^k \left(p_r^k T_{A_1^0} V^r(n_1, n_2) + (1 - p_r^k) T_{A_2^0} V^r(n_1, n_2) \right). \end{aligned}$$

Under the assumed service setting, we can further write the departure operator $T_{D_i^r}$ as

$$\begin{aligned} T_{D_1^r} f(n_1, n_2) &= \sum_{k=1}^J q_1^k \left(l_1^{k,+} T_{DR_1^{r,k,+}} f(n_1, n_2) + l_1^{k,-} T_{DR_1^{r,k,-}} f(n_1, n_2) \right) \\ &+ q_1^0 T_{D_1^0} f(n_1, n_2), \quad \text{if } n_1 > 0, \\ T_{D_2^r} f(n_1, n_2) &= q_2^0 T_{D_2^0} f(n_1, n_2) + q_2^1 T_{D_2^1} f(n_1, n_2), \quad \text{if } n_2 > 0, \end{aligned}$$

Here, q_i^k is the proportion of class k patients in station i from the multinomial distribution where $q_i^k \mu_i(s_i)$ is the rate with which a departure of class k occurs. $T_{D_1^0}$, $T_{D_2^0}$ and $T_{D_2^1}$ are the two-dimensional analogues of the previously presented departure operators, and $T_{DR_1^{r,k,+}}$, $T_{DR_1^{r,k,-}}$ follow from the assumption of direct departure, for example,

$$T_{DR_1^{r,k,+}} f(n_1, n_2) = c_1^{j,+} + f(n_1 - 1, n_2).$$

We treat these probabilities, $p_r^{k'}$ s, as tunable parameters. In the implementation, we start from the $p_r^{k'}$ s obtained from the policy that mimics the current practice (i.e., routing low-risk patients to station 1 and other patients to station 2) and then fine tune these $p_r^{k'}$ s with line search to minimize the average cost from the decomposition algorithm.

One-step policy improvement. For class j , we perform a one-step policy improvement based on $V^r(n_1, n_2)$ by replacing the static operators with the original operators. That is, we replace $T_{AR^{r,j}}$, $T_{DR_1^{r,j,+}}$, and $T_{DR_1^{r,j,-}}$ in Equation (27) with the original operators T_{AR^j} , $T_{DR_1^{j,+}}$, and $T_{DR_1^{j,-}}$ for class j and apply them to $V^r(n_1, n_2)$. Then, we obtain the class-dependent arrival and departure routing actions for class j by finding the actions that are the minimizers of:

$$\begin{aligned} T_{AR^j} V^r(n_1, n_2) &= \min\{V^r(n_1 + 1, n_2), V^r(n_1, n_2 + 1)\}, \\ T_{DR_1^{j,+}} V^r(n_1, n_2) &= \min\{c_1^{j,+} + V^r(n_1 - 1, n_2), \\ &V^r(n_1 - 1, n_2 + 1)\}. \end{aligned}$$

We do the same for $T_{DR_1^{j,-}}$. These class-specific operators allow state-dependent routing for each class j patient, which can be seen from the minimization operation that does not appear in the static routing policy formulation.

Multi-step policy improvement. For the multi-step policy improvement, we use $V^r(n_1, n_2)$ as an initial solution and then iteratively update the value function using the following equation. In iteration $t + 1$, we get $V^{j,(t+1)}(n_1, n_2)$ for class j by

$$\begin{aligned} \omega + V^{j,(t+1)}(n_1, n_2) &= \mathbb{E}[h_1(n_1, \underline{q}_1)] + \mathbb{E}[h_2(n_2, \underline{q}_2)] \\ &+ \lambda_1^{\text{ex}} T_{A_1^0} V^{j,(t)}(n_1, n_2) + \lambda_2^{\text{ex}} T_{A_2^0} V^{j,(t)}(n_1, n_2) \\ &+ \lambda^j T_{AR^j} V^{j,(t)}(n_1, n_2) + \sum_{k \neq j} \lambda^k T_{AR^{r,k}} V^{j,(t)}(n_1, n_2) \\ &+ \mu_1(s_1) T_{D_1^j} V^{j,(t)}(n_1, n_2) + \mu_2(s_2) T_{D_2^j} V^{j,(t)}(n_1, n_2) \\ &+ (\bar{\mu}_1 - \mu_1(s_1)) V^{j,(t)}(n_1, n_2) + (\bar{\mu}_2 - \mu_2(s_2)) V^{j,(t)}(n_1, n_2), \end{aligned} \quad (29)$$

where $V^{j,(0)} = V^r(n_1, n_2)$, and

$$\begin{aligned} T_{\bar{D}_1} f(n_1, n_2) &= \sum_{k=1}^J q_1^k \left(l_1^{k,+} T_{DR_1^{k,+}} f(n_1, n_2) \right. \\ &\quad \left. + l_1^{k,-} T_{DR_1^{k,-}} f(n_1, n_2) \right) \\ &\quad + q_1^0 T_{D_1^0} f(n_1, n_2). \end{aligned} \quad (30)$$

Equation (29) is obtained from (27) by replacing the arrival operator for class j with T_{AR^j} , and the departure operators for each class with $T_{DR_1^{k,+}}$ and $T_{DR_1^{k,-}}$ in $T_{\bar{D}_1}$. We solve (29) for τ iterations to get $V^{j,(\tau)}(n_1, n_2)$. Applying the class- j operators to $V^{j,(\tau)}(n_1, n_2)$ in the last iteration τ , we obtain the arrival and routing actions for class j as we did in the one-step improvement. This multi-step policy improvement algorithm is in the same spirit as the multi-step greedy approach presented in Efroni et al. (2018a, b). Recent papers report empirical success of the multi-step policy improvement over one step in the performance of the original MDP (Efroni et al. 2018a); our numerical analyses support these findings. We leave it to future research to rigorously justify the performance improvement of this multi-step approach.

REMARK 2. One drawback of the algorithm presented here is that, in the departure operator $T_{\bar{D}_1}$, the proportions q_1 and q_2 remain the same no matter which class j we are looking at. This may lead to a larger optimality gap when patients of a certain class k constitute the majority of the patient population. That is, when q_k is close to 1, the class j departure decisions are solved mostly based on the characteristics of class k patients, rather than those of class j . To remedy this, we propose a refinement, which, as mentioned, explicitly tracks the number of patients from the target class in station 1; details are provided in Appendix A.

REMARK 3. Using a technique based on event-based dynamic programming, it is possible to show that a threshold policy is optimal when there is only a

single class of patients. This threshold result is relevant to our decomposition algorithm that considers only one target class j in each of the J subproblems. Specifically, if both stations use the $M/M/1$ queuing structure, it is optimal to route the patient to station 1 if the number of patients in station 2 exceeds some threshold that depends on the number of patients in station 1, $n_2(n_1)$. However, we are not able to extend this result to $M/M/c$ queues, as Koole (2007) pointed out that proving the threshold policy for multiple-server queues remains an open problem.

5.2. Numerical Validation of the Decomposition Algorithm

To validate the performance of our decomposition algorithm, we numerically compare the long-run average cost of our algorithm with the “true” optimal policy obtained from value iteration. We refer to the policy obtained from the decomposition algorithm introduced in section 5 as the *basic approximate policy*, and to that obtained from the refined algorithm detailed in Appendix A as the *refined approximate policy*. Due to the curse of dimensionality, value iteration can only solve a three-class patient problem (i.e., $J = 3$) without arrivals of exogenous patients to station 1. Thus, our comparison below focuses on this setting.

Table 1 reports the average cost under different system load conditions. The parameters are derived from data at our partner ED, with details on parameter estimation given in section 6.1. The optimality gap for each approximate policy is defined as $(\text{approximate policy cost} - \text{optimal cost}) / \text{optimal cost}$. Optimality gaps are $<1.5\%$ under all load conditions.

We also compare the policies from our algorithms with several other naïve policies, including a point-threshold policy and a static policy that sends all suspected patients to the new test first; see Appendix C. In most parameter settings, the benefit gained from the refined heuristic policy is more than 5% compared with the threshold policy, and 10% compared with several other static policies. Furthermore, in Table 5 in Appendix C, we extend our study beyond our

Table 1 Long-Run Average Daily Costs of Different Policies

	Lightly loaded	Intermediately loaded	Heavily loaded
Optimal	6.90±0.01	11.49±0.02	15.74±0.03
Basic approx policy	6.98±0.01	11.65±0.02	15.86±0.04
Optimality gap	1.29%	1.33%	0.76%
Refined approx policy	6.92±0.01	11.50±0.02	15.75±0.02
Optimality gap	0.31%	0.05%	0.09%

Notes: We set $J = 3$, $p_0^1 = 0.05$, $p_0^2 = 0.35$, $p_0^3 = 0.65$. For $j = 1, 2, 3$, $b_1^{j,+} = 0.9819$, $b_1^{j,-} = 0.4249$, $\lambda^j = 0.2$. $\lambda_1^{\text{ex}} = 0$, $\lambda_2^{\text{ex}} = 0.6$, $\mu_1(x) = 0.9x^{0.5}$. $c_1^{\text{TP}} = c_1^{\text{TN}} = c_2 = 0$, $c_1^{\text{FP}} = 100$, $c_1^{\text{FN}} = 800$, $h_1(x) = 6s_1$, $h_2(x) = 6s_2$. Routing probabilities for the heuristic policies are uniformly set as (0.6, 0.6, 0.6). Left panel: $\mu_2(x) = 2.2$; middle panel: $\mu_2(x) = 1.7$; right panel: $\mu_2(x) = 1.45$. The number after the \pm sign is the half-width of the corresponding 95% confidence interval

partner ED by comparing the optimality gaps under a variety of system conditions. We observe that the optimality gaps of the basic heuristic policy are <3% under most settings, and the maximum optimality gap is 5.2%. The optimality gaps of the refined heuristic policy are <1.5% under all settings.

6. PE Testing in Emergency Departments

In this section, we apply our framework to analyze adoption and integration of a new PE test into the ED at our partner hospital. In our partner ED, all patients with pre-test scores (from a simple scoring tool) above a predefined threshold are sent for a CT to confirm the diagnosis. The coarseness of this initial screening results in too many patients being sent to CT, which has both operational and medical consequences, including overuse of highly utilized CT machines and potential hazards to the patients due to radiation exposure or kidney damage from the diagnostic dye. To avoid overuse of the CT, our collaborator has proposed using a diagnostic biomarker test—the D-dimer test—as a *rule out* tool for PE, given its high true negative rate. However, this test also has a high false positive rate. If the D-dimer is not discerning enough, excessive double-testing can increase the overall time a patient spends in the ED, which increases the ED physicians' workload. Our collaborator's concern is that if new tests are seen by physicians as creating extra work, they will not be adopted into clinical practice.

In this case study, we address our collaborator's primary concerns and, more broadly, study how test characteristics and the manner of integration affect the adoption and performance of a new test. In section 6.1, we introduce the dataset and parameterization for our analytical framework. In section 6.2, we study adoption decisions under the medical and operational criteria, and we demonstrate how the manner of test integration affects the adoption regions.

6.1. Data Description and Model Parameterization

Dataset. The data were obtained from a large teaching hospital in the state of Indiana. The data consist of a test order log and ED patient movement data, where the data entries were obtained from eight randomly selected weeks between July 2013 and July 2015 according to the admission time of the patients, where the random selection is to conform with IRB requirements. The two datasets were merged using de-identified patient IDs. Each record in the merged dataset contains a patient's entire care history in the hospital, from arrival to the ED until discharge from the hospital. Patient arrival data contains arrival date and time,

visit reason, and main complaints, etc. The testing data include test name, test order date and time, and test completion date and time. The discharge data include discharge date and time, discharge diagnosis, and discharge condition.

In total, 10,254 patients visited the ED during the selected weeks, accounting for 103,514 ED testing entries. Tests that were performed during the inpatient stay are excluded from our ED study. Out of the 10,254 patients, 6,698 patients (65.3%) were discharged after the ED stay, 2,535 patients (24.7%) were admitted as inpatients, 998 patients (9.7%) were admitted under hospital observation status, and 23 patients (0.2%) were bedded outpatients.

Arrival rates. Four main complaints of patients with PE are chest pain, shortness of breath, dyspnea, and difficulty breathing. First, we included patients sent to the CT (chest test) that exhibited these four symptoms as PE suspects, which is recommended by our clinical collaborator. Additionally, we considered all patients sent to the D-dimer as PE suspects, as this test is used exclusively for diagnosing PE. The total arrival rate of PE patients to our partner ED is determined to be $\lambda = 0.65$ patients per hour. For arrivals to CT, non-PE suspects accounted for $\lambda_2^{\text{ex}} = 1.06$ patients per hour. We assume time-stationary arrivals in the experiments presented in section 6, and we consider time-varying arrivals in section 7.2 as sensitivity analysis.

Service rates and departure routing. We illustrate the empirical distribution of CT occupancy levels in Figure 4 using a solid line, where the occupancy level at a given time includes all patients who are waiting for or are receiving the CT test. This empirical occupancy distribution is calculated from data on each patient's CT request time and test finish time. In our data, we only have the time when a test was requested and the time the test was finished, but not when the test was started. In other words, we cannot estimate the queue length at the CT test. Hence, we choose to estimate a service rate that best matches the empirical occupancy distribution from the data. Among different distributional fittings, we find that using a processor-sharing queue with a service rate function $\mu_2(x) = 0.9x^{0.5}$ best replicates the empirical occupancy distribution curve (minimizing the mean squared errors); see the dashed line in Figure 4. For comparison, we also plot the occupancy distribution using the conventional $M/M/1$ queue with $\mu_2(x) = 1.91$, which clearly deviates from the empirical distribution. We believe that the processor-sharing queue provides a better fit because the CT testing equipment is used in complex ways in the ED. For example, doctors sometimes order priority STAT tests (non-FIFO order), and the CTs in our partner ED have a time-varying number of servers as some of the

machines are not staffed at certain times or are occasionally dedicated for inpatients. These complexities could make traditional $M/M/1$ and $M/M/k$ queueing systems a poor fit for the empirical occupancy distribution, whereas the processor-sharing queue is more flexible to accommodate these complexities commonly seen in the ED testing environment.

We estimate the service rate function for the new test (D-dimer) in a similar way. We find that $\mu_1(x) = 0.96x$ is the best fit for the empirical occupancy distribution. Thus, in the rest of the numerical study, we adopt a processor-sharing queue for both tests. We perform sensitivity analysis on alternative queueing models in section 7.2. In addition, we follow the current practice and fix the departure routing, that is, patients getting a positive result at station 1 must be sent for follow-up at station 2.

Patient characteristics. Physicians usually stratify PE-suspected patients into low, intermediate and high-risk groups based on their pretest scores (Kline et al. 2014, Kohn et al. 2017). We infer the pretest probability and the proportion of each class from the data in Kline et al. (2014), who conducted clinical trials at three EDs and one community hospital.

D-dimer characteristics. Most hospitals adopt a uniform D-dimer cutoff of $0.5\mu\text{g/mL}$ for all PE suspected patients (Kohn et al. 2017, Linkins et al. 2013). This D-dimer cutoff is a threshold to distinguish between negative and positive test results. Kohn et al. (2017) study the effect of different D-dimer cutoffs on diagnosing PE. Using their data, we derived the sensitivity and specificity of the D-dimer test at a $0.5\mu\text{g/mL}$ cut-off, which are reported in Table 2.

Cost parameters. There are two basic types of diagnostic costs: (1) the cost of an inaccurate diagnosis (false negative or false positive) and (2) the cost of performing a test. For the first type, we normalize c_1^{TP} and c_1^{TN} to be 0. For the second type, we normalize the cost of performing the PE test to 0, since blood tests are cheap relative to a CT. Following Kohn et al. (2017), who suggest an upper bound of 3% post-test probability for PE diagnostic tests, we set c_1^{FP} , c_1^{FN} , and c_2 to satisfy (16) with $\alpha = 3\%$. For the congestion cost,

we adopt a linear form with $h_i(x) = hs_i$, $i = 1, 2$. In the baseline experiment, we set $h = 30$. Later we will show that these cost parameters, c 's and h , illustrate the weight that hospital managers put on congestion vs. individual patient outcomes and can be used as tuning parameters for managers to achieve desired performance metrics.

6.2. Adoption Decision in our Partner ED

In this section, we apply our framework for evaluating test adoption and integration for diagnosing PE in the parameter regime of our partner ED. We demonstrate the interplay between medical and congestion effects for test adoption in section 6.2.1, and how the manner of test integration plays an important role in the adoption regions in section 6.2.2. A key insight from this section is that considering the clinical environment, rather than the test in isolation, leads to the conclusion that the impact of specificity is likely to have been overlooked by the medical community, because the current evaluation of diagnostic tests ignores how the test will be used in practice. From the operational standpoint, specificity is a key driver in how a test is implemented and how effectively it will function in combination with the existing test. Measuring the test in isolation downplays the importance of specificity since sensitivity is the primary driver of the misdiagnosis rate Equation (3) under the medical criterion. Our operational framework reveals the important role that specificity plays in both the operational and medical value of a test, whereas sensitivity—the primary consideration for most medical research—primarily impacts the medical criterion with little impact on doctor workloads and patient delays.

We summarize the major findings from the rest of this section. First, a test with high sensitivity and low specificity may meet medical standards but could be rejected in practice due to increased workloads. Conversely, a test with lower sensitivity and high specificity may be rejected even when the benefits of reduced workload and CT avoidance may outweigh the slightly higher false negatives. Second, we find

Table 2 Base Parameters for PE Setting

Queueing parameters	$\mu_1(x)$ 0.96x	$\mu_2(x)$ $0.96x^{0.5}$	λ 0.65	λ_1^{ex} 0	λ_2^{ex} 1.06
Test parameters and patient characteristics	j	p_0^j	p^j	$b_1^{j,+}$	$b_1^{j,-}$
	1	5%	80%	98.19%	42.49%
	2	15%	16%	98.19%	42.49%
	3	20%	4%	98.19%	42.49%
Diagnostic costs	c_1^{FP}	c_1^{FN}	c_1^{TP}	c_1^{TN}	c_2
	100	800	0	0	24
Unit holding costs	$h_1(x) = hs_1; h_2(x) = hs_2; h = 30.$				

Notes: The service rate functions, arrival rates and unit holding costs are on the hourly scale.

the current medical criterion to be misleading because (1) the new test will not be used on all patients and (2) false positives can be ruled out by subsequently sending the patient to the existing test. This causes the myopic medical criterion to overestimate the true misdiagnosis rate in the system context since some patients will be sent directly to the existing test and receive a correct diagnosis, which reduces the misdiagnosis rate relative to just measuring the misdiagnosis rate at the new test alone. In addition, the medical criterion alone is too restrictive (unnecessary rejection) at low specificities because false positives cause a larger misdiagnosis rate when considering the individual test, whereas in practice, false positives eventually become true negatives when the patient is sent to the CT (avoiding the misdiagnosis). Third, we find that sensitivity has little impact on congestion, which is not considered in the medical criterion, whereas specificity has a strong impact. Moreover, when using a dynamic routing policy, specificity has a stronger impact on the medical value of the test because it drives greater usage of the test due to medical benefits of CT avoidance as well as operational benefits in reduced patient delays. Finally, we find that the dynamic policy is able to adapt usage to test characteristics while the static policy does not. This adaptability is particularly important for reducing CT usage and doctor workloads. However, improving test speed mitigates the inefficiencies of a simpler routing policy, making tests more easily and broadly adoptable.

6.2.1. Medical vs. Operational Considerations.

Figure 5 presents an example of a primary output of this research that partitions the test characteristic space into adoption vs. non-adoption regions under different criteria. We call the line that partitions the space the *adoption boundary* for the corresponding criterion. The dashed and dotted lines in Figure 5 illustrate the levels of sensitivity and specificity that a new test for PE must achieve to meet the medical

criterion (for low-risk patients only, using $\alpha = 3\%$) and the operational criterion ($\gamma = 7.5\%$), respectively; the choices of α and γ are based on the medical literature Kohn et al. (2017) and input from our clinical collaborator. We also plot the congestion adoption boundary as a solid line using $\beta = \gamma = 7.5\%$. The three boundaries in Figure 5 reflect how the adoption boundary shifts from medical to congestion considerations. To generate these adoption boundaries we use the optimal routing policy calculated from the MDP with proper cost parameterizations as introduced in section 3.3.3. Next, we discuss the regions formed by the partitions in Figure 5.

Overvalue vs. undervalue. Two interesting regions in the partition are the upper left and lower right quadrants. In the upper left quadrant colored light gray (medical only region), the sensitivity is high and specificity is low, in which case considering only the medical criterion overvalues the new test; researchers may consider the test sufficiently accurate to suggest its use in EDs only to find that it would be rejected in practice due to increased patient delays and doctor workloads. The high sensitivity means few false negatives and hence a low misdiagnosis rate from the new test, which is important from the medical standpoint. However, the low specificity generates many false positives and hence a significant amount of double testing, which increases physician workload and delays patient discharge.

In the lower right quadrant shaded dark gray (operational only region), the sensitivity is lower and specificity is high, which leads the medical criterion to undervalue the new test; researchers may not consider the test worthy of pursuing due to concerns over medical accuracy, although adopting it in practice may alleviate congestion at the CT and avoid potential harm from unnecessary CTs. Undervaluing a test may be particularly relevant in cases where a condition is not imminently life threatening (e.g., testing for glaucoma, which is a very slowly progressing disease), since there is less risk of an immediate misdiagnosis and the improvement in patient throughput/avoidance of a potentially harmful or arduous test may outweigh an initial misdiagnosis.

Tradeoff between medical and congestion considerations. To further demonstrate the potential of relaxing one criterion for significant gains in the other, we compare several points that fall outside either the medical or congestion adoption region; see Points 1–3 in Figure 5. Table 3 reports the performance metrics for these points using routing policies solved from the MDP. Note that the misdiagnosis rate reported here is for *system-level* misdiagnosis, which is the proportion of patients who get a false negative result (incorrectly discharged) among all the PE-suspected patients.¹ Our main observations are:

Figure 5 Test Adoption Decisions with $\alpha = 3\%$ (for Low-risk Only) and $\beta = \gamma = 7.5\%$. The performance metrics for points 1–3 are reported in Table 3

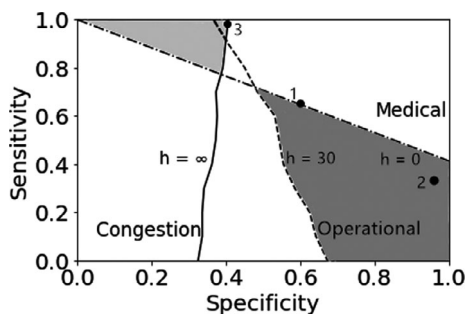


Table 3 Performance Metrics Under Different Test Characteristics. Points 1–3 are Plotted in Figure 5

No	% CT avoidance	Misdiag - Sick	$E[T]$	$E[T_{PE}]$	CT TAT (hr)	Avg cost
1	49%	2%	2.02	1.89	2.08	146.47
2	82%	3%	1.65	1.21	1.91	129.65
3	31%	0.07%	2.29	2.37	2.24	154.92
CT Only	0%	0%	2.44	2.44	2.44	166.13

- Comparing Points 1 and 2, Point 2 relaxes the medical criterion and would be rejected under the current medical guidelines. However, Point 2 still lies within the operational adoption region. We see that Point 2 only increases the system-level misdiagnosis rate from 2% to 3% while providing significant operational gains compared to Point 1. The throughput time for all ED patients and PE patients can be reduced by 22 and 41 minutes respectively. Furthermore, CT turnaround (time to get a CT) is reduced by 10 minutes per patient, and more unnecessary CT scans are avoided (82% rather than 49%), potentially saving lives by avoiding a reaction to the CT dye and exposure to radiation.
- In contrast, Point 3 lies in the region where the new test is not adopted under the congestion criterion but is adopted under both the medical and operational criteria. This is because misdiagnosis rate is dramatically reduced (to 0.07%) although the throughput time is slightly increased, though it is still better than using the CT alone. This further illustrates how the operational criterion weighs the tradeoff between the congestion and the system-level misdiagnosis rate.

Insights for medical research. A key reason for the significantly larger adoption region based on the operational cost-based criterion, in contrast to the adoption region under the medical criterion only, is that this operational criterion accounts for the consequences of using the CT (e.g., stroke and death)—this was a primary motivation for our clinical collaborator in developing a new test for PE. The medical criterion only focuses on the individual-level misdiagnosis rate. This is reflected in the fact that, despite the medical criterion being relaxed more significantly in Point 2, the misdiagnosis rate is counterbalanced by the CT avoidance and the shortened throughput time. This analysis highlights the shortsightedness of considering only the individual-level misdiagnosis rate in evaluating the benefits of new medical research; new tests should be considered in the broader, *system* context in which they will be used, not just as standalone diagnostics. Our framework allows managers to qualitatively weigh the costs of misdiagnosis with the benefits of CT avoidance as well as the system benefits to

ED patients needing a CT and doctor workload levels, all of which impact quality of care.

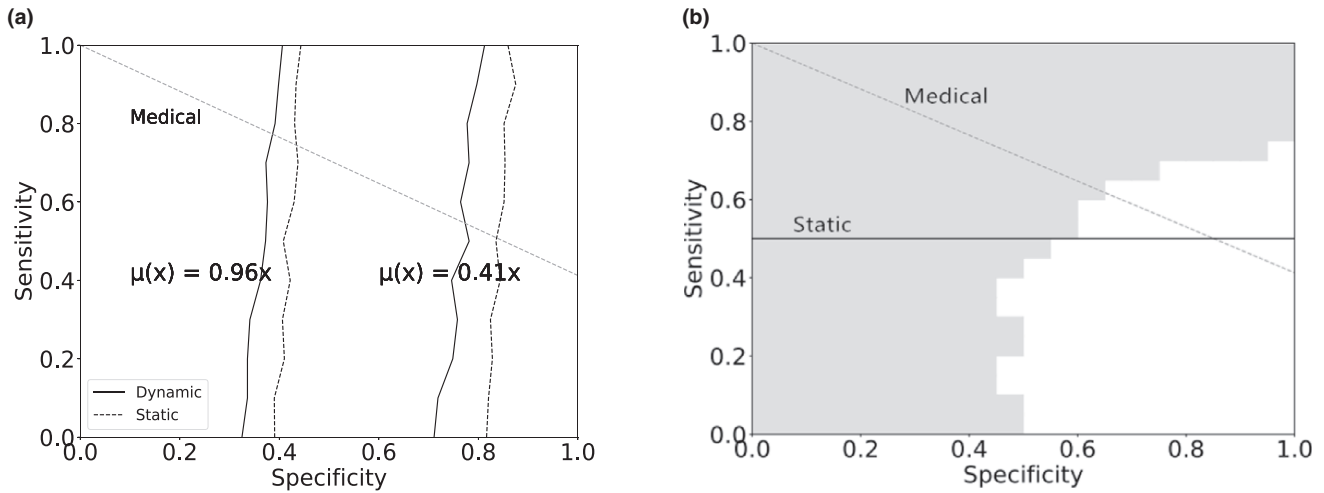
More broadly, from the shape of these adoption regions, one can derive focused guidelines regarding which aspects of a new test medical researchers should target to increase chances of adoption into clinical practice. Interestingly, specificity impacts both medical and congestion adoption, whereas sensitivity has little impact on reducing doctor workloads and patient delays. This indicates that, perhaps specificity has been undervalued in medical research development. Specifically, if rejection by physicians due to increased workload is a major concern, researchers should focus on almost exclusively on *specificity* to reduce double testing as sensitivity has little impact on reducing doctor workloads; see the slope of the congestion boundary. On the other hand, if misdiagnosis is more of a concern, then researchers should focus more on *sensitivity*, though, while specificity has less impact on the medical adoption it still plays a role.

6.2.2. Impact of Integration of Test into ED Workflow. We conclude the discussion of medical research adoption by examining the impact of how new diagnostic tests are integrated into clinical workflow. For comparison with the MDP's dynamic routing policy, we consider a *static policy* that mimics the practice suggested by the medical literature for diagnosing PE; this policy routes only low-risk patients to the new test (D-dimer) and all others directly to the current test (CT).

Impact of integration on congestion boundary. Figure 6(a) shows the adoption region under the congestion criterion for different integration policies and different service rates. The solid (dotted) line shows the adoption boundary under the dynamic (static) routing policy. This partition shows that the simpler static policy requires a higher level of specificity due to the impact of double testing of patients who do not actually have the suspected disease (PE). In contrast, the dynamic policy avoids routing patients to the new test when the new test is highly congested because the long waiting time for the new test will significantly increase the impact of double testing.

Moreover, when the test is slower, the manner of integration into practice has a more significant impact on the level of accuracy that the test must achieve to

Figure 6 Integration into Clinical Workflow (a) Impact of routing policies on the congestion adoption boundary. the boundaries when $\mu_1(x) = 0.96x$ are shown in the left two curves and the ones when $\mu_1(x) = 0.41x$ are shown in the right two curves (b) Impact of routing policies on medical adoption boundary. the grey boxes are for regions when the system-level misdiagnosis is below 2% for the optimal routing policy, while the solid line is the boundary for 2% under the static policy



be suitable for adoption, indicated by the larger separation between the static and dynamic policies (the two vertical lines to the right). Thus, focusing on test speed during the research process can mitigate some of the drawbacks of using a simpler operational policy in clinical practice and make the test more widely adoptable for hospitals without the means to develop and implement more sophisticated routing policies.

Impact of integration on medical boundary. Figure 6b shows the adoption regions based on system-level and individual-level (myopic) misdiagnosis rates. The light grey region shows the adoption region using a system-level misdiagnosis rate of <2% and the optimal, dynamic policy (using the baseline cost parameters in Table 2). We use a 2% system-level misdiagnosis rate because this roughly corresponds to a 3% individual-level misdiagnosis rate using the static policy (averaged across different sensitivity and specificity combinations). The area above the solid line represents the adoption region using the 2% system-level misdiagnosis rate and the static policy. For comparison, we also plot the dashed line for the medical criterion, that is, the area above the dashed line represents the adoption region where the individual-level misdiagnosis rate is <3%.

This figure demonstrates the myopic nature of the individual-level misdiagnosis criterion. Compare the two lines (dashed and solid) that represent the static policy under system and individual-level criteria. The individual-level medical criterion underestimates the benefit of the new test at lower specificities, while it overestimates the benefit of the new test at higher specificities compared with the more holistic system-level misdiagnosis criterion. This is because the individual-level criterion is conditional

on a patient receiving a negative test result, which is, in some ways misleadingly, a function of the number of false positives. Reducing false positives increases the total number of negatives and increases the chance of adoption under the medical criterion (it reduces the left-hand side of Equation (3)). However, at the system level, false positives are eventually ruled out by the CT so the overall number of patients that are misdiagnosed remains unchanged regardless of the specificity. This leads the individual-level criterion to overvalue a test with high specificity and undervalue a test with low specificity in the larger context of the testing system. Hence, a systematic evaluation framework such as the one developed in this study is critical for evaluation of a new test in the system context.

Comparing the gray region with the solid line, we see that the dynamic policy accentuates the mismatch between the system-level and individual-level criteria. This is because when specificity is high, the test is used more frequently by the dynamic policy to better avoid the negative health consequences of unnecessary CT scans; high specificity implies less chance of ordering an unnecessary CT. With greater usage of the test, a higher sensitivity is needed to ensure the misdiagnosis rate remains low, particularly if the test is used on medium and high risk-level patients. This further highlights the myopic nature of considering a new test in a vacuum. When specificity is low, static policies undervalue the new test because they are not able to adjust the usage of the test to account for the implications of double testing. The dynamic policy, on the other hand, will use a lower specificity test less frequently, leading to lower overall misdiagnosis rates by increasing the usage of the more accurate CT.

7. Sensitivity Analyses: Impact of Test Characteristics and ED Environment

In section 7.1, we conduct a detailed performance analysis to show how different test characteristics affect medical and congestion performance metrics. When varying each characteristic dimension, we compare the dynamic and the static policy. These analyses shed light on what features researchers should focus on to improve certain performance metrics and whether using a simple routing policy is sufficient (or not) to derive much of the value of a new test. In section 7.2, we generalize our analysis to clinical settings beyond our collaborating ED, by considering a broad set of clinical environments.

The key findings from this section are summarized as follows. First, we find that improving test characteristics may not improve the performance measures most closely related to that characteristic. For example, improving sensitivity and specificity can lead to greater misdiagnosis rates and improving test speed can lead to longer throughput times. Second, sensitivity primarily affects medical value while specificity affects operational value, providing guidance on which features to focus on depending on adoption concerns. Third, when the new test is sufficiently fast, the static policy performs almost as well as the dynamic policy, which has implications for ease of adoption. Finally, extending our analysis to a broad range of clinical environments, we find that the only major driver of the size of the adoption regions is the speed of the existing test; the results are relatively insensitive to all other system parameters.

7.1. Impact of Test Characteristics

In this section, we consider introducing an arbitrary new test for PE (D-dimer being a special case) and study how test characteristics affect different operational and medical performance metrics in the ED. We plot the performance changes when varying one dimension of the test characteristics ($\mu_1(x)$, b^+ , or b^-) over a continuous spectrum while fixing the other two characteristics at selected levels. All other parameters remain the same as the baseline scenario shown in Table 2. In the interest of space, we relegate all the detailed analysis to Appendix D.4 and only highlight the main findings below.

The key finding in this section is that improvements in each of the test characteristics have *mixed* results on performance in the clinical environment; positive or negative. Surprisingly, improved test speed can actually increase throughput time and doctor workload, while improved sensitivity and specificity can increase the system misdiagnosis rate. This occurs because improvements in test characteristics universally increase the usage of the new test, which brings

with it the drawbacks of misdiagnosis (not being as good as the primary test) and double testing. These drawbacks are balanced by avoidance of the adverse effects associated with unnecessary CT scans and an eventual reduction in workload by avoiding the more congested CT. Once a threshold is reached, further improving test characteristics counteracts these drawbacks, resulting in increasing reductions in workload and misdiagnosis.

7.2. Extension to Other Clinical Environments

In this section, we evaluate our framework under settings beyond our collaborating ED. We vary multiple environmental factors (arrival rates, service rate of the existing primary test) and patient characteristics (pretest probability and patient mix). Table 4 summarizes the clinical environments we study.

Key environmental factors. We perform a full-factorial analysis with all combinations of the parameter settings in Table 4, and then conduct ANOVA to identify the factors that affect the adoption regions most significantly; see Appendix D.5 for the complete details. In general, the adoption regions are not sensitive to the environmental factors and patient characteristics in our analysis. This indicates that the insights found in the previous sections (under the baseline ED settings) are generalizable to a broader range of clinical environments.

The ANOVA shows that the service rate at the existing (primary) test, $\mu_2(x)$, is the most important factor for all adoption regions; the influence of other factors is much smaller. Our sensitivity analysis also shows that the queueing structure of station 2 plays an important role in the system performance. Next, we explore the impact of these two significant factors: (i) $\mu_2(x)$ in the processor-sharing setting; and (ii) different queueing structures for station 2. We also study the impact of time-varying arrivals at the end of this section.

Impact of $\mu_2(x)$. To illustrate the impact of the speed of the *current test* (baseline setting vs. a slower speed $\mu_2(x) = 0.47x^{0.5}$), we vary the $\mu_1(x)$, b^+ and b^- of the *new test* and observe the performance metrics of cost reduction, throughput time, and misdiagnosis.

Table 4 Parameter Settings

	Low	High
$\mu_2(x)$	$0.47x^{0.5}$	$0.9x^{0.5}$
Patient mix	20%,44%,36%	80%,16%,4%
Pretest probability	5%,15%,20%	5%,30%,45%
Total arrival rate	1.71	1.91
Proportion of external arrival	0	0.62

Notes: The Bold Value in Each Row is the Same as that in the PE Setting in Section 6. Diagnostic Costs and Unit Holding Cost are Same as in Table 2, and $\mu_1(x) = 0.96x$

Compared to the baseline case, we observe the largest changes when varying $\mu_1(x)$; varying b^+ and b^- of the new test gives similar insights as in the baseline setting.

Figure 7 plots the performance metrics—cost reduction, throughput time, and misdiagnosis—under the slower speed $\mu_2(x) = 0.47x^{0.5}$ when we vary the service rate of the new test, $\mu_1(x)$. We make two main observations. First, when the current test is slower, routing just a fraction of patients to the new test can help alleviate overall congestion due to the convex nature of waiting times as a function of utilization at the existing test (e.g., CT). Thus, introducing a new test and/or improving the new test speed has a greater impact on cost reduction; see Figure 7a. The significant improvement in the congestion metrics can also be observed from the significant reduction in the ED throughput time in Figure 7b. Second, in contrast with the baseline case (Figure 12c in Appendix D.4), we see in Figure 7c that, if the current test is slower, the dynamic policy dominates the static policy in terms of current test avoidance. This is because the dominating effect of workload reduction, combined with the benefits of current test avoidance, leads to far greater usage of the new test and hence far fewer patients being sent to the existing test. These observations again highlight the importance of considering a new test within the overall clinical environment.

Impact of queueing structure at station 2. Next, we consider alternative queueing structures $M/M/1$ (with $\mu_2(x) = 1.91$) and $M/M/k$ for station 2. Figure 8 plots the performance metrics—cost reduction, throughput time, and misdiagnosis—when we vary the service rate of the new test, $\mu_1(x)$. Comparing Figure 8a with Figure 7a, we observe an even more significant cost reduction using the dynamic and static routing policies over the CT only policy. This is because the $M/M/1$ queue is more sensitive to the workload since the single server suffers more from the variability in service time. As a result, diverting a

small number of patients to the new test can greatly alleviate the congestion at CT. This is also shown in Figure 8b: even for the slowest service rate of the new test ($\mu_1(x) = 0.2x$), introducing the new test reduces the throughput time. In other settings the throughput time actually increases over the CT only policy when the new test is slow, for example, Figure 7b. When using an $M/M/3$ queue for station 2, the performance metrics are closer to those under the baseline setting. This is likely due to the fact that the $M/M/3$ queue produces a similar CT occupancy level as the processor-sharing queue used in the baseline experiments; see Figure 10 in Appendix D.3. In addition, we also consider an $M/M/K(t)$ setting for station 2 where the number of servers is time-varying. Specifically, we create a model with three CTs available during the day shift and two CTs during the night shift. This setting produces an occupancy distribution that matches the empirical distribution better than any of the $M/M/n$ models. The performance in this setting, though, is similar to the performance in the $M/M/3$ setting. See Appendix D.3 for the detailed numerical results. The key takeaway from this analysis of different queueing models is that our framework is flexible enough to capture a wide variety of testing structures that may occur in practice. This suggests that our methods can be applied to a wide range of testing systems.

Impact of time-varying arrivals. Lastly, we consider the impact of time-varying arrivals. Incorporating time-varying arrivals requires expanding the state space to include a time component, leading to a continuous state space. It is known that exactly solving time-varying MDPs is non-trivial (Marecki et al. 2006, McMahon 2008) except for settings with restricted value functions (Boyan and Littman 2000). Thus, we adopt a commonly used technique, Pointwise Stationary Approximation (PSA), to tackle this challenge when the arrival rate function is periodic. The main idea of PSA is to first partition the time period into several intervals, and then solve the stationary

Figure 7 Impact of New Test Service Rate when $\mu_2(x) = 0.47x^{0.5}$ (a) Cost reduction (b) Throughput time (c) % of Patients Avoiding CT

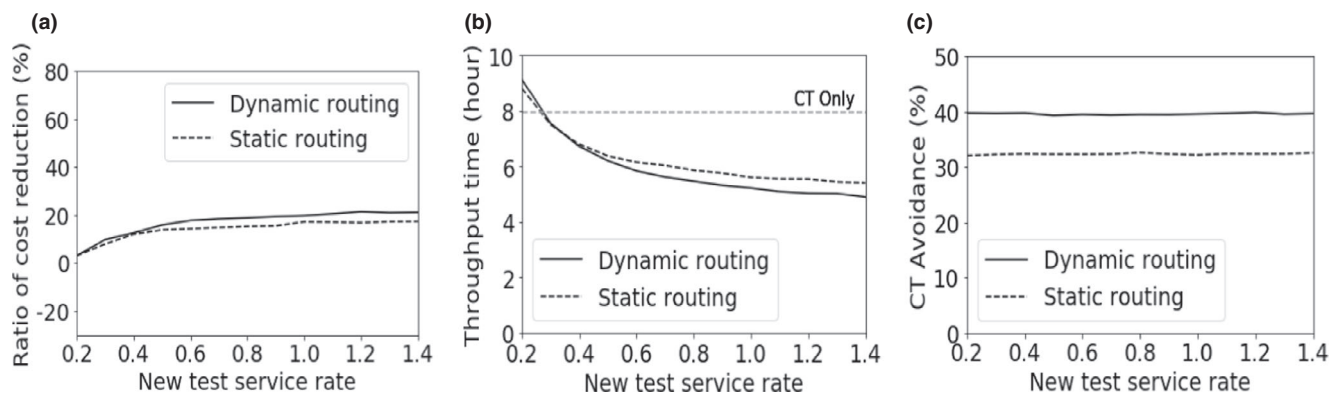
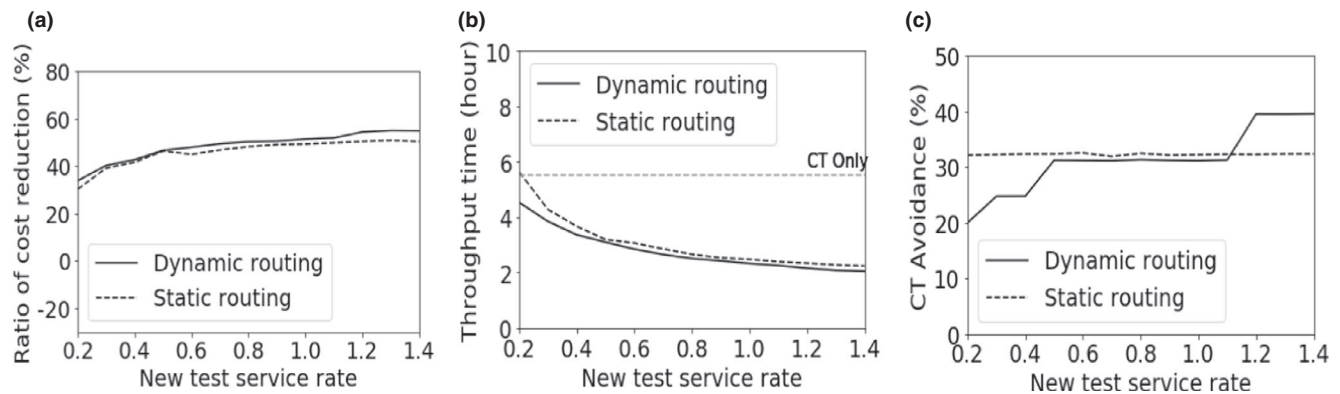


Figure 8 Impact of New Test Service Rate when Using the $M/M/1$ Queue Setting for Station 2 with $\mu_2(x) = 1.91$ (a) Cost Reduction (b) Throughput time (c) % of PATIENTS avoiding CT



version of the MDP using the average arrival rate from each interval. When making a decision, we use the policy from the corresponding time interval. Figure 9 in Appendix D.2 plots the time varying arrivals from our partner ED as well as the arrival rates used in our PSA, which leads to a piecewise-linear policy for different time intervals. Yoon and Lewis (2004) provide mathematical justification for the use of PSA admission control in time-varying MDPs.

In the numerical experiments, we apply the PSA solution to the simulated system where the time-varying arrival process is used, and we compare its performance with those from the CT only policy and the static policy. We find that, under time-varying arrivals, the dynamic policy achieves a larger magnitude of improvement over the CT only policy and the static policy. Hence, the adoption region is larger than that in the stationary setting. However, the magnitude of improvement is not overly large, which is consistent with our previous findings from the ANOVA in which the adoption regions were found to be relatively insensitive to the arrival rate. The detailed numerical results are relegated to Appendix D.3.

8. Conclusions

We study the bridge between medical research on diagnostic testing and the clinical workflow in which the diagnostic test will be used. The gap between the two can lead to rejection of valuable medical research by the clinical community due to increased workloads and/or a misunderstanding of the system-level benefits of the new test. We define a new operational criterion for adopting medical research in practice that accounts for doctor workload and patient delays in addition to medical outcomes, including outcomes that are not currently considered in evaluating a new test. Using an MDP to model the clinical workflow and to capture complex tradeoffs on both the individual and system levels, we identify adoption regions

that specify the combination of test characteristics that test must achieve for the research to be feasible for adoption in practice, which can provide guidance for the medical research development process. The MDP also allows us to study the impact of how the new test is used in practice on adoption decisions. To solve the MDP, we develop a decomposition heuristic and derive structural properties that shed light on when and for which patients a new diagnostic test should be used.

Via a comprehensive case study with real-world data leveraging the computationally efficient heuristic, we show that the conventional medical criterion can lead to poor decision making in both research development and clinical practice. This myopic view of medical research can lead to overvaluing or undervaluing new tests. A test with high sensitivity can be overvalued by researchers as it is found to be acceptable in the laboratory only to be rejected in practice due to operational inefficiency, whereas a test with lower sensitivity may be undervalued and rejected in the laboratory setting despite potential benefits in the clinical environment, where it can lead to a reduction in the use of expensive and potentially harmful existing tests. Moreover, the current medical criterion not only ignores the operational impact of the new test, it also surprisingly overestimates the new test's system-level misdiagnosis rate, leading to unnecessary rejection of new research. Furthermore, it fails to account for system-level implications of the new test, which can help avoid adverse effects of unnecessary testing with the existing diagnostic equipment. All these mismatches are accentuated when a simple static policy is used to integrate research into the clinical environment instead of using our more nuanced MDP policy. Similarly, the individual level misdiagnosis criterion used in practice underestimates the benefit of new tests with lower specificities, while overestimating the benefit of new tests with higher specificities, with similar consequences for adoption in practice. We find

that specificity has a much larger impact than previously understood, and that using specificity as only a secondary consideration in medical research development can lead to poor adoption decisions in practice.

Our framework provides easily interpretable guidelines for medical research development, diagnostic test design, and clinical adoption decisions in the form of adoption regions. These regions can guide medical research as to which test characteristics to focus on to improve chances of adoption in practice, bridging the gap in the current research practice that often fails to consider the broader clinical context in which the test will be used. In general, we find that increasing sensitivity supports medical value creation, whereas increasing specificity (which is not currently a primary focus of medical research) supports operational value and greater opportunities for adoption, which is especially important for slower tests. On the medical side, specificity has mixed results: trading off increased individual misdiagnosis rates with improved avoidance of an expensive and/or harmful primary test. When tests are slower, we also find that the manner of integration plays an important role, whereas a simple static policy (as in current practice) can gain much of the benefits if the new test is sufficiently fast. These findings can also guide the medical research process by focusing efforts based on specific concerns about adoption.

Note

¹The medical boundary in Equation (3) is the proportion who received a false negative result, *conditioning* on the result being negative. This is why, although point 1 is on the medical boundary for 3%, the misdiagnosis rate is not 3% since not all patients are sent to the new test and not everyone gets a negative result. The system-level misdiagnosis rate reflects the impact of routing (which patients are sent to the new test), while the medical misdiagnosis rate is unaffected by routing policy.

References

- Alizamir, S., F. De Véricourt, P. Sun. 2013. Diagnostic accuracy under congestion. *Management Sci.* 59(1): 157–171.
- Argon, N. T., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manuf. Serv. Oper. Manag.* 11(4): 674–693.
- Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stoch Syst.* 5(1): 146–194.
- Barrows, H. S., G. C. Pickell. 1991. *Developing Clinical Problem-Solving Skills: A Guide to More Effective Diagnosis and Treatment*. Norton Medical Books, New York, NY.
- Bordage, G. 1994. Elaborated knowledge: A key to successful diagnostic thinking. *Acad. Med.* 69(11): 883–885.
- Boyan, J. A., M. L. Littman. 2000. Exact solutions to time-dependent MDPs. Proceedings of the 13th International Conference on Neural Information Processing Systems. MIT Press, pp. 982–988.
- Buitenhek, R., G.-J. van Houtum, J.-K. van Ommeren. 1997. On first-come first-served versus random service discipline in multiclass closed queueing networks. *Prob. Eng. Inform. Sci.* 11(3): 313–326.
- Dai, T., S. Tayur. 2018. *Handbook of Healthcare Analytics: Theoretical Minimum for Conducting 21st Century Research on Healthcare Operations*. John Wiley & Sons, Hoboken.
- Davis, T. M. E., H. A. Karunajeewa, K. F. Ilett. 2005. Artemisinin-based combination therapies for uncomplicated malaria. *Med. J. Aust.* 182(4): 181–185.
- Efroni, Y., G. Dalal, B. Scherrer, S. Mannor. 2018a. Beyond the one step greedy approach in reinforcement learning. *arXiv preprint arXiv:1802.03654*.
- Efroni, Y., G. Dalal, B. Scherrer, S. Mannor. 2018b. Multiple-step greedy policies in approximate and online reinforcement learning. *Adv. Neural. Inf. Process. Syst.* 5238–5247.
- Fesmire, F. M., M. D. Brown, J. A. Espinosa, R. D. Shih, S. M. Silvers, S. J. Wolf, W. W. Decker. 2011. Critical issues in the evaluation and management of adult patients presenting to the emergency department with suspected pulmonary embolism. *Ann. Emerg. Med.* 57(6): 628–652.
- Guyatt, G. H., P. X. Tugwell, D. H. Feeny, R. B. Haynes, M. Drummond. 1986. A framework for clinical evaluation of diagnostic technologies. *Can. Med. Assoc. J.* 134(6): 587.
- Guyatt, G., D. Rennie, M. Meade, D. Cook, et al. 2002. *Users' Guides to the Medical Literature: A Manual for Evidence-based Clinical Practice*, Vol. 706. AMA Press, Chicago.
- Hajek, B. 1984. Optimal control of two interacting service stations. *IEEE Trans. Autom. Control*, 29(6): 491–499.
- Hasija, S., E. J. Pinker, R. A. Shumsky. 2005. Staffing and routing in a two-tier call centre. *Int. J. Oper. Res.* 1(1–2): 8–29.
- Hu, Y., C. W. Chan, J. Dong. 2019. Optimal scheduling of proactive care with patient deterioration. Technical report, Columbia University Working Paper.
- Kassirer, J. P., R. I. Kopelman, J. B. Wong. 1991. *Learning Clinical Reasoning*. Williams & Wilkins, Baltimore.
- Kline, J. A., A. E. Jones, N. I. Shapiro, J. Hernandez, M. M. Hogg, J. Troyer, R. D. Nelson. 2014. Multicenter, randomized trial of quantitative pretest probability to reduce unnecessary medical radiation exposure in emergency department patients with chest pain and dyspnea. *Circ. Cardiovasc. Imaging*, 7(1): 66–73.
- Kohn, M. A., F. A. van EsKlok. 2017. D-dimer interval likelihood ratios for pulmonary embolism. *Acad. Emerg. Med.* 24(7): 832–837.
- Koole, G. 2007. Monotonicity in Markov reward and decision chains: Theory and applications. *Found. Trends Stoch. Syst.* 1(1): 1–76.
- Leeflang, M. M. G., A. W. S. Rutjes, J. B. Reitsma, L. Hooft, P. M. M. Bossuyt. 2013. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ*, 185(11): E537–E544.
- Linden, A. 2006. Measuring diagnostic and predictive accuracy in disease management: An introduction to receiver operating characteristic (ROC) analysis. *J. Eval. Clin. Pract.* 12(2): 132–139.
- Linkins, L.-A., S. M. Bates, E. Lang, S. R. Kahn, J. D. Douketis, J. Julian, S. Parpia, P. Gross, J. I. Weitz, F. A. Spencer, et al. 2013. Selective d-dimer testing for diagnosis of a first suspected episode of deep venous thrombosis: A randomized trial. *Ann. Intern. Med.* 158(2): 93–100.
- Marecki, J., Z. Topol, M. Tambe, et al. 2006. A fast analytical algorithm for MDPs with continuous state spaces. AAMAS-06 Proceedings of 8th Workshop on Game Theoretic and Decision Theoretic Agents.

- McMahon, J. J. 2008. Time-dependence in Markovian decision processes chapter 8. Ph.D. thesis, The University of Adelaide.
- Miller, B. L. 1969. A queueing reward system with several customer classes. *Management Sci.* **16**(3): 234–245.
- Mulherin, S. A., W. C. Miller. 2002. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann. Intern. Med.* **137**(7): 598–602.
- Örmeci, E. L., E. D. Güneş, D. Kunduzcu. 2015. A modeling framework for control of preventive services. *Manuf. Serv. Oper. Manag.* **18**(2): 227–244.
- Penaloza, A., P.-M. Roy, J. Kline, F. Verschuren, G. Le Gal, S. Quentin-Georget, N. Delvau, F. Thys. 2012. Performance of age-adjusted d-dimer cut-off to rule out pulmonary embolism. *J. Thromb. Haemost.* **10**(7): 1291–1296.
- Qaseem, A., V. Snow, P. Barry, E. R. Hornbake, J. E. Rodnick, T. Tobolic, B. Ireland, J. B. Segal, E. B. Bass, K. B. Weiss, et al. 2007. Current diagnosis of venous thromboembolism in primary care: A clinical practice guideline from the American Academy of Family Physicians and the American College of Physicians. *Ann. Intern. Med.* **146**(6): 454–458.
- Segal, J. B., J. Eng, L. J. Tamariz, E. B. Bass. 2007. Review of the evidence on diagnosis of deep venous thrombosis and pulmonary embolism. *Ann. Fam. Med.* **5**(1): 63–73.
- Shumsky, R. A., E. J. Pinker. 2003. Gatekeepers and referrals in services. *Management Sci.* **49**(7): 839–856.
- Stidham, S. 1985. Optimal control of admission to a queueing system. *IEEE Trans. Autom. Control*, **30**(8): 705–713.
- Stidham Jr., S., R. Weber. 1993. A survey of Markov decision models for control of networks of queues. *Queueing Syst.* **13** (1–3): 291–314.
- Sun, Z., N. T. Argon, S. Ziya. 2017. Patient triage and prioritization under austere conditions. *Management Sci.* **64**(10): 4471–4489.
- Wang, X., L. G. Debo, A. Scheller-Wolf, S. F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Sci.* **56** (11): 1873–1890.
- Whitt, W., X. Zhang. 2017. A data-driven model of an emergency department. *Oper. Res. Health Care*, **12**: 1–15.
- Yoon, S., M. E. Lewis. 2004. Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Syst.* **47**(3): 177–199.
- Zhang, B., H. Ayhan. 2013. Optimal admission control for tandem queues with loss. *IEEE Trans. Autom. Control*, **58**(1): 163–167.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Refinement to the Basic Decomposition Algorithm.

Appendix B: Major Proofs.

Appendix C: Numerical Validation of Decomposition Algorithm.

Appendix D: Numerical Experiments in Partner ED Setting.