Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Modeling Framework for Control of Preventive Services

E. Lerzan Örmeci, Evrim Didem Güneş, Derya Kunduzcu

To cite this article:
E. Lerzan Örmeci, Evrim Didem Güneş, Derya Kunduzcu (2016) A Modeling Framework for Control of Preventive Services. Manufacturing & Service Operations Management 18(2):227-244. http://dx.doi.org/10.1287/msom.2015.0556

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 2, Spring 2016, pp. 227–244
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2015.0556

A Modeling Framework for Control of Preventive Services

E. Lerzan Örmeci
Department of Industrial Engineering, Koç University, 34450 Istanbul, Turkey, lormeci@ku.edu.tr

Evrim Didem Güneş
College of Administrative Sciences and Economics, Koç University, 34450 Istanbul, Turkey, egunes@ku.edu.tr

Derya Kunduzcu
Credit Analytics, Akbank, 34774 Istanbul, Turkey, dkunduzcu@gmail.com

We present a modeling framework for facilities that provide both screening (preventive) and diagnostic (repair) services. The facility operates in a random environment that represents the condition of the population that needs screening and diagnostic services, such as the disease prevalence level. We model the environment as a partially endogenous process: the population's health can be improved by providing screening services, which reduces future demand for diagnostic services. We use event-based dynamic programming to build a framework for modeling different kinds of these facilities. This framework contains a number of service priority policies that are concerned with prioritizing screening versus diagnostic services. The main trade-off is between serving urgent diagnostic needs and providing screening services that may decrease future diagnostic needs. Under certain conditions, this trade-off reverses the famous $c\mu$ rule; i.e., the patients with lower waiting cost are given priority over the others. We define appropriate event operators and specify the properties preserved by these operators. These characterize the structure of optimal policies for all models that can be built within this framework. A numerical study on colonoscopy services illustrates how the framework can be used to gain insights on developing good screening policies.

Keywords: healthcare management; public policy; dynamic programming; stochastic methods
History: Received: August 3, 2010; accepted: June 22, 2015. Published online in *Articles in Advance* November 4, 2015.

1. Introduction

A famous quote from Benjamin Franklin reads, "An ounce of prevention is worth a pound of cure." Although it was originally firefighting advice to Philadelphians, it applies to many situations now, especially to healthcare. As healthcare costs increase, prevention has gained more importance in the struggle to improve the health of the population and reduce healthcare costs. For example, recent healthcare policy reforms in the United States promote preventive services by removing copayments for screening services.[1] Improved consciousness increases the demand for preventive services. On the other hand, when the demand for preventive services competes with the demand for curative services for utilizing resources, it is not clear how priorities should be assigned.

Optimal scheduling models tend to schedule the more urgent patient first, i.e., the one who is most vulnerable to waiting, which in this case would be the patient demanding curative services. However, although the patients demanding preventive services do not present themselves with urgency, a reasonable service level should be provided for prevention if any benefits are to be realized. The real benefit of prevention is a decreased demand for curative services in the future, which is not considered in the context of queueing control models. These models generally focus on the delay sensitivities of the customers, and therefore they generally suggest prioritizing the urgent need. In this paper, we provide a modeling framework that accounts for the future benefits of prevention, and we analyze optimal scheduling policies within this framework. One of our main objectives is to generate insights concerning the situations when prevention, the less-urgent need, would obtain priority.

Our motivating example is colonoscopy services for colorectal cancer screening and diagnosis. Colorectal

[1] According to the Patient Protection and Affordable Care Act, signed in March 2010 by President Obama, insurers are prohibited from charging copayments or deductibles for level A or level B preventive care and medical screenings on all new insurance plans (https://www.healthcare.gov/preventive-care-benefits/, accessed October 8, 2015).

cancer is one of the most common cancers and a leading cause of cancer-related deaths in the United States (American Cancer Society 2008). The U.S. Preventive Services Task Force recommends screening for colorectal cancer in adults of ages between 50 and 75 (U.S. Preventive Services Task Force 2009). Colonoscopy has gained popularity in recent years as the most accurate test, because it allows the doctor to see the entire colon and remove polyps that may cause cancer (Levin et al. 2008, Kolata 2003). Therefore, it not only detects cancers early but also prevents future cancers. Furthermore, colonoscopy is essential for diagnosing colorectal cancer in symptomatic patients. Diagnostic colonoscopies are demanded by patients with a symptom, who may come from all age groups and who may or may not have cancer. Some diagnostic colonoscopies may also serve the purpose of screening and prevention. On the other hand, the proportion of diagnostic colonoscopies that serve for prevention depends on various factors. Therefore, relying on only diagnostic colonoscopies for prevention is not preferred, and screening campaigns call the target population to have a screening test. The aim is to serve both types of patients so that costs to the society are minimized, which may be challenging in the presence of scarce resources.

Colonoscopy is an expensive procedure; most countries have insufficient resources to screen the entire target population, and patients complain of long waiting times for colonoscopies (Kolata 2003, Donnellan 2008). On the other hand, there is no universal rule applied for prioritization of patients. Indeed, a study of colonoscopy centers in New Hampshire reports that, among the 36 centers surveyed, only one implemented a formal priority system based on the indication for colonoscopy to prioritize scheduling (Butterly et al. 2007). This situation calls for developing good screening policies and control schemes to allocate scarce colonoscopy resources.

In this paper, we present our results in the context of colonoscopy services. The decision maker in our model is a health maintenance organization that focuses on the total cost of care for the enrolled population and can recommend operational policies to the service providers in their network. Our approach can also be valid for single-payer healthcare systems, such as the national health services in the United Kingdom and in Canada. The colonoscopy services have three main characteristics that are critical for our model: (1) screening and diagnostic colonoscopies are performed by the same resources, (2) providing colonoscopies decreases the frequency of diagnostic needs, and (3) screening colonoscopy decreases the frequency of diagnostic needs more than diagnostic colonoscopy does. Hence, the main trade-off in this context is between the urgent diagnostic needs and the less-urgent screening needs, whose benefits will be realized in the future through a lower cancer rate and lower demand rate for diagnosis.

To the best of our knowledge, this paper is the first to model this endogenous relationship between screening and diagnostic services to address the problem of dynamically allocating a shared server for these two activities. In essence, our model links the short-term decision of scheduling to a long-term impact of reduced disease incidence. Using our framework, we show that this interaction may in certain settings favor serving the less-urgent screening demand rather than diagnosis.

More specifically, we build a modeling framework based on Markov decision processes (MDPs) to represent a facility consisting of a colonoscopy suite that provides screening and diagnostic services. Our construction is completed in two steps: (1) modeling the endogenous relationship between providing colonoscopies and the future arrivals of diagnostic needs and (2) constructing an event-based dynamic programming (EBDP) framework for systems employing different service priority rules, including optimal scheduling policy.

We model the facility as a single server receiving two types of patients: one is asymptomatic patients who need periodic screening, the other is symptomatic patients who have symptoms of colorectal diseases and need a diagnostic colonoscopy. Patients with diagnostic needs are more costly to the system and should be served more quickly. The facility operates in a random environment that represents the symptomatic demand rate in the population: in a better environment, the diagnostic needs occur less frequently. The distinguishing feature of our framework lies in the fact that the state of the random environment can be changed by performing screening or diagnostic colonoscopies, with different probabilities.

We then introduce a modular representation of all possible events in such a facility through EBDP with a special emphasis on the effects of operational control policies, which determine how to allocate the capacity between screening versus diagnosis. Further, we characterize the properties preserved by the events in this framework. As a result, the optimal policy structures of the models built within this framework can easily be deduced. Finally, we use the framework to numerically analyze the effects of screening in colonoscopy services, where we consider a simple system, which operates in an environment with four states.

The main contributions of the paper are summarized as follows.

1. The first contribution is in modeling. The key feature of the model is to capture the relationship between service provision and the future demand for

diagnostic needs through a partially endogenous random environment. To the best of our knowledge, our study is the first to develop an MDP model that represents this feature in a queueing control setting. A general framework is developed by defining a common state and event operators corresponding to a library of events using EBDP. Structural properties preserved by the event operators are identified.

2. Our second contribution is in implementing our framework to analyze the colonoscopy services, with parameters based on the medical literature. After a thorough numerical analysis we conclude that, practically, prioritizing demand for diagnostic services is a viable policy.

3. Finally, with further numerical experiments we explore the settings in which the reversal of the $c\mu$ rule (the well-known scheduling rule that gives priority to jobs with the highest product of holding cost (c) and service rate (μ)) performs well. This provides insights for situations in which prioritizing screening demand improves the total cost substantially.

The rest of this paper is organized as follows. We give a brief review on the related literature in §2. Section 3 presents the modeling framework. Section 4 characterizes the structure of optimal policies for models that can be generated within our framework. Section 5 discusses numerical examples based on colonoscopy services. In §6, we summarize insights and discuss possible extensions of the framework.

2. Literature Review

Our work is related to various research areas in the operations literature. Our model represents the effects of preventive maintenance (screening) on the demand of repair services (diagnosis) using a two-class queue operating in a random environment that can be improved by preventive services. The aim is to characterize the optimal scheduling policies for preventive and repair services that minimize the total cost. Accordingly, the closest research areas are cancer screening, preventive maintenance, and stochastic control.

Alagoz et al. (2011) present a review of the literature on cancer screening, including colorectal cancer. There are a number of recent studies that consider colorectal cancer screening: Erenay et al. (2014) develop a partially observable Markov decision process to optimize colonoscopy screening policies for the objective of maximizing total quality-adjusted life years. This paper does not consider the limited capacity of colonoscopy that has to be shared by demand for screening and diagnosis. However, Güneş et al. (2015) develop a compartmental modeling framework where the colonoscopy capacity is allocated between demand for screening and diagnosis. They explicitly model the relationship between colonoscopy services and the demand for diagnosis via a population dynamics approach, as opposed to our framework, which implicitly models this feedback via an intricate definition of an environment. Their approach supports tactical-level decisions of capacity allocation, whereas our approach analyzes optimal dynamic scheduling policies, which guide operational-level decisions. As a result, these two papers complement each other. We also note that in the colorectal cancer setting these papers agree in prioritizing the demand for diagnostic services at both tactical- and operational-level decisions.

There is a vast literature on preventive maintenance. Pierskalla and Voelker (1976) provide an early review of the subject, whereas Wang and Pham (2006) present a comprehensive survey of the latest theories and methods in the field. Wu and Zuo (2010) review studies that model the relationships of the hazard functions with the preventive maintenance policies. This is in parallel to our model of interaction between screening and diagnostic services. However, to the best of our knowledge, there is no study that explicitly models the capacity that can be shared by maintenance and repair services to optimize scheduling decisions.

The MDP approach, the basis of our framework, is frequently used to analyze stochastic control policies for different systems. For an overview of MDP models, see Puterman (1994). Here, we focus on MDP applications in healthcare systems. Schaefer et al. (2004) discuss modeling issues related to MDP models for medical treatment decisions with application areas such as epidemic control, transplantations, and breast cancer screening and treatment. Alagoz et al. (2007) analyze the decision process related to acceptance of cadaveric livers, whereas Shechter et al. (2008) use an MDP model to answer the question of when to initiate HIV treatment. This line of research models the progress of patient health through time to identify the best treatment policy, without a reference to capacity constraints, as opposed to our framework, which considers scheduling the demand for screening and diagnosis in a shared server.

Our research falls into the general category of queueing control via MDPs, and within this into event-based dynamic programming (EBDP). Koole (2006) and Çil et al. (2009) present EBDPs that model a variety of queueing control problems and analyze their properties. There are a number of such problems formulated for healthcare systems: Green et al. (2006) and Patrick et al. (2008) provide examples for dynamic scheduling of multipriority patients in a diagnostic medical facility. Argon and Ziya (2009) consider priority assignments in a triage system. In all these studies, the patient types are defined according to the

delay sensitivities of the patients. Our work, on the other hand, acknowledges an additional characteristic of patients through endogeneity, so that serving the less-urgent need can be more effective in decreasing the future arrival rate of the more-urgent need.

We model the interaction between colonoscopy services and the demand for diagnosis implicitly through a fluctuating environment, which improves by screening services and deteriorates over time. The use of a fluctuating environment is common in various contexts; see, e.g., Prabhu and Zhu (1989) and Gayon et al. (2009) for control of queues, Özekici and Soyer (2003) for reliability, and Ozkan et al. (2013) for revenue management. In most studies, the environment process is defined as an exogenous Markov process that modulates certain parameters of the system under consideration. In contrast, our work considers a partially endogenous environment. Hordijk and Koole (1993) are closer to our approach, because they also introduce an indirect dependence between the arrival process and the number of customers in the system. The state of the random environment in their model can be optimized, but it is affected by the decisions regarding service only through the queue lengths. In our model, the scheduling decisions influence the arrival process.

The feedback mechanism between service and demand processes can also be represented explicitly as in the literature of retrial queues, where customers who need to repeat the service join an orbit before requesting service again; see Artalejo (2010) for a comprehensive bibliography. This literature generally concentrates on performance evaluation, and there are a few studies that consider the optimal control, such as Bhulai et al. (2014) and Efrosinin and Breuer (2006). Moreover, these studies assume one type of customer, as opposed to our model with two patient types.

3. Modeling Framework

Throughout the paper, we consider a basic model that consists of a single-server facility. The facility receives symptomatic patients who need diagnostic services, as well as asymptomatic patients who request screening periodically. For screening with colonoscopy, screening is recommended every 10 years for patients of ages 50–75 if they are not considered to have a high risk of colorectal cancer. If they are in the high-risk group, then they should be screened every five years. The condition of a symptomatic patient is diagnosed through a diagnostic colonoscopy and the patient is referred to treatment, if necessary. Colonoscopy has nearly perfect specificity (Forde 2006) and so will not generate additional need for other diagnostic services including colonoscopy, unlike some less conclusive screening tests (such as the fecal occult blood test). We model only colonoscopy services, excluding all treatment activities, so that a diagnostic colonoscopy does not have a further effect on the system. The key feature of this model is representing the relationship between the colonoscopy services that take place now and the future arrivals of diagnostic needs. Both types of patients have the potential to decrease future arrivals, but the effect of screening is higher than that of the diagnostic colonoscopy. Our motivation in using such a basic model is to isolate the effects of this relationship on the service priority decisions. In the conclusion, we briefly discuss different settings that can be modeled by the framework.

Now let us describe the basic model in more detail: both patient types require a random service time, assumed to be exponential with rate μ. The duration of the colonoscopy exam depends on patient and physician characteristics as well as the number of polyps removed during the exam, rather than the purpose of the test. The number of polyps does not depend on the patient type; i.e, an asymptomatic patient may have many polyps, whereas a symptomatic patient may have few. Hence, the procedure takes 30–60 minutes, regardless of the patient type (Mayo Clinic 2011). Therefore, it is plausible to assume equal service rates for the colonoscopy for all patients.

The facility can use different priority rules when scheduling diagnosis and screening services. We consider the following rules: giving strict priority to symptomatic or asymptomatic patients; allocating a certain proportion of service capacity to each type, where these proportions may vary by the environment state; and employing an optimal scheduling policy that gives priority to one of the patient types according to the current state of the system.

A fixed service cost of s is charged per patient, regardless of the type of service. Moreover, each symptomatic patient incurs a fixed cost of t_H, which represents the average treatment cost of a symptomatic patient to the system. A diagnostic colonoscopy may reveal that a symptomatic patient has colorectal cancer, or some other colon disease such as colitis, or a polyp, or no serious problems. We can consider t_H as the expected cost of these outcomes. A similar cost is incurred for asymptomatic patients, denoted by t_L.

Finally, linear holding costs c_H and c_L are incurred per unit time for each symptomatic or asymptomatic patient in the system, respectively. These costs represent the cost of delay in terms of treatment costs. For example, if an early-stage cancer advances to late stage during the waiting period of a symptomatic patient, then the treatment cost difference between the early and late stages will be reflected in c_H. Symptomatic patients are more likely to have cancer so that delaying their treatment has more serious and more

costly consequences. Since the delay affects a symptomatic patient much more than it does an asymptomatic patient, we assume that symptomatic patients have higher waiting costs than the asymptomatic patients do; i.e., $c_H \geq c_L$.

In this paper, we calculate all the costs in terms of treatment costs. However, it is also possible to estimate them in terms of the quality-adjusted life year (QALY) value loss due to cancer. In that case, the fixed costs would correspond to the expected QALY loss of patients, whereas the waiting costs would reflect the effect of delay on QALY.

Section 3.1 explains the arrival processes for both patient types, as well as the relationship between colonoscopy services and the environment process in depth. We describe how to build a discrete-time MDP model through uniformization in §3.2. This section also presents the event operators that are necessary to represent the basic system.

3.1. Modeling Endogeneity in Colonoscopy Services

To understand the endogeneity between colonoscopy services and the future symptomatic demand, we need to consider the outcomes of a colonoscopy service. There are three possible outcomes regardless of the patient type: (1) the patient may be healthy with no polyps, (2) the patient may have polyps that are removed during the colonoscopy, or (3) the patient may have cancer or some other colon disease such as colitis that is detected during the test and the patient is referred for treatment. Outcome (1) has no consequence regarding the healthcare system. Outcome (2) can prevent future cancers and reduce the incidence rate, as it is known that more than 90% of colorectal cancers grow from polyps (Levin et al. 2008). Outcome (3) brings additional need for treatment services, but not for diagnostic colonoscopy, since colonoscopy is a conclusive test. Outcome (2) creates endogeneity between colonoscopy services and the future diagnostic demand, since serving patients with polyps will reduce the number of symptomatic patients in the future. We aim to capture the relationship between colonoscopy services and the colorectal cancer incidence rate through a subtle definition of environment. We note that, although the possible outcomes are the same for all patients, their corresponding probabilities vary significantly by patient type.

To model this endogeneity, we define an appropriate environment and set the necessary relationships between the environment process and the colonoscopy services. The colonoscopy facility operates in a random environment, which represents the population health with respect to colorectal cancer. The exact definition of population health is a matter of modeling choice and can take different forms. However, the resulting model should satisfy certain requirements, such as observability of the environment and estimatability of its parameters by the available data.

In the colorectal cancer setting there is strong evidence that supports the relationship between the percent screened and the incidence rate. Thus, we link the environment definition to the percentage of target population screened, because it satisfies all the above requirements. Each environment e is the rank order of a certain range of "percent screened," where ranges are defined mutually exclusively to cover (0%–100%). We let π_e be the average of the corresponding range. Then, the environment will take values in a finite state space $\{1, 2, \ldots, E\}$, where e will indicate that on average π_e percent of the target population is screened. We assume that $\pi_e < \pi_{e+1}$, so that more people are screened in environment $e+1$ than those in environment e.

In reality, every person screened is a step toward improving the environment; however, in our model the environment state is improved when the average number of people screened exceeds a certain number. Hence, we assume that screening an asymptomatic patient improves the environment with probability $p^L_{e,e+1}$, whereas the environment state remains the same with probability $1 - p^L_{e,e+1}$. Hence, each screening is a Bernoulli trial with probability of success $p^L_{e,e+1}$, and the expected number of trials until success is $1/p^L_{e,e+1}$, where success is defined as improving the environment. We link the expected time to improve the environment state with the environment definition as follows: we assume that a colonoscopy facility serves a target population with a given size N. If an environment e corresponds to π_e percent of the population screened, then an average of n_e patients should be screened in environment e; i.e., $n_e = N\pi_e/100$. Consequently, $\Delta_e = n_{e+1} - n_e$ screenings are needed to improve environment e to $e+1$ so that $\Delta_e = N(\pi_{e+1} - \pi_e)/100$. Thus, the model captures the long-term effect of the screening process by setting $1/p^L_{e,e+1} = \Delta_e$, which ensures that the expected number of screenings until improvement is equal to the actual number of screenings required. Note that this estimation of $p^L_{e,e+1}$ assumes perfect screening. If the sensitivity of the test were less than perfect, then more patients would have to be screened to achieve the same effect. This would effectively reduce $p^L_{e,e+1}$. If the physician assesses that the diagnostic service can be counted as screening, this service also increases the percent screened. Hence, we let $p_d = P(\text{diagnosis serves for screening})$ and set $p^H_{e,e+1} = p_d p^L_{e,e+1}$.

Consider the case when a patient who is required to be screened in seven years was screened eight years ago. Then, this patient cannot be counted as a part of the percent screened in the population. Hence, he or

she cannot be counted as a part of the percent screened in the population. Moreover, the target population for screening is ages 50–75, so that every year new patients, not yet screened, enter the target population while others, who are possibly screened, leave. The deterioration in the population's health takes place in years, so that the environment should gradually change its state. To represent this phenomenon, we assume that, unless there is a screening, the environment stays in a state e for an exponential amount of time with rate $\gamma_{e,e-1}$ before it moves to state $e-1$. Hence, the expected time for the effect of screening to dissipate in environment e is given by $1/\gamma_{e,e-1}$. We use the average interval suggested for periodic screening to estimate the parameters $\gamma_{e,e-1}$. This interval is recommended as 5–10 years for colorectal cancer screening, depending on the risk group. As a result, $\gamma_{e,e-1}$ models the factors that create long-run impacts on the process.

The demand rate for diagnostic needs depends on the health condition of the population. Hence, we assume that the diagnostic needs occur according to a Poisson process with rate $\lambda_H(e)$ when the environment is in state e. When the population is in good health, i.e., when the percent screened is high in our formulation, the arrivals for diagnosis will occur less frequently. Accordingly, we have $\lambda_H(e) < \lambda_H(e-1)$, so that environment $e-1$ is worse than environment e, in terms of the arrival rates for diagnosis. Note that parameters $\gamma_{e,e-1}$ and $p_{e,e+1}$ affect the evolution of the environment process, where $\lambda_H(e)$ is the result of the environment state.

Patients with screening requests arrive at the facility according to a Poisson process with rate λ_L. This rate represents the maximal possible screening demand generated by the target population, since we consider the case where everyone in the target population asks for screening. This serves our purpose of modeling the trade-off between screening and diagnostic needs when demand for screening is maximized. Table 1 summarizes the notation introduced in this section.

Table 1 Definition of Parameters

Parameters	Definitions
π_e	Average percent of the target population screened in environment e
$\lambda_H(e)$	Arrival rate of symptomatic patients in environment e (patients/day)
λ_L	Arrival rate of asymptomatic patients (patients/day)
μ	Service rate (patients/day)
$p^L_{e,e+1}$	Probability of improving an environment e to $e+1$ upon a screening colonoscopy, $e < E$
$p^H_{e,e+1}$	Probability of improving an environment e to $e+1$ upon a diagnostic colonoscopy, $e < E$
$\gamma_{e,e-1}$	Deterioration rate of environment e to $e-1$, $e > 1$
t_L	Cost of an asymptomatic patient ($/patient)
t_H	Cost of a symptomatic patient ($/patient)
c_H	Holding cost of a symptomatic patient ($/patient/day)
c_L	Holding cost of an asymptomatic patient ($/patient/day)
s	Service cost ($/service)

The behavior of the environment process and the endogeneity between colonoscopy services and demand for diagnosis are the key elements of our modeling framework. The colonoscopy services prevent future cancer cases by removing polyps from the colon. Therefore, colonoscopy services decrease the proportion of symptomatic individuals in the population in the future, which, in turn, decreases the disease prevalence and the arrival rate of symptomatic patients. Although both types of patients have the potential of decreasing future cancers, the overall probability that screening colonoscopy improves the population health is higher than that of diagnostic colonoscopy.

Our model does not incorporate disease progression or patient death explicitly. However, disease progression is considered indirectly. (1) Screening and diagnostic services may prevent disease progression, which is represented by their ability to decrease the arrival rate of symptomatic patients through improving the environment. (2) If a person is not screened, existing polyps may turn into cancer, which increases the long-run arrival rate of symptomatic patients. This is modeled through the environment deterioration process. (3) When a patient is delayed, the disease may progress during the wait. The waiting costs represent the cost of delay in terms of treatment costs, which accounts for the disease progression. To verify that this stylistic model is a reasonable approximation of reality, we developed a stochastic simulation model that incorporates the population dynamics explicitly through four health states (healthy, with polyp, early cancer, and late cancer) as well as two independently operating facilities (rather than one, as considered here). We assume that patients join the facility with the shortest queue; i.e., the simulation model does not account for the effects of more complex patient behavior and appointment scheduling. The simulation results confirm our recommendations concerning the scheduling policies. More information on this simulation study is available in the online appendix (available as supplemental material at http://dx.doi.org/10.1287/msom.2015.0556).

3.2. Event Operators

This section describes how to build a discrete-time equivalent of the basic model using uniformization and defines the event operators, which are the building blocks of the EBDP framework.

The state of the system is defined as (e, x_H, x_L), where e represents the state of the environment, and x_H and x_L denote the number of symptomatic

(in need of diagnosis) and asymptomatic (in need of screening) patients in the system, respectively. We use labels H and L to denote symptomatic and asymptomatic patients, associated with their high and low costs respectively. We let $\mathcal{S}$ be the state space of the system and define it as follows: $\mathcal{S} = \{(e, x_H, x_L): 1 \le e \le E, 0 \le x_H, 0 \le x_L\}$. To simplify the notation we set $\mathbf{x} = (x_H, x_L)$. Then, we let $f(e, \mathbf{x})$ be a generic function defined on the state space $\mathcal{S}$, $f: \mathcal{S} \to \mathbb{R}$.

We consider an infinite-horizon model, so that the objective can be to minimize the long-run average cost or the total expected discounted cost over an infinite horizon. The original process takes place in continuous time, where the time between two events is exponential, and the total rate out of each state is bounded by A, where we set $A = \lambda_L + \lambda_H(1) + \mu + \hat{\gamma}$ with $\lambda_H(1) = \max_e\{\lambda_H(e)\}$ and $\hat{\gamma} = \max_{2 \le e \le E}\{\gamma_{e,e-1}\}$. Therefore, the state of the original system can change in every exponential time distributed by rate A. We refer to these points as potential transition epochs. By using an appropriate time scale, we can set $A = 1$, which transforms the rate of an event J in the original process to the probability that event J occurs in the next transition epoch (Lippman 1975). At a potential transition epoch, a fictitious transition can also occur; i.e., the system state remains the same at this epoch. For example, in a state $(E, \mathbf{x})$, the actual symptomatic arrival rate is $\lambda_H(E) < \lambda_H(1)$, so that with probability $\lambda_H(1) - \lambda_H(E)$ there will not be an arrival and the system state will not change. From now on, we will refer to the rates in the original process as probabilities. We note that the uniformization method is necessary in EBDP frameworks; see e.g., Koole (2006) and Çil et al. (2009). The operator for uniformization is given as follows.

Uniformization (T_{UNIF}). We define the operator, T_{UNIF}, to form the uniformized MDP models:

$$T_{UNIF}(\{f_j(e, \mathbf{x})\}; \{p_j\}) = \sum_j p_j f_j(e, \mathbf{x}),$$

where p_j is the probability of observing an event j, and f_j represents the consequence of event j. We can model discounting criterion by satisfying the condition $\sum_j p_j < 1$, and long-run average criterion by $\sum_j p_j = 1$.

3.2.1. Operators for the Service Control. The following operator corresponds to serving type K patients, where $x_K > 0$ and $K \in \{H, L\}$:

$$S_K f(e, \mathbf{x}) = p^K_{e,e+1} f(e+1, \mathbf{x} - \mathbf{1}_K) + (1 - p^K_{e,e+1}) f(e, \mathbf{x} - \mathbf{1}_K) + s, \quad (1)$$

where $\mathbf{1}_K$ is the unit vector, which has a 1 in its Kth component and 0 elsewhere, and $p^K_{E,E+1} = 0$. When a type K patient is served, the environment state improves from e to $e+1$ with probability $p^K_{e,e+1}$ and remains the same with probability $1 - p^K_{e,e+1}$. This function will be used to define the service operators, which represent the priority rules we described in the beginning of this section.

Recall the features of the service process: A service is completed in an exponentially distributed time with rate μ. Hence, at a transition epoch of the uniformized MDP model, a service completion is observed with probability μ. If there are no customers in the system, the service completion is fictitious so that it does not change the state of the system. Now we define the service operators.

Strict service priority to type K patients (T_K). This operator models the service discipline that always gives priority to type K patients, so that the other type of patients, denoted by $\bar{K}$, receive service only when $x_K = 0$. For $K \in \{H, L\}$, the operator, T_K, is defined as

$$T_K f(e, \mathbf{x}) = \begin{cases} S_K f(e, \mathbf{x}), & x_K > 0, \\ S_{\bar{K}} f(e, \mathbf{x}), & x_K = 0, \quad x_{\bar{K}} > 0, \\ f(e, \mathbf{x}), & x_K = 0, \quad x_{\bar{K}} = 0. \end{cases} \quad (2)$$

The first case corresponds to serving a type K patient, the second case represents serving a type $\bar{K}$ patient, and the third corresponds to a fictitious transition. This operator corresponds to the well-known $c\mu$ rule in scheduling when $K = H$, since $c_H \ge c_L$.

Dedicated screening capacity (T_{DED}). The operator T_{DED} models the service process of a system that allocates certain proportions of the service capacity to screening and diagnosis. When the system has both symptomatic and asymptomatic patients, i.e., when $x_H > 0$ and $x_L > 0$, $\phi(e)$ of the total service capacity is allocated for symptomatic patients, while the rest, $1 - \phi(e)$, is used for asymptomatic patients. If $x_H = 0$ ($x_L = 0$), then the whole service capacity is devoted to asymptomatic (symptomatic) patients. Now, we define T_{DED} for $0 \le \phi(e) \le 1$:

$$T_{DED} f(e, \mathbf{x}) = \phi(e) T_H f(e, \mathbf{x}) + (1 - \phi(e)) T_L f(e, \mathbf{x}),$$

where T_H and T_L are defined by Equation (2). Setting $\phi(e)$ constant in all environments models systems that always allocate a fixed proportion of their capacity to each type. We note that the proportions dedicated to symptomatic patients, $\phi(e)$, can be optimized through a line search.

Scheduling control (T_{SCH}). The scheduling operator, T_{SCH}, models the optimal choice of whether to serve an asymptomatic patient or a symptomatic patient in each state:

$$T_{SCH} f(e, \mathbf{x}) = \begin{cases} \min\{S_H f(e, \mathbf{x}), S_L f(e, \mathbf{x})\}, & x_H > 0, \quad x_L > 0, \\ f(e, \mathbf{x}), & x_H = 0, \quad x_L = 0, \\ S_K f(e, \mathbf{x}) & x_K > 0 \text{ and } x_{\bar{K}} = 0. \end{cases}$$

This operator dynamically compares serving a symptomatic patient with an asymptomatic patient.

3.2.2. Operators for the Exogenous Events. In this section we define the operators that model the events that are beyond the control of the decision maker.

Environment deterioration (T_{DET}). We define the deterioration operator, T_{DET}, as follows:

$$T_{DET}f(e,\mathbf{x}) = \beta(e)f(e-1,\mathbf{x}) + (1-\beta(e))f(e,\mathbf{x}),$$

so that with probability $\beta(e)$ the environment state deteriorates from e to $e-1$, while the environment state remains the same with probability $1-\beta(e)$, corresponding to a fictitious event.

The appropriate value of $\beta(e)$ can be found for the setting described in §3.1 as follows: at each transition epoch, the system observes a potential environment deterioration with probability $\hat{\gamma}$, because we have $\hat{\gamma} = \max_{2\le e\le E}\{\gamma_{e,e-1}\}$ in uniformization. Consequently, the environment changes its state from e to $e-1$ with probability $\gamma_{e,e-1}/\hat{\gamma}$, i.e., $\beta(e) = \gamma_{e,e-1}/\hat{\gamma}$, and the environment does not change its state with probability $1-\beta(e)$, which corresponds to a fictitious event.

Arrival of symptomatic patients (T_{ARR_H}). The arrival operator, T_{ARR_H}, is defined as follows:

$$T_{ARR_H}f(e,\mathbf{x}) = \psi(e)(f(e,\mathbf{x}+\mathbf{1}_H)+t_H)+(1-\psi(e))f(e,\mathbf{x}),$$

where a symptomatic patient arrives at the system with probability $\psi(e)$, which increases x_H by 1 and incurs a fixed cost of t_H, and the system remains in the same state with probability $1-\psi(e)$. We assume that $\psi(e) < \psi(e-1)$ for all $e \ge 1$.

We compute the appropriate values of $\psi(e)$ similarly to $\beta(e)$ in T_{DET}: the probability of a potential symptomatic patient arrival at a transition epoch is $\lambda_H(1) = \max_{1\le e\le E}\{\lambda(e)\}$ due to uniformization. Then, a symptomatic patient arrives in environment e with probability $\lambda_H(e)/\lambda_H(1)$ so that $\psi(e) = \lambda_H(e)/\lambda_H(1)$, whereas a fictitious arrival occurs with probability $1-\lambda_H(e)/\lambda_H(1)$.

Arrival of asymptomatic patients (T_{ARR_L}). We define the arrival operator, T_{ARR_L}, as follows:

$$T_{ARR_L}f(e,\mathbf{x}) = f(e,\mathbf{x}+\mathbf{1}_L)+t_L,$$

where the arrival of an asymptomatic patient increases x_L by 1 and incurs a fixed cost t_L. Note that an asymptomatic patient arrives at the system with probability λ_L at a transition epoch.

Holding costs (c_H *and* c_L). The cost operator incurs the holding costs c_H and c_L:

$$T_{COST}f(e,\mathbf{x}) = f(e,\mathbf{x}) + c_H x_H + c_L x_L.$$

To illustrate the EBDP framework, we use the basic model, which employs optimal scheduling policy with the long-run average criterion. Letting g be the long-run average cost and $v(e,\mathbf{x})$ be the relative value function for a system starting in state $(e,\mathbf{x})$, we have the Bellman equations as

$$v(e,\mathbf{x}) = g + T_{COST}(T_{UNIF}(\{T_{ARR_H}v, T_{ARR_L}v, T_{SCH}v, T_{DET}v\}; \{\lambda_H(1), \lambda_L, \mu, \hat{\gamma}\})),$$

where the state $(e,\mathbf{x})$ is suppressed on the right-hand side. Here, $\lambda_H(1)+\lambda_L+\mu+\hat{\gamma}=1$ by uniformization. The fictitious transitions occur in operators T_{ARR_H}, T_{SCH}, and T_{DET}, with probabilities $\lambda_H(1)-\lambda_H(e)$, μ when the system is empty and 0 otherwise, $\hat{\gamma}-\gamma(e)$, respectively. The transitions regarding the environment changes model the long-term effects, whereas the symptomatic and asymptomatic patient arrivals, as well as the service completions, represent the short-term effects.

Note that event operators to represent admission control of asymptomatic patients and nonpreemptive service can be defined within this framework. The structural properties shown in §4 hold for those operators as well; see Kunduzcu (2009).

4. Characterizing Optimal Scheduling Policies

This section investigates how the optimal service priorities change with respect to the state variables and the parameters. For this purpose, we analyze the underlying value functions in §4.1 and then the structure of optimal scheduling policies in §4.2.

4.1. Monotonicity of the Value Functions

We let $v(e,\mathbf{x})$ denote a typical value function of an MDP model within the framework. This section derives certain monotonicity properties of $v(e,\mathbf{x})$ in three groups: (I) the monotonicity of the value functions in the state variables (x_H, x_L, e), (II) the effect of changing a symptomatic patient to an asymptomatic patient on the value functions, and (III) the monotonicity of the value functions in the parameters (the service rate μ and the probability of improving an environment state $\hat{e}$ to $\hat{e}+1$, $p^K_{\hat{e},\hat{e}+1}$ for $K = H, L$). Table 2 presents the definitions of these properties.

Properties in group II are directly linked to the scheduling decisions, which compare the outcomes of scheduling screening and diagnostic services. For this purpose, we consider two systems: scheduling a screening service or a diagnostic service. Then, we couple these two systems so that all the departure and arrival times are the same in both systems. Moreover, the environment state upon service completion of the scheduled service is determined as follows: both services improve the environment with probability $p^H_{e,e+1}$, only screening improves the environment with probability $p^L_{e,e+1} - p^H_{e,e+1}$, or finally none of the services

Table 2 Definitions of Properties Used to Analyze Optimal Policy Structures

Group	Property	Definition
(I)	$Inc(x_H)$	$v(e, x_H, x_L) \le v(e, x_H+1, x_L)$
	$Inc(x_L)$	$v(e, x_H, x_L) \le v(e, x_H, x_L+1)$
	$Dec(e)$	$v(e, x_H, x_L) \le v(e-1, x_H, x_L)$
(II)	$Diag(\mathbf{x})$	$v(e, x_H, x_L+1) \le v(e, x_H+1, x_L)$
	$IDiag_e(\mathbf{x})$	$v(e+1, x_H+1, x_L) \le v(e, x_H, x_L+1)$
(III)	$Dec(\mu)$	$v_{\mu+\varepsilon}(e, x_H, x_L) \le v_\mu(e, x_H, x_L)$, for any $\varepsilon > 0$, and v_μ is the value function of the model with μ.
	$Dec(K, p, \hat{e})$ $(K = H, L)$	$v^K_{p+\varepsilon, \hat{e}}(e, x_H, x_L) \le v^K_{p, \hat{e}}(e, x_H, x_L)$, for any $\varepsilon > 0$, and $v^K_{p, \hat{e}}$ is the value function of the model with $p^K_{\hat{e}, \hat{e}+1} = p$.

improves the environment with probability $1 - p^L_{e, e+1}$. Property $Diag(\mathbf{x})$ reflects the trade-off between providing the two services when the environment results in the same state upon the completion of both services, whereas property $IDiag_e(\mathbf{x})$ corresponds to the same trade-off when the resulting environments are different.

To establish these properties for the general framework, we identify the properties preserved by the operators defined in §3.2, and the conditions to guarantee these results, if any. More explicitly, let $f(e, \mathbf{x})$ be a function with a given property, *PropA*. If $Tf(e, \mathbf{x})$ also has *PropA*, we say that operator T preserves *PropA*. MDP models can generally be solved by using a value iteration algorithm, which is based on induction. Hence, if an MDP model is constructed by a number of operators that preserve the same property, then the value function corresponding to the MDP model will also have that property by induction.

Table 3 presents our results, whereas the corresponding proofs are placed in the online appendix. Ticks in Table 3 imply that the event operator preserves the corresponding property. Whenever additional conditions are required, these are indicated by superscripts. For example, T_{SCH} preserves $Diag(\mathbf{x})$, $IDiag_e(\mathbf{x})$, and $Dec(K, p, \hat{e})$, if function f satisfies property $Dec(e)$. Since T_{SCH} preserves $Dec(e)$, this additional condition is satisfied so that T_{SCH} preserves all of these properties. The following proposition summarizes our conclusions.

Proposition 1. (i) *Operator T_{UNIF} preserves a property if all functions constituting T_{UNIF} preserve this property.*

(ii) *All properties given in Table 2, except for property $IDiag_e(\mathbf{x})$, are preserved by all operators.*

(iii) *Property $IDiag_e(\mathbf{x})$ is preserved by all operators, except for T_{DET} and T_{COST}.*

(iv) *Operators T_{DET} and T_{COST} preserve property $IDiag_e(\mathbf{x})$ under certain conditions given in Table 3.*

Property $IDiag_e(\mathbf{x})$ compares the effect of improving the environment by screening an asymptomatic patient with the effect of reducing waiting costs by prioritizing a symptomatic patient. In the context of colonoscopy, when property $IDiag_e(\mathbf{x})$ holds, having a higher percent screened, which reduces the symptomatic arrival rate, is better than changing a symptomatic patient to a less costly asymptomatic patient. As pointed out in Proposition 1, all operators preserve property $IDiag_e(\mathbf{x})$, except for T_{DET} and T_{COST}.

Let us first consider systems that do not include operators T_{DET} and T_{COST}. When T_{COST} is excluded, the system does not incur holding costs. Then, there is no need to prioritize one patient class over another, so property $IDiag_e(\mathbf{x})$ holds. If T_{DET} is excluded, the underlying system has no deterioration, so that the effect of screening on the population's health is permanent. Vaccination of a group of people in a certain region for a specific disease within a given time period can be a typical example. In this case, property $IDiag_e(\mathbf{x})$ also holds, since screening has a great impact on the system. In the context of colonoscopy, it is not possible to exclude these operators, since we know that the waiting costs of the two patient classes are significantly different, and a person may develop colorectal cancer in 10 years after a screening with no trace of cancer.

We consider the condition required for T_{COST} to preserve property $IDiag_e(\mathbf{x})$, i.e., $c_H - c_L \le t_H(\lambda_H(e-1) - \lambda_H(e))$. The right-hand side of the inequality is the

Table 3 Properties Preserved by the Operators When the Function $f(e, \mathbf{x})$ Has the Corresponding Property

Properties	Operators								Conditions
	T_{ARR_H}	T_{ARR_L}	T_H	T_L	T_{DED}	T_{SCH}	T_{COST}	T_{DET}	
$Inc(x_H)$	✓	✓	✓	✓	✓	✓	✓	✓	$^{\dagger}\gamma_{e,e-1} \le \lambda_H(e-1) - \lambda_H(e)$
$Inc(x_L)$	✓	✓	✓	✓	✓	✓	✓	✓	$^{\bullet}c_H - c_L \le (\lambda_H(e-1) - \lambda_H(e))t_H$
$Dec(e)$	✓	✓	✓	✓	✓	✓	✓	✓	$^{\times}f(e, \mathbf{x})$: $Dec(e)$
$Diag(\mathbf{x})$	✓	✓	✓	✓	✓	✓$^{\times}$	✓	✓	$^{+}f(e, \mathbf{x})$: $Inc(x_L)$
$IDiag_e(\mathbf{x})$	✓$^{\diamond}$	✓	✓	✓	✓	✓$^{\times}$	✓$^{\bullet}$	✓$^{\dagger+\times\diamond}$	$^{\diamond}f(e, \mathbf{x})$: $Inc(x_H)$
$Dec(K, p, \hat{e})$	✓	✓	✓	✓	✓	✓$^{\times}$	✓	✓	
$Dec(\mu)$	✓	✓	✓$^{\times+\diamond}$	✓$^{\times+\diamond}$	✓$^{\times+\diamond}$	✓$^{\times+\diamond}$	✓	✓	

additional expected cost of symptomatic patients per day when the environment changes from e to $e-1$, with t_H being the expected cost of one symptomatic patient. When a symptomatic patient has cancer, the corresponding treatment cost is very high, which causes t_H to be high when compared to the holding costs c_H and c_L. Hence, $t_H(\lambda_H(e-1)-\lambda_H(e))$ is expected to exceed the difference in the holding costs in cancer-related settings. We note that property $IDiag_e(\mathbf{x})$ always holds in systems incurring the same holding cost for all patients, i.e., when $c_H = c_L$.

For T_{DET} to preserve property $IDiag_e(\mathbf{x})$, the system parameters need to satisfy $\gamma_{e,e-1} \le \lambda_H(e-1)-\lambda_H(e)$. The rate at which the environment worsens by moving from state e to $e-1$, $\gamma_{e,e-1}$, models the dissipating effects of screening, a very slow process when compared with the operational-level processes. As we discussed in §3.1, the natural unit of measurement for $1/\gamma_{e,e-1}$ is years, whereas it is the number of patients per day for $\lambda_H(e-1)-\lambda_H(e)$. Even when $\lambda_H(e-1)-\lambda_H(e)$ is small, we do not generally expect it to be as small as $\gamma_{e,e-1}$. As a result, property $IDiag_e(\mathbf{x})$ requires certain conditions on system parameters, and these conditions are expected to hold in many practical situations. We note that the parameters estimated for colonoscopy services (§5.1) satisfy the conditions for $IDiag_e(\mathbf{x})$ property.

4.2. Structure of Optimal Scheduling Policies

In §3.2, we introduced the following service priority rules by defining the corresponding operators: strict priority to symptomatic patients (the well-known $c\mu$ rule in scheduling), strict priority to asymptomatic patients (the opposite of the $c\mu$ rule), dedicating a portion of capacity to asymptomatic patients, and optimal scheduling, which dynamically assigns priorities. In this section, we identify the conditions that ensure that these strict priorities are optimal. Then, we illustrate the structure of optimal scheduling policies when these conditions do not hold.

The main trade-off stems from the advantages of serving the two service types: when a symptomatic patient is scheduled, the future holding costs will be decreased, whereas scheduling an asymptomatic patient may decrease the future arrival rate of symptomatic patients at the expense of higher holding costs. When the system is in the best environment, environment E, there is no possibility of improving the environment. Hence, scheduling asymptomatic patients does not bring any benefit. Consequently, it is optimal to prioritize symptomatic patients in environment E, which coincides with the $c\mu$ rule, as shown in Proposition 2.

Proposition 2. *In the best environment, E, symptomatic patients are always scheduled first.*

Proof. In state E, T_{SCH} compares the two quantities $v(E, x_H-1, x_L)$ and $v(E, x_H, x_L-1)$. However, $Diag(\mathbf{x})$ guarantees $v(E, x_H-1, x_L) \le v(E, x_H, x_L-1)$. By Table 3, property $Diag(\mathbf{x})$ is preserved by all operators. Hence, it is optimal to prioritize symptomatic patients in environment E. □

For colonoscopy services, this result implies that, if the percentage of the target population screened is high, then it is better to prioritize patients waiting for diagnostic colonoscopy.

This result characterizes the behavior of optimal scheduling policies in environment E: operator T_{SCH} will always schedule symptomatic patients in environment E, and in operator T_{DED}, the optimal proportion dedicated to symptomatic patients in environment E, say $\phi^*(E)$, is 1.

Now we focus on systems for which the opposite of the $c\mu$ rule is optimal in an environment e. This rule can be optimal only if there is room for improvement in the environment state, implying $e < E$. By definition, we have $p^H_{e,e+1} \le p^L_{e,e+1}$, so that the probability of improving the environment is higher if an asymptomatic patient is served. In the next proposition, we consider the extreme case where $p^L_{e,e+1} = 1$ and $p^H_{e,e+1} = 0$, and we identify sufficient conditions to observe the optimality of the opposite $c\mu$ rule. Note that $e < E$ since $p^L_{e,e+1} > 0$.

Proposition 3. *Assume that $p^L_{e,e+1} = 1$ and $p^H_{e,e+1} = 0$ for an environment e and that the following condition is satisfied*:

$$\max\left\{\frac{c_H - c_L}{t_H}, \gamma_{e,e-1}\right\} \le \lambda_H(e-1)-\lambda_H(e) \quad \text{for all } e \in \{1,\dots,E\}. \tag{3}$$

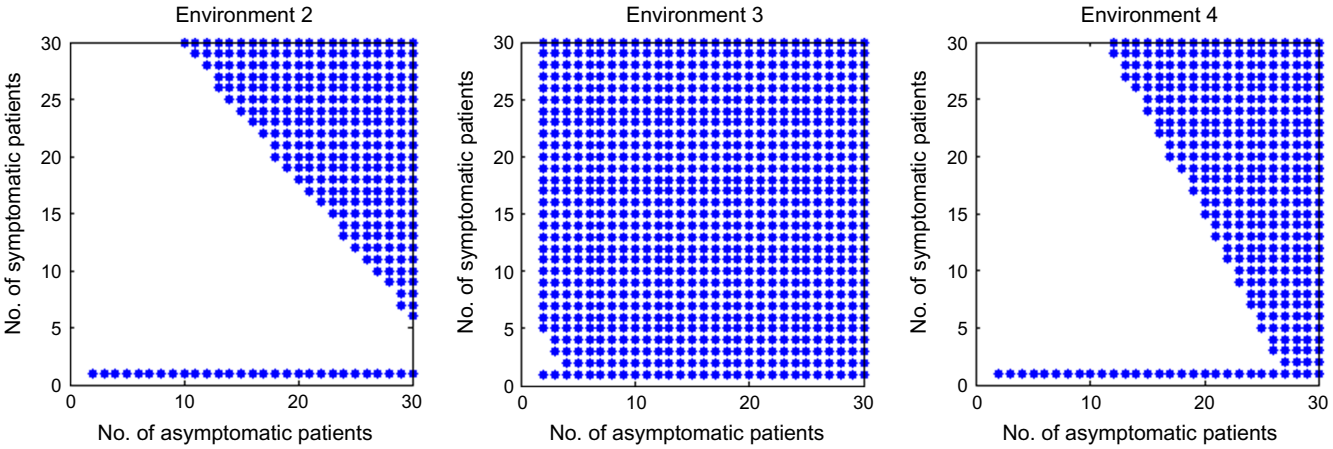
Then, asymptomatic patients are scheduled first in environment e.

Proof. Assume that condition (3) holds. Consequently, all operators preserve $IDiag_e(\mathbf{x})$ property (see Table 3). Let $p^L_{e,e+1} = 1$ and $p^H_{e,e+1} = 0$ for an environment e. Then, it is optimal to schedule an asymptomatic patient if $v(e+1, x_H, x_L-1) \le v(e, x_H-1, x_L)$, which is true by $IDiag_e$ property. □

This proposition applies to situations in which the preventive power of screening is at the highest (i.e., screening definitely improves the environment, $p^L_{e,e+1} = 1$) while the preventive power of diagnostic colonoscopy is at the lowest (i.e., diagnostic colonoscopy does not improve the environment, $p^H_{e,e+1} = 0$). In such cases, when the parameters satisfy the condition given in Equation (3), screening services obtain strict priority.

Propositions 1 and 2 characterize the scheduling policy when it gives strict priorities. However, optimal scheduling policies can also be dynamic, as shown in the next example.

Figure 1 (Color online) Optimal Scheduling Policies in Environments 2, 3, and 4 in Example 1, Which Show a Dynamic Structure



Notes. • indicates scheduling an asymptomatic patient. Parameters: $\lambda_L = 0.86, \lambda_H(1) = 2.13, \lambda_H(2) = 1.63, \lambda_H(3) = 1.13$, $\lambda_H(4) = 0.63$, $\mu = 1.8$, $p^L_{e,e+1} = 0.005$ and $p^H_{e,e+1} = 0$ for all $e < E$, $\gamma_{e,e-1} = 0.00055$ for all $e > 1$, $t_H = 5{,}000$ $t_L = 1{,}430$, $c_H = 5.2$, $c_L = 1.5$, $s = 0$.

Example 1. This example considers a system with four environments, so that $\pi_1 = 12.5\%$, $\pi_2 = 37.5\%$, $\pi_3 = 62.5\%$, and $\pi_4 = 87.5\%$. Figure 1 presents the optimal scheduling policy, as well as the parameters of the example. We observe that the optimal scheduling policy has a dynamic structure in environments $e \geq 2$.

The scheduling policy in Figure 1 can be identified by thresholds, $l(e, x_H)$: it is optimal to prioritize asymptomatic patients in a state $(e, \mathbf{x})$ when $x_L > l(e, x_H)$. These thresholds are not necessarily monotonic in e, as seen in Figure 1. Moreover, the scheduling policy in our framework changes the environment state as well as the number of patients in the system. Therefore, we could not establish the existence of an optimal threshold policy within this framework.

The value functions decrease in μ and $p^K_{\hat{e},\hat{e}+1}$ because of properties $Dec(\mu)$ and $Dec(K, p, \hat{e})$, since both properties are preserved by all the operators. We will investigate how the structures of optimal scheduling policies change with respect to parameters μ and $p^K_{e,e+1}$ numerically in §5.2.

To summarize, in the best environment, when demand for diagnostic services is at the lowest possible level, diagnosis should always obtain priority in accordance with the $c\mu$ rule. If there is room for improvement by prevention, represented by $e < E$, then the incentive to improve the environment may take over, which may violate the $c\mu$ rule in two ways: (I) the exact opposite of the $c\mu$ rule holds, so that it is always optimal to prioritize preventive services; and (II) dynamic scheduling is optimal, so that screening and diagnostic needs are prioritized depending on the actual state of the system. We prove that case (I) is true when screening services definitely improve the environment ($p^L_{e,e+1} = 1$), and diagnostic services do not improve the environment ($p^H_{e,e+1} = 0$), under the condition given by (3) in Proposition 3.

5. Implementation of the Modeling Framework

In this section we demonstrate the use of our framework for the case of colonoscopy services. The objective is to minimize the long-run average cost. We focus on the performance of the optimal scheduling policy and the $c\mu$ rule, which strictly prioritizes symptomatic demand (represented by operator T_H). The $c\mu$ rule represents an intuitive and practical solution to the scheduling problem. We also analyze the effects of parameters on these policies.

We present the parameter estimation of the model for the colonoscopy services in §5.1. Section 5.2 analyzes how the policies and their performances change with respect to system parameters. In §5.3, we consider a hypothetical system in which the optimal scheduling policy significantly improves the performance of the $c\mu$ rule.

5.1. Parameter Estimation for Colonoscopy Scheduling

This section illustrates how parameters of a real case can be estimated to use our framework. We estimate the corresponding parameters from published data and medical literature whenever possible and make reasonable assumptions (accompanied with a sensitivity analysis) otherwise. A summary of the parameter values is given in Table 4, and the references used in estimating these parameters are presented in the online appendix.

Environment. We represent the state with a random environment, which is linked to the percent of target population screened, and an environment e can be

Table 4 Parameter Values for the Base Case Example

Parameters	Values	References
π_e	$\pi_1 = 12.5,\ \pi_2 = 37.5,$ $\pi_3 = 62.5,\ \pi_4 = 87.5,$	By definition of environments
$\lambda_H(e)$	$\lambda_H(1) = 1.089,\ \lambda_H(2) = 0.936,$ $\lambda_H(3) = 0.784,\ \lambda_H(4) = 0.631$	Butterly et al. (2007), Lieberman et al. (2005)
λ_L	$\lambda_L = 1.72$	Butterly et al. (2007)
μ	$\mu = 1.73, 2.4$	Butterly et al. (2007)
$p^L_{e,e+1}$	$p^L_{e,e+1} = p = 0.00125$ for all e	Butterly et al. (2007)
p_d	$p_d = 0.33$	Butterly et al. (2007), Lieberman et al. (2005)
$\gamma_{e,e-1}$	$\gamma_{e,e-1} = \gamma = 0.00055$ for all e	Butterly et al. (2007)
t_H	$t_H = 5,000$	Vijan et al. (2001), de Bosset et al. (2002)
t_L	$t_L = 1,430$	Vijan et al. (2001), Regula et al. (2006)
c_H	$c_H = 3.2$	Vijan et al. (2001), de Bosset et al. (2002), Gopalappa et al. (2011)
c_L	$c_L = 1.5$	Vijan et al. (2001), Gopalappa et al. (2011), Brenner et al. (2007)
s	$s = 0$	Assumption
α	$\alpha = 0.70$	Winawer et al. (1993)

improved to $e+1$ by screening an additional 25% of the target population. The ranges defining the environments are as follows: 1, [0%–25%]; 2, (25%–50%]; 3, (50%–75%]; and 4, (75%–100%].

Impact of the Environment on Demand Rate. As discussed in §3.1, there is strong evidence that colonoscopy screening reduces the cancer rate in the population. The literature reports a range of values for the effect of percent screened, such as a reduction in incidence of 76%–90% by 100% screening (Winawer et al. 1993). We set a 70% reduction in the cancer rate by 100% screening as a reasonable approximation. This reflects the effect of screening on reducing the cancer rate, averaged over all risk groups in the population, which we denote as $\alpha = 0.70$. We assume that the reduction depends linearly on the screening rate; i.e., screening 1% of the population decreases the demand from symptomatic patients by 0.70%. We have also checked the impact of this assumption through a sensitivity analysis, where $\alpha = 0.80, 0.95$. As expected, average system costs decrease and the benefit of optimal scheduling increases when α increases.

Transitions Between Environments. We base our estimations on population size, demand, and capacity on a study of New Hampshire by Butterly et al. (2007). A person should be screened on average every 7.4 years. Hence, we assume that the expected duration in an environment e ($e \geq 2$) is 7.4 years, and we set a deterioration rate of environment as $\gamma_{e,e-1} = 1/(7.4 \times 250) = 0.00055$ per day, independent of the environment.

The total population size, which plays a role in calculating the probability of improving the environment, is found as $N = 3,194$. The environment is defined symmetrically so that the number of screenings to be completed to improve an environment e, Δ_e, is the same for all environments $e < 4$. This leads to equal $p^L_{e,e+1}$ since we define $p^L_{e,e+1} = 1/\Delta_e$ in §3.1. Hence, we set $\Delta = \Delta_e$ and $p = p^L_{e,e+1}$ for all $e < 4$. Then, Δ is found as $3,194 \times (0.25) = 798$, since $\pi_{e+1} - \pi_e = 25$ and $\Delta_e = N(\pi_{e+1} - \pi_e)/100$. Thus, we set $p = 1/798 = 0.00125$.

The probability that environment e is improved by a diagnostic colonoscopy, $p^H_{e,e+1}$, is estimated through p_d, which is defined as the probability that a diagnostic colonoscopy serves the purpose of screening. It is not easy to estimate this parameter. Butterly et al. (2007) also comment on "the unknown number of non-screening (symptom-related diagnostic and therapeutic) colonoscopies that contribute to meeting the demand for screening" (p. 29). We assume $p_d = 0.33$ and perform a sensitivity analysis to discuss its effects.

Service Capacity and Demand. According to the New Hampshire data, in year 2002, a single endoscopist could perform 1.73 colonoscopies per day on the average. Hence, we set $\mu = 1.73$ patients per day. Note that the assumption in this paper is that there is 100% compliance. In our numerical experiments, we let μ change in the range $(1.73, 2.4)$. For $\mu = 1.73$, the service capacity is not sufficient for the expected total demand under any of the priority policies, although it is sufficient for symptomatic demand in all environments. When capacity is increased to $\mu = 2.4$, the expected total demand can be satisfied under all priority policies. The expected number of screening colonoscopies to be demanded daily is calculated as $\lambda_L = 1.72$ patients per day.

Using the information from Butterly et al. (2007), we estimate the demand for diagnostic colonoscopy as 0.86 patients per day. However, a portion of this demand cannot be influenced by screening. Therefore, we partition the symptomatic patients so that the impact of screening can be reflected on the demand. As a result, the number of patients who have symptoms that could have been prevented by screening is estimated as 0.57 patients per day.

The arrival rate of symptomatic patients in our model, $\lambda_H(e)$, is a function of the environment. Our estimation of 0.57 patients per day reflects the demand when 50% of the target population is screened, as estimated by Butterly et al. (2007). We then calculate the demand for different environments assuming that screening 1% of the population will decrease the demand from symptomatic patients by 0.7%, and that the arrival rate is 0.57 patients per day when 50% of target population is screened. The resulting arrival rates for each environment can be seen in Table 4. Note that we do not model different risk groups and assume that the screening demand is homogenous. This implies that each screening colonoscopy has the same polyp detection rate and hence the same expected preventive effect. Thus, linearly decreasing demand rates is a reasonable assumption for this setting. If high-risk patients were to be served first, then we would expect a diminishing return to increasing the screening rates.

Treatment and Service Costs. We estimate the cost of a symptomatic patient as $t_H = \$5{,}000$ per patient. The cost of an asymptomatic patient is found as $t_L = \$1{,}430$ per patient. We assume the colonoscopy cost to be the same for both types of demand, so we set $s = 0$.

Waiting Costs. The waiting costs for both patient types, c_L and c_H, are estimated by calculating the expected negative consequences of waiting in terms of the health state and treatment cost. The waiting incurs three cost components: (1) additional treatment cost when an early-stage cancer progresses to late stage, (2) additional treatment cost when a polyp that can turn into cancer (can be named as endoscopic lesion or adenoma or neoplasia) progresses to early-stage cancer, and (3) cost of anxiety or loss of goodwill. We estimate the holding costs as the sum of the expected cost of these three components to find $c_L = 1.5$ and $c_H = 3.2$.

5.2. Implementation for Colonoscopy Scheduling

We define the base case as the colonoscopy example with the parameters estimated in §5.1. In this section, we discuss the structure of optimal scheduling policies for the base case. In addition, we compare the performance of optimal policy with the $c\mu$ rule. Finally, sensitivity of the optimal scheduling policy and the improvement in performance over the $c\mu$ rule to parameters is explored.

5.2.1. Optimal Scheduling Policy. For this implementation we focus on two service capacity values: $\mu = 1.73$ and $\mu = 2.4$. For the case of $\mu = 1.73$, the optimal scheduling policy strictly prioritizes asymptomatic patients in environments 1 and 2 (the worst environments), and strictly prioritizes symptomatic patients in environments 3 and 4. That is, a complete reversal of the $c\mu$ rule is observed in $e \le 2$. This suggests that, until a relatively good environment is reached (i.e., at least 50% of the target population is screened), it is optimal to prioritize screening colonoscopies. We should note, however, that with the optimal scheduling policy, the system would be in environments 1 or 2 only 8% of the time (see Table 5 for the environment distribution). This implies that, practically, diagnosis would obtain priority most of the time. As a result, optimal scheduling policy can provide only 0.33% improvement in total cost over the $c\mu$ rule, which always prioritizes diagnosis. Indeed, our implementation of the two policies in the simulation model did not show any significant difference in their performance.

For $\mu = 1.73$, approximately 92% of the time, the system is in environments 3 and 4 and diagnosis obtains priority. In these environments, the demand rate from symptomatic patients is low (with a maximum of 0.784 in environment 3), and therefore, for more than 50% of the time, there are no symptomatic patients in the system. Hence, more than 50% of the capacity is utilized for screening patients. This shows that, even with the $c\mu$ rule, a significant portion of the screening patients are served. Note that the data from New Hampshire stated that approximately 50% of the capacity was used for screening, which suggests that our model is a reasonably good representation of reality.

If the capacity is increased to $\mu = 2.4$, then screening obtains priority only in a few states in environment 1, and optimal policy cannot improve the $c\mu$ rule. One can conclude that, in an environment with sufficient capacity for the total demand, the best policy is to prioritize diagnosis and serve the screening demand whenever the colonoscopy suite is idle.

In conclusion, we recommend the $c\mu$ rule as a scheduling policy for colonoscopy services. This was also verified by the simulation model.

5.2.2. Sensitivity to Capacity, μ. As μ increases, the incentive to prioritize asymptomatic patients decreases. This is intuitive; with more service capacity, some screening can still be provided in addition to diagnosis, even when diagnosis is prioritized. Consequently, it becomes preferable to prioritize symptomatic patients, since their waiting costs are higher than screening patients. To illustrate this observation, we compare the expected arrival rate of symptomatic patients under the two policies for varying μ.

Prioritizing asymptomatic patients causes the environment process to spend more time in better environments, hence it decreases the expected arrival rate of symptomatic patients. Table 5 presents the long-run probability distribution of the environment process for the optimal policy, and the expected arrival rate of symptomatic patients, denoted by $\bar{\lambda}_H$ and $\bar{\lambda}_H^{c\mu}$ for optimal policy and the $c\mu$ rule, respectively.

Table 5 **Comparison of the Two Policies for Different Values of μ for Base Case Parameters**

	Optimal policy					
μ	$e=1$	$e=2$	$e=3$	$e=4$	$\bar{\lambda}_H$	$\bar{\lambda}_H^{c\mu}$
1.73	0.0158	0.0625	0.2476	0.674	0.6952	0.7127
1.8	0.0145	0.0586	0.2387	0.6882	0.6921	0.7067
2	0.0151	0.0649	0.2101	0.7099	0.6899	0.6934
2.2	0.0158	0.0533	0.1921	0.7388	0.6839	0.6839
2.4	0.0115	0.044	0.1789	0.7656	0.6771	0.6771

We observe that, as μ increases, the optimal policy prioritizes screening patients in fewer states, as expected. Although diagnosis obtains priority in a higher number of states, more screening can be provided, and the expected arrival rate of symptomatic patients decreases in μ. This is merely because there will be more cases in which there is no symptomatic patient waiting and hence screening patients can be served. In conclusion, as μ increases, the difference between the policies diminishes, which can also be seen in Figures 2 and 3.

5.2.3. Sensitivity to Cost Parameters, c_H and t_H. We consider $c_H \in \{3.2, 5.2, 7.2\}$ and $t_H \in \{5{,}000, 10{,}000, 20{,}000\}$. As the waiting cost of symptomatic patients, c_H, increases, screening patients obtain less priority since the waiting cost of symptomatic patients becomes dominant. As a result, the improvement by optimal scheduling policy diminishes. The fixed cost associated with a symptomatic patient, t_H, has the reverse effect: as t_H increases, screening becomes more attractive since the benefit of decreasing the future arrival rate increases. Consequently, improvement obtained by optimal policy over the $c\mu$ rule increases as well. This is illustrated in Figure 2.

5.2.4. Sensitivity to Probability of Improving the Environment. This section first explores the effect of $p^L_{e,e+1}$, which we shortly refer to as p here, on the optimal scheduling policy, where $p \in \{0.000625, 0.00125, 0.0025, 0.005, 0.01\}$. We observe that this effect depends on μ, c_H, and t_H. Keeping all the cost parameters the same as the base case, when $\mu = 1.73$, as p increases the optimal scheduling policies do not change, so asymptomatic patients are prioritized in the worst environments, 1 and 2. For $\mu \geq 2$, as p increases, we observe that the system gives asymptomatic patients more priority in all environments. However, for higher values of the waiting cost, c_H, the optimal policy becomes less sensitive to p, since the effect of c_H dominates the scheduling decision. In contrast, as t_H increases, the optimal policy becomes more sensitive to p.

The benefits of optimal scheduling decrease with increasing p: when screening can improve the environment with a high probability, occasional screening provided by the $c\mu$ rule becomes sufficient for decreasing the arrival rate of symptomatic patients. Hence, the improvement that the optimal scheduling policy offers decreases. Figure 3 illustrates this behavior, for $t_H = 20{,}000$.

Diagnostic colonoscopy can also improve the environment, with probability $p_d p^L_{e,e+1}$. For the base case, $p_d = 0.33$. As p_d decreases, screening becomes more important and obtains more priority since then diagnosis cannot improve the environment as well as screening. As p_d increases, we observe that diagnosis obtains more priority, since the difference between screening and diagnostic colonoscopy in terms of preventive power decreases. For the base case parameters with $\mu = 1.73$ ($\mu = 2.4$), the $c\mu$ rule becomes optimal when $p_d \geq 0.55$ ($p_d \geq 0.4$). Note that the most conservative estimate for p_d would be 0.55 for the base case,

Figure 2 **(Color online) Percent Improvement Over the $c\mu$ Rule in the Long-Run Average Cost for the Base Case Parameters with $t_H = 5{,}000$ and $t_H = 20{,}000$**

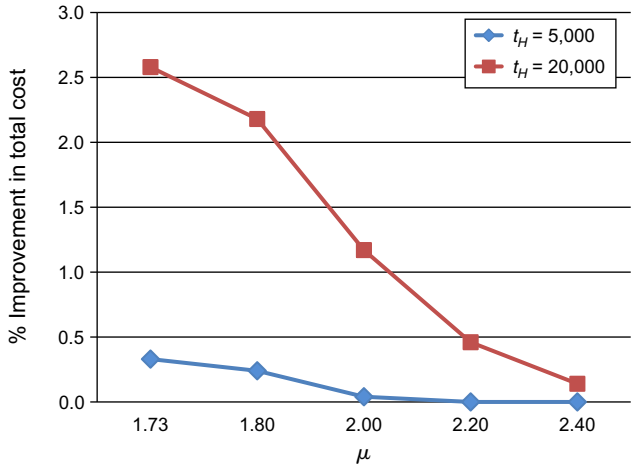
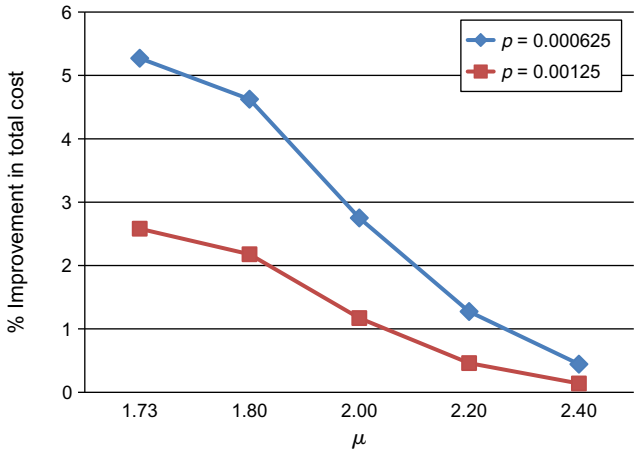


Figure 3 **(Color online) Percent Improvement Over the $c\mu$ Rule in the Long-Run Average Cost for the Base Case Parameters with $t_H = 20{,}000$, $p = 0.000625$, and $p = 0.00125$**



which would be obtained if all symptomatic patients aged 50–75 are assumed to count for screening.

To summarize, we observe that the optimal scheduling policy improves both the average costs and the total arrival rate only marginally when compared to the $c\mu$ rule. The main reason is that screening colonoscopy can be provided in all environments, since the demand for diagnostic services utilizes only a portion of the available capacity. As a result, with both policies it is always possible to screen enough patients to realize its preventive power.

5.3. Comparison of Different Service Priorities for a Hypothetical Case

In this section, we explore situations in which the optimal scheduling policy can achieve a significant improvement over the $c\mu$ rule in the performance measures, long-run average cost, and total expected arrival rate. Our motivation is to provide insights on different settings for which our framework can be used. Breast cancer screening with mammography and scheduling a general practitioner's time for treatment and health promotion activities can be such examples. To this end, we construct a hypothetical case by modifying the parameters of the colonoscopy example.

5.3.1. Parameter Setting. We have run numerous experiments to gain insights on the effects of parameters. Here we summarize these by discussing the construction of the hypothetical case.

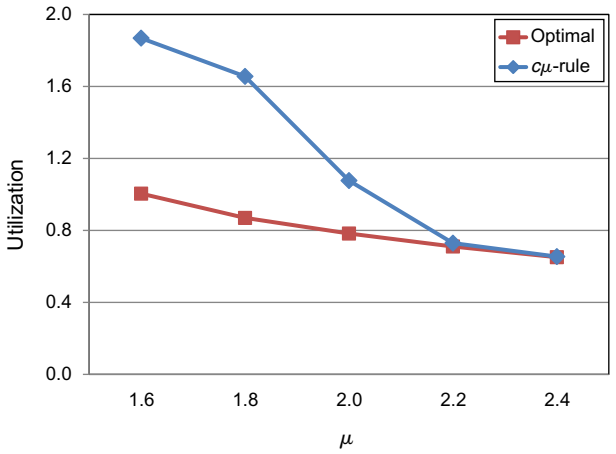
First, if the difference in the probability of improving the environment by the two services is high, then screening can make a higher impact in performance measures. Thus we set $p_d = 0$ and $p^L_{e,e+1} = 0.005$ for all e.

Second, if the impact of environment on the arrival rate of symptomatic demand is high, then controlling the environment via scheduling policies becomes more important. This corresponds to a larger range of values for $\lambda_H(e)$ and also larger differences $\lambda_H(e) - \lambda_H(e-1)$. Accordingly, we set $\lambda_H(1) = 2.13$, $\lambda_H(2) = 1.63$, $\lambda_H(3) = 1.13$, $\lambda_H(4) = 0.63$.

Third, when the system may become stable in the better environments (e.g., $\lambda_H(4) + \lambda_L < \mu$), while unstable in the worse environments (e.g., $\lambda_H(1) + \lambda_L > \mu$), then controlling the environment becomes important, since by screening more patients, the total expected arrival rate can be decreased to a level below the service rate. Consequently, we set $\lambda_L = 0.86$ and experiment with μ in $[1.6, 2.4]$. Notice that the service rate is not sufficient for the two worst environments ($e = 1, 2$) for all μ.

Finally, we set $c_H = 5.2$ in this section and keep the remaining parameters the same with the colonoscopy example.

Figure 4 (Color online) Utilization Values of Different Policies

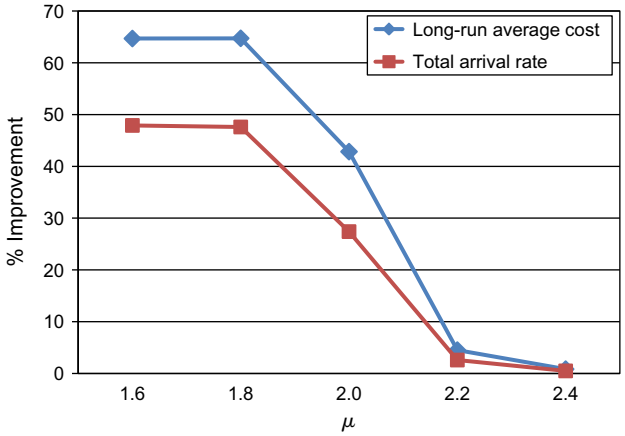


5.3.2. Comparison of the Optimal Policy and the $c\mu$ Rule. For this hypothetical example, the optimal scheduling policy always shows a dynamic structure, similarly to Figure 1 of Example 1 (see §4.2).

Now let us consider the issue of system stability. Note that in the computational experiments, we limit the queue size to 50. We define the utilization of the system as the ratio of the total expected arrival rate, $\bar{\lambda}_H + \lambda_L$, to the service rate μ. For optimal scheduling policies, utilization is below 1 for all values of μ, whereas the $c\mu$ rule generates utilizations that are higher than 1 (Figure 4). When the service rate is low, the $c\mu$ rule cannot finish serving symptomatic patients in the worst environment so that it does not have much opportunity to screen.

Figure 5 presents the percent improvement that the optimal scheduling policy brings over the $c\mu$ rule in the total arrival rate and the long-run average cost. We observe that, when the service capacity is low (for $\mu \leq 2$), there are significant differences between the policies for both performance measures. Note that for

Figure 5 (Color online) Percent Improvement in Long-Run Average Cost and Total Expected Arrival Rate



$\mu \leq 2$ the utilization induced by the $c\mu$ rule is greater than 1.

To summarize, when the arrival rate of one class is modulated by an environment, the system may be stable in one environment state and unstable in the others, as the example constructed in this section shows. If the preventive power of screening is high in the sense that the probability of improving the environment and the effect of this improvement are both high, then the scheduling policy can induce a stable system by regulating screening service provision efficiently. As a result, it improves the performance of the system significantly.

6. Conclusion

This paper introduced a modeling framework using event-based dynamic programming, for facilities providing preventive and diagnostic services. Our main contribution in modeling is the key feature of our framework, which is representing a partial endogeneity between the demand for diagnostic services and the provision of colonoscopy services (both screening and diagnosis). This is essential for capturing the long-term effects of prevention. Defining an environment that can be affected by the decisions of the system itself is a novel idea, which can be possibly used in different contexts.

Structural properties of the model components are identified that can be used to characterize the structure of optimal scheduling policies for systems built within the framework. The model can be extended using the same framework to incorporate admission control of asymptomatic patients, as well as a more general cost function and nonpreemptive service, while keeping all the structural properties shown in this paper; see Kunduzcu (2009).

One interesting result we show is that the endogeneity in the system can induce an optimal scheduling policy that favors serving the less-urgent screening demand as opposed to the demand for diagnostic services. This provides an example for the reversal of the well-known $c\mu$ rule in scheduling.

We also implement the framework for colonoscopy services. The parameter estimation illustrates how this stylized model can represent a real case. In this implementation, the optimal scheduling policy prioritizes diagnosis when at least 50% of the target population is screened, and prioritizes screening otherwise. However, we observe a marginal benefit with the optimal scheduling policy compared to the $c\mu$ rule. Therefore, a practical recommendation is to apply the $c\mu$ rule, i.e., to prioritize diagnostic demand for colonoscopy services.

Further, a hypothetical case illustrates the situations in which the optimal scheduling policy can have a higher effect on the performance. These cases are characterized by the following features: high probability of improving the environment by screening service, low probability of improving the environment by diagnostic service, a significant effect of the environment on the arrival rates, and a medium service capacity which is between the total arrival rate in the best and the worst environments. In such cases, the scheduling policy can induce a stable system by regulating the screening efficiently, whereas the $c\mu$ rule cannot. Consequently, the performance of the system is improved significantly by the optimal scheduling policy.

Our modeling framework can be used in other contexts as well. Machine and facility maintenance are examples of sharing a resource for repair and preventive services, where a machine's failure rate can be affected by preventive maintenance. Healthcare facilities, on the other hand, provide a variety of examples for such contexts. For example, mammography is used for both screening and diagnosis purposes for breast cancer, and family doctor duties include vaccination and health promotion in the community in addition to regular treatment and diagnosis. The purpose of the former duties is prevention of serious illnesses, whereas the latter correspond to curative services.

This paper considers only a negative correlation between colonoscopy services and symptomatic demand, since colonoscopy is a nearly perfect test and does not create unnecessary demand on the diagnostic resources. A positive correlation would be observed when the screening test has less than perfect specificity, i.e., if the result could be a false positive. When the probability of a false positive is significant, then additional tests would be required to verify the positive result. One example is the use of mammography in breast cancer screening and diagnosis. When a screening mammography result is positive, then a diagnostic mammography is required. Our model can be extended easily to represent such situations if necessary. In such cases screening will have two opposing effects on the arrival rate of symptomatic demand: an increasing effect as a result of the false positives, and a decreasing effect as a result of prevention. The balance between the two would determine the optimal scheduling policy.

In this paper we assume that the service rates for screening and diagnostic services are equal. This is a realistic assumption for colonoscopy services. However, as in the example of mammography, diagnostic tests may take longer than the screening tests. Modeling such cases in our framework is possible. However, the structural properties should be further explored. We conjecture that longer diagnostic services (lower μ

for diagnosis) would make these services less attractive. As a result, incentives to provide screening may increase since benefits of screening would remain the same. Note that the $c\mu$ rule would also suggest such a change in the incentive structure. Another simplifying assumption in our model is equal service costs for the two services. When diagnostic services are more expensive, the relative cost of symptomatic demand increases and it becomes important to decrease that arrival stream. Consequently, we expect that incentives to screen increase.

Our model focuses on a single-server case and recommends the $c\mu$ rule for scheduling in colonoscopy services. In reality, patients can receive service from multiple colonoscopy facilities. We tested the performance of the recommended policies when the system has two facilities in the stochastic simulation model, and we observed no significant changes in our conclusions. However, the simulation model does not account for patient behavior such as jockeying and making multiple appointments. A new modeling approach is required to represent patient behavior, which can be explored in future research. Finally, a more realistic approach to modeling the system could include an appointment scheduling component that provides a scheduling date. The waiting time for screening appointments may affect patient compliance behavior. A more detailed analysis of such systems can be explored via simulation models.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2015.0556.

Acknowledgments

The authors gratefully acknowledge Javad Lessan for his technical assistance in developing the simulation model. They are also grateful to the anonymous review team for constructive feedback that improved the paper significantly.

References

Alagoz O, Ayer T, Erenay FS (2011) Operations research models for cancer screening. Cochran JJ, Cox LA Jr, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).

Alagoz O, Maillart LM, Schaefer AJ, Roberts MS (2007) Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Oper. Res.* 55(1):24–36.

American Cancer Society (2008) Colorectal cancer facts and figures 2008–2010. Report, American Cancer Society, Atlanta.

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

Artalejo J (2010) Accessible bibliography on retrial queues: Progress in 2000–2009. *Math. Comput. Modelling* 51(9–10):1071–1081.

Bhulai S, Brooms AC, Spieksma FM (2014) On structural properties of the value function for an unbounded jump markov process with an application to a processor sharing retrial queue. *Queueing Systems* 76(4):425–446.

Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U (2007) Risk of progression of advanced adenomas to colorectal cancer by age and sex: Estimates based on 840,149 screening colonoscopies. *Gut* 56(11):1585–1589.

Butterly L, Olenec C, Goodrich M, Carney P, Dietrich A (2007) Colonoscopy demand and capacity in New Hampshire. *Amer. J. Preventive Medicine* 32(1):25–31.

Çil EB, Örmeci EL, Karaesmen F (2009) Effects of system parameters on the optimal policy structure in a class of queueing control problems. *Queueing Systems* 61(4):273–304.

de Bosset V, Froehlich F, Rey J, Thorens J, Schneider C, Wietlisbach V, Vader J, Burnand B, Muhlhaupt B, Fried M, Gonvers J (2002) Do explicit appropriateness criteria enhance the diagnostic yield of colonoscopy? *Endoscopy* 34(5):360–368.

Donnellan E (2008) Nine-month wait for bowel cancer test. *Irish Times* (November 15), http://www.irishtimes.com/newspaper/ireland/2008/1115/1226700611205.html.

Efrosinin D, Breuer L (2006) Threshold policies for controlled retrial queues with heterogeneous servers. *Ann. Oper. Res.* 141(1):139–162.

Erenay FS, Alagoz O, Said A (2014) Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing Service Oper. Management* 16(3):381–400.

Forde KA (2006) Colonoscopic screening for colon cancer. *Surgical Endoscopy* 20(2):S471–S474.

Gayon J, Talay-Değirmenci I, Karaesmen F, Örmeci E (2009) Optimal pricing and production policies of a make-to-stock system with fluctuating demand. *Probability Engrg. Inform. Sci.* 23(2):205–230.

Gopalappa C, Aydogan-Cremaschi S, Das T, Orcun S (2011) Probability model for estimating colorectal polyp progression rates. *Health Care Management Sci.* 14(1):1–21.

Green LV, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. *Oper. Res.* 54(1):11–25.

Güneş ED, Örmeci EL, Kunduzcu D (2015) Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. *Production Oper. Management* 24(1):1–20.

Hordijk A, Koole GM (1993) On the optimality of lept and μc rules for parallel processors and dependent arrival processes. *Adv. Appl. Probab.* 25(4):979–996.

Kolata G (2003) 50 and ready for colonoscopy? Doctors say wait is often long. *New York Times* (December 8), http://www.nytimes.com/2003/12/08/us/50-and-ready-for-colonoscopy-doctors-say-wait-is-often-long.html.

Koole GM (2006) Monotonicity in Markov reward and decision chains: Theory and applications. *Foundations Trends Stochastic Systems* 1(1):1–76.

Kunduzcu D (2009) Admission and scheduling control for preventive services. M.S. thesis, Koç University, Istanbul.

Levin B, Lieberman D, McFarland B, Smith R (2008) Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the U.S. Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J. Clinicians* 58(3):130–160.

Lieberman D, Holub J, Eisen G, Kraemer D, Morris C (2005) Utilization of colonoscopy in the united states: Results from a national consortium. *Gastrointestinal Endoscopy* 62(6):875–883.

Lippman SA (1975) Applying a new device in the optimization of exponential queueing systems. *Oper. Res.* 23(4):687–710.

Mayo Clinic (2011) Information on colonoscopy. Accessed October 2, 2015, http://www.mayoclinic.com/health/colon-cancer-screening/MY00935/NSECTIONGROUP=2.

Özekici S, Soyer R (2003) Network reliability assessment in a random environment. *Naval Res. Logist.* 50(6):574–591.

Ozkan C, Karaesmen F, Özekici S (2013) Structural properties of Markov-modulated revenue management problems. *Eur. J. Oper. Res.* 225(2):324–331.

Patrick J, Puterman M, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* 56(6):1507–1525.

Pierskalla WP, Voelker JA (1976) A survey of maintenance models: The control and surveillance of deteriorating systems. *Naval Res. Logist. Quart.* 23(3):353–388.

Prabhu NU, Zhu Y (1989) Markov-modulated queueing systems. *Queueing Systems* 5(1–3):215–245.

Puterman M (1994) *Markov Decision Processes* (John Wiley & Sons, Hoboken, NJ).

Regula J, Rupinski M, Kraszewska E, Polkowski M, Pachlewski J, Orlowska J, Nowacki M, Butruk E (2006) Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. *New England J. Medicine* 355(18):1863–1872.

Schaefer AJ, Bailey MD, Shechter SM, Roberts MS (2004) Modeling medical treatment using markov decision processes. Brandeau M, Sainfort F, Pierskalla W, eds. *Operations Research and Health Care: A Handbook of Methods and Applications* (Kluwer, Dordrecht, Netherlands), 597–616.

Shechter SM, Bailey MD, Schaefer AJ, Roberts MS (2008) The optimal time to initiate HIV therapy under ordered health states. *Oper. Res.* 56(1):20–33.

U.S. Preventive Services Task Force (2009) Screening for colorectal cancer, topic page. Accessed October 2, 2015, http://www.ahrq.gov/clinic/uspstf/uspscolo.htm.

Vijan S, Hwang E, Hofer T, Hayward R (2001) Which colon cancer screening test? A comparison of costs, effectiveness, and compliance. *Amer. J. Medicine* 111(8):593–601.

Wang H, Pham H (2006) *Reliability and Optimal Maintenance* (Springer, New York).

Winawer SJ, Zauber AG, Ho MN, O'Brien MJ, Gottlieb L, Sternberg SS, Waye JD, et al. (1993) Prevention of colorectal cancer by colonoscopic polypectomy. *New England J. Medicine* 329(27):1977–1981.

Wu S, Zuo MJ (2010) Linear and nonlinear preventive maintenance models. *IEEE Trans. Reliability* 59(1):242–249.