PRODUCTION AND OPERATIONS MANAGEMENT POMS
Vol. 27, No. 12, December 2018, pp. 2226–2250
ISSN 1059-1478 | EISSN 1937-5956 | 18 | 2712 | 2226
DOI 10.1111/poms.12858

Missed Opportunities in Preventing Hospital Readmissions: Redesigning Post-Discharge Checkup Policies

Xiang Liu*
Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, Michigan 48109, USA, liuxiang@umich.edu

Michael Hu
Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA, hum@mit.edu

Jonathan E. Helm
W. P. Carey School of Business, Arizona State University, 300 E Lemon Street, Tempe, Arizona 85287, USA, jehelm2@asu.edu

Mariel S. Lavieri
Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, Michigan 48109, USA, lavieri@umich.edu

Ted A. Skolarus
Department of Urology, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109, USA, tskolar@med.umich.edu

Hospital readmissions affect hundreds of thousands of patients every year, negatively impacting patients and placing a tremendous burden on the national healthcare system. Post-discharge checkup policies can reduce readmissions through early detection of health conditions, however, the methods behind designing effective checkup policies are poorly understood. Under current practice, up to 67% of readmitted patients return to the hospital before their first scheduled office visit. This work aims to develop effective checkup plans to monitor patients following hospital discharge, using a variety of checkup methods, including phone calls and office visits. We develop and analyze a new delay-time analysis model to identify the optimal type and timing of checkups to implement post-discharge monitoring plans. By analyzing the structure of optimal policies, we develop checkup schedules that can detect up to 43.7% more readmission-causing conditions experienced by readmission-bound patients. Further, we uncover simple rules of thumb that can help doctors design and improve monitoring plans even in the absence of advanced computer software or complex computations.

Key words: disease screening; machine maintenance delay-time model; hospital readmissions; telemedicine
History: Received: January 2016; Accepted: November 2017 by Sergei Savin, after 3 revisions.

1. Introduction

In the United States, hospital readmissions are heavily scrutinized as a driver of healthcare costs. According to Weinberger et al. (1996), up to half of all hospitalizations are readmissions. Furthermore, it is estimated that up to 75% of all readmissions are preventable by patient education, pre-discharge assessment, and domiciliary aftercare (Benbassat and Taragin 2000). In effect, preventable hospital readmissions represent approximately $25 billion in annual healthcare costs (PwC Health Research Institute 2010). One in eight Medicare patients are readmitted within 30 days of discharge after surgery (Dartmouth Atlas Project 2013), and 56.5% of readmitted Medicare patients are readmitted through the Emergency Department (ED) (Kocher et al. 2013), contributing to high costs. These readmissions represent not only preventable healthcare costs, but also a tremendous burden on patients and their families.

In order to address this problem, policies such as the Affordable Care Act (ACA) have been implemented (Koh and Sebelius 2010). Following the ACA, the Centers for Medicare and Medicaid

Services (CMS) now penalize hospitals with worse than expected 30-day readmission rates (Joynt and Jha 2012). For example, Section 3025 of the Affordable Care Act added Section 1886(q) to the Social Security Act establishing the Hospital Readmissions Reduction Program. This program requires CMS to reduce payments to the Inpatient Prospective Payment System (IPPS) hospitals beginning in October 2012 (James 2013). These circumstances encourage healthcare professionals to more actively search for and implement solutions to minimize hospital readmissions (Wong et al. 2013).

Numerous interventions have been proposed to prevent readmissions (including better pre-discharge care and improved discharge instructions). Post-discharge checkups such as phone calls, home visits, and office visits have been independently shown in the clinical literature to significantly reduce hospital readmissions (Dudas et al. 2001, Wong et al. 2013) and offset increases in demand for physician services (Green et al. 2013). The purpose of these checkups is to detect developing conditions before they worsen and cause either an unnecessary trip to the ED and/or an inpatient readmission.

Although checkups can mitigate the readmissions crisis, the methods behind designing effective checkup policies are poorly understood. Specifically, healthcare providers remain uncertain about how many checkups to schedule, what types of checkups to schedule, and when to schedule those checkups. In practice, checkup policies currently implemented by hospitals are designed and based on unsupported heuristics. For example, current practice recommends that doctors first follow-up with cystectomy (a major surgery for bladder cancer) patients with an office visit approximately 2 weeks after their hospital discharge; however, 40% of readmitted cystectomy patients are readmitted within 1 week of discharge, and as many as 67% of readmitted cystectomy patients are already readmitted before the first scheduled office visit (Hu et al. 2014, Skolarus et al. 2015). Our research seeks to reclaim this missed opportunity by identifying the optimal timing as well as the type of checkups to perform after discharge. It also provides guidance for how many visits would be most effective. This will give healthcare professionals (both clinicians and non-physicians) an increased chance of detecting a patient's health condition before it causes a readmission.

Because most hospitals do not yet have a systematized mechanism for managing follow-ups for their cohort of patients, much of the follow-up decision making is left to the treating surgeon, and it is typically determined on a case-by-case basis. This work seeks to improve the efficacy of these personalized follow-up plans. This approach has been confirmed as having low barriers to implementation relative to a larger scale, system-wide approach that considers costs and savings relative to total hospital resources. This is because medical professionals currently make decisions on a per-patient basis (hence no major culture change required) by weighing the expected benefit (e.g., early detection, readmission reduction, improved quality, etc.) vs. the amount of time the practice is able/willing to commit to follow-ups. Cost-based calculations are not frequently used in these individual patient decisions, in part because it is difficult to assign a monetary value to early detection of a condition. This study provides analytical, data-based methods and decision guidelines (medical professionals are comfortable with both) to better personalize these decisions that doctors already make on a daily basis.

To provide contextual grounding for our practice-focused readmission detection approach, we develop our models in close collaboration with a urological practice, with a focus on cystectomy, which is one of the highest readmission rate surgeries in the United States. Other papers have shown similarities in the readmission characteristics of cystectomy patients and other types of surgical patients (Jacobs et al. 2017). This approach could hence be generalizable to other types of surgery and other patient conditions by changing the model parameterization based on historical data, as long as the processes for follow-ups and underlying disease dynamics remain similar. More information about the key assumptions that must be verified before applying our models to other diseases is provided in subsequent sections.

The post-discharge monitoring process after cystectomy proceeds as follows. At the time of discharge, a monitoring schedule is determined by the discharge team and the patient is made aware of when they will be receiving phone calls and when they are scheduled to return for an office visit to check on their recovery. During a phone call or office visit, the patient will be tested to see if they have developed a condition that is likely to lead to readmission. For cystectomy, the two most common conditions are infection and failure to thrive (unable to eat enough food), which account for the majority of readmissions (see Hu et al. 2014, Skolarus et al. 2015). These conditions exhibit important characteristics that are suited to early detection and mitigation: (i) these types of conditions are readily detectable via phone call, telemedicine, or office visit, (ii) the window for detection is long enough to make a follow-up potentially effective (e.g., patients stay at home with an infection for several days before becoming sick enough for readmission), and (iii) early detection can be effective in mitigating the condition on an outpatient basis or at the very least result in a reduced cost ED visit or readmission (e.g., providing antibiotics to treat infection, or early detection means the condition is less serious when treatment begins leading to reduced cost and better patient outcomes).

If a condition is detected early by a follow-up, steps to mitigate the condition can be immediately undertaken. These steps can include starting antibiotics to eliminate infection, or IV treatment for patients suffering from failure to thrive. Hence, early detection may avoid the readmission entirely, prevent an expensive ED visit, or at the very least lessen the time and cost of overcoming the condition while improving the quality of the outcome by catching the condition before it becomes too severe. At the suggestion of our clinical collaborator, we do not attempt to directly quantify the monetary value of such outcomes in our model, but instead leave the decision to the clinician/practice as to the amount of follow-up effort that is reasonable relative to the increased likelihood of early detection.

To capture this personalized follow-up process, we develop a delay-time modeling approach adapted from the machine maintenance literature to analyze and optimize post-discharge checkup policies. Several unique features of readmission dynamics require new extensions of the traditional framework, providing new insights into the structure of delay-time machine maintenance problems and broadening the scope of problems in which delay-time analysis can be applied. In addition to theoretical implications, this study contributes beneficial insights for physicians and other healthcare decision makers to help them improve post-discharge monitoring for patients.

As a proof of concept, we calibrate, test, and validate our models on nationwide data for cystectomy patients. Cystectomy, often performed on bladder cancer patients, is a type of surgery that involves removal of all or part of the urinary bladder. Cystectomy patients experience one of the highest readmission rates of any surgery, as approximately 25% of cystectomy patients are readmitted within 30 days of discharge from the hospital (Hu et al. 2014, Jacobs et al. 2013).

The structure of this study is as follows. In sections 3 and 4, we develop and analyze our model to understand key properties of the optimal checkup policies. We identify the importance of checkup timing, and how checkup timing is affected by the stochasticity of how long patients are ill prior to readmission (delay-time), as well as the detection rate of checkups. In section 5, we verify our findings through numerical analyses by applying our model to national State Inpatient Database (SID) patient cohorts. The numerical analyses also demonstrate that our model is robust to the system parameters and consistently outperforms current checkup policies. Using the same number of checkups, current practice (which is expected to detect only 16% of the conditions experienced by readmitted patients) can be improved by up to 43.7%. In section 6, we summarize the theoretical and practical implications of our study. In particular, we highlight how our model provides valuable extensions to the traditional delay-time analysis framework and how our findings can effectively detect readmission-causing conditions and improve the quality of patient care, thereby mitigating the national readmissions crisis.

2. Literature Review

Readmissions play a critical role in recent clinical literature. It is estimated that up to 75% of readmissions are preventable by patient education, pre-discharge assessment, and domiciliary aftercare (Benbassat and Taragin 2000), and post-discharge checkups such as phone calls, home visits, pharmacists' visits, and doctors' office visits can significantly reduce hospital readmissions (Bellone et al. 2012, Costantino et al. 2013, D'Amore et al. 2011, Dudas et al. 2001, Wong et al. 2013). Within the healthcare operations research literature, models have been created to improve post-discharge health outcomes, including reducing readmissions and mortality rates: Bartel et al. (2016) analyzes how the initial hospitalization length of stay impacts post-discharge mortality rate; Chan et al. (2012) studies the impact of ICU discharge strategies on readmissions; Kim et al. (2014) analyzes how ICU admission control strategies impact readmission rate. Bayati et al. (2014) builds a classification model to predict readmissions and analyzed intervention decisions. However, this work does not address the timing of interventions. Leeds et al. (2015) conducts a statistical analysis to study how surgeons make discharge decisions and the effect of decision-support tools for discharge. None of those models directly address how patients should be monitored after hospital discharge. To address that question, two areas in the operations research literature are especially relevant to our study: (i) machine maintenance and inspection, and (ii) disease screening.

Machine maintenance and inspection: The literature of machine maintenance and inspection is very well established. Literature surveys (Barlow and Proschan 1996, Wang 2002) categorize maintenance policies into two groups: preventive maintenance (PM) and corrective maintenance (CM). Our problem aligns more closely with PM frameworks since PMs proactively prevent failure, whereas CMs are only performed after failures occur. PMs can be scheduled in the following fashion: (i) age-dependent policies perform PM at a fixed time T; (ii) periodic and sequential policies schedule multiple PMs in fixed or variable intervals; and (iii) failure limit policies perform PMs when the failure rate of a machine exceeds a predetermined threshold. The dynamics of machine deterioration are typically modeled by (i) Markovian processes (Sim and Endrenyi 1993), (ii) semi-Markovian processes (Milioni and Pliska 1988, Yeh 1997), (iii) hidden Markov models (White 1977), and (iv) delay-time models (Wang 2012). More specifically, Sim and Endrenyi (1993) models the deterioration as a continuous time Markov chain and considers multiple failure types and repair/maintenance actions. They minimize the long-run average downtime and cost, which is not suitable for our problem. Yeh (1997) uses phase-type distributions to approximate general distributions of a semi-Markovian model. They develop algorithms for optimal state-age-dependent policies that also minimize long-run average cost. White (1977) develops a POMDP model for the machine inspection/maintenance problem which minimizes the long-run average cost. These models are not suitable for our problem because they assumed Markovian deterioration and optimized long-run average cost and downtime.

Wang (2012) gives a thorough survey on delay-time models, which are a special case of semi-Markovian models with three states. Traditional delay-time analysis is based on renewal theory and reliability which assumes the unit lifetime has increasing failure rate. The goal of those models is typically to determine an inspection schedule that minimizes long-run costs (Christer and Jack 1991, Jardine and Tsang 2005) or minimizes expected downtimes (Dagpunar 1994) given identical units that can be replaced. Our problem necessitates several extensions: (i) unlike interchangeable machine components, patients cannot be "replaced"; (ii) our objective is to maximize the probability of a checkup (inspection) detecting a patient's condition; (iii) readmission rates depend on time since discharge, so we have a time-varying failure rate; and (iv) existing models do not allow for policies composed of different types of inspections with varying inspection detection rates (Christer 1999). Monitoring policies composed of inhomogeneous checkups (e.g., phone calls, office visits, etc.) are particularly valuable because empirical evidence indicates that policies consisting of mixed checkup methods are more effective than policies consisting of only one checkup method (Holland et al. 2005, Wong et al. 2013).

Close to our work is Milioni and Pliska (1988), where a semi-Markovian model with three states was used to model machine deterioration and catastrophic failure (i.e., no repair/replacement after failed). They considered two objectives: minimize the cost of inspections, false positives, and treatment; and minimize the probability of failure. Existence of optimal solutions and algorithms for solving the problems were established. However, the authors did not provide insights into the structure of the optimal policies. Moreover, they assumed perfect inspections in the sick state. Although this model is somewhat similar to our model, the key difference is that this model is still a long-run steady-state planning model in both objective functions.

Disease screening: Within the healthcare operations research field, models have been developed to establish medical inspection schedules that detect the onset and progressions of diseases such as chlamydia infection (Teng et al. 2011), diabetes (Brandeau et al. 2004), AIDS (Deo et al. 2014, Sanders et al. 2005), hepatitis (Fu et al. 2012), breast cancer, (Ayer et al. 2012, Ayer et al. 2015, Brailsford et al. 2012, Maillart et al. 2008), colorectal cancer (Erenay et al. 2014, Güneş et al. 2015, Harper and Jones 2005) cervical cancer (Myers et al. 2000), prostate cancer (Pinsky 2004, Tsodikov et al. 2006, Zhang et al. 2012a), bladder cancer (Kent et al. 1989), and glaucoma (Helm et al. 2015). Delay-time models are used to model hepatitis progression (Fu et al. 2012), and vascular patency loss (Zhang et al. 2012b). Most of the models are based on discrete time Markovian assumptions (Ayer et al. 2012, Ayer et al. 2015, Erenay et al. 2014, Kent et al. 1989, Maillart et al. 2008, Myers et al. 2000, Zhang et al. 2012a), which do not fit into our problem since the deterioration dynamics of the readmitted patients are not necessarily Markovian.

Bavafa et al. (2013) studies a three-state continuous time Markov model in the context of primary care routine visits. The authors examine the effectiveness of office visits as well as e-visits as a cost-effective preventative action. However, the model assumes Markovian deterioration and focuses on steady-state planning from the perspective of the primary care providers. Fu et al. (2012) applies delay-time models on hepatitis screening. However, they focus on optimal statistical estimation rather than the optimal monitoring schedule planning. Closest to our work is Zhang et al. (2012b), where follow-up checkups are scheduled to minimize the probability that the time between patency loss and its detection exceeds some length of time. The results on the timing of checkups under the assumptions of deterministic delay-time and Weibull-distributed failure rate are generally consistent with our findings. However, the authors consider perfect checkups only and do not consider general distributions. Their work focuses on the timing of checkups only and does not study how quantity, quality, or mix of different checkups impact monitoring schedules. Moreover, they estimate the distributions using maximum likelihood methods assuming Erlang and exponential distributions, whereas we use best-fit distributions obtained directly from the data. The novelties of our work leverage the composition of different checkup methods (e.g., office visits and phone calls) and address the tradeoffs in scheduling checkups with both perfect and imperfect inspections under inhomogeneous failure rates. Our work differentiates from Zhang et al. (2012a) in the following aspects. (i) In contrast to their model, we analyze the optimal structural of the checkup policies (consisting of perfect checkups) without assuming a specific parametric family. (ii) For imperfect checkups, we show that imperfect checkups (such as phone calls) can affect the timing and detection probability significantly by considering the detection rate of checkups. Moreover, (iii) we incorporated various sources of data to estimate the hidden time-to-develop the condition distributions using numerical Laplace inverse transform. Helm et al. (2016) developed a mixed integer programming (MIP) approach to solving a planning problem for how many healthcare professionals to staff to implement a follow-up program. This model, however, assumed a homogeneous population(s) of patients and was designed as a static planning model for a cohort of patients taking the hospital's perspective. Our model, on the other hand, is patient centered and can be tailored based on each individual's projected readmission density curve—focusing on the operational level rather than a steady-state planning model. Our delay-time modeling approach also enables us to identify structural properties, which is not possible using their MIP formulation. Personalized prediction can be incorporated in our model by (i) estimating a population-based survival function to model the time to readmission curve using a Weibull regression and associated risk factors (socio-demographic, hospital admission and stay characteristics, etc.), and then (ii) applying transfer learning techniques to individualize each patient's readmission curve (Helm et al. 2016). In this study, we start with a population-based readmission curve to focus more on the structural insights into the optimization model given the readmission curve. To demonstrate how our model performs on patients with different readmission risk, we obtained optimal policies for three risk profiles (low, medium, and high) from Helm et al. (2016) (see Appendix A). With sufficient data, our model fully supports a personalized monitoring approach.

3. Model for Optimizing Post-Discharge Checkup Policies

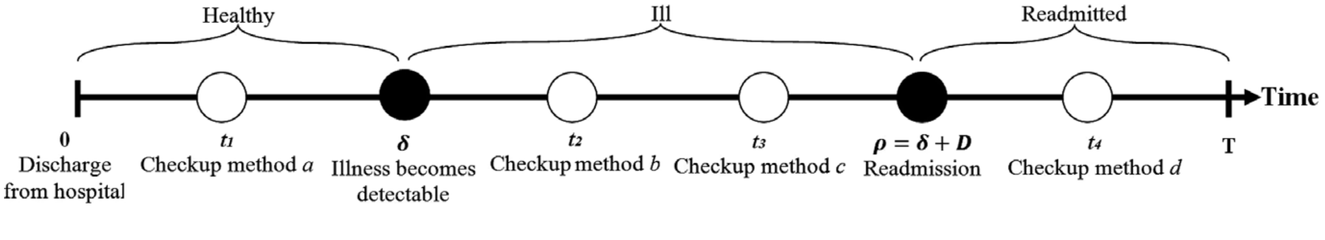
In this section, we develop and analyze a general model for designing monitoring plans for patients after they are discharged from the hospital. First, we introduce our model notation and parameters (a summary of the notation can be found in Appendix B). Next, we develop our general model.

3.1. Delay-Time Model for Readmissions

Based on our field research, the dynamics of an inpatient readmission occur as follows. After a patient is discharged, he/she may develop a readmission-causing condition. When this condition first develops, it does not necessarily cause an immediate readmission (e.g., an infection). Instead, the patient's condition will degrade over time, eventually becoming so severe that he/she must return to the hospital and be readmitted. These dynamics are identical to those found in machine failure models, which have been shown in the machine maintenance literature to be well modeled by a delay-time model. Unlike Markovian models, our model handles general distributions under mild conditions (see section 3.3). Moreover, since our problem has a short planning horizon (30 days) and a transient nature (patient-centric not steady-state planning), continuous delay-time models allow us to keep track of how long a patient has been in each state and we can tailor the objective function as we shall see later. As seen in Figure 1, we consider individual patients stochastically progressing through three sequential states upon discharge: healthy, ill, and readmitted. Thus, within the framework of traditional delay-time analysis models used in preventative maintenance, the patient represents the system, illnesses represent defects, and readmissions represent failures.

Remark 1. ("Ill State"). It is important to note here that the ill state is defined as identifying a patient in a state that causes them to be at risk for a future readmission. This includes conditions such as infection and failure to thrive, but also includes conditions such as when the patient has failed to fill a prescription, is taking their medicine incorrectly, or

Figure 1 Patient State Progression and Checkup Policy



has not understood or followed post-discharge treatment plans such as exercise or nutritional guidelines. Both medical and compliance issues can be checked for during a phone call or office visit and incorporated into our modeling framework.

At time 0, we assume a patient is discharged in a healthy state. After a stochastic amount of time, δ, the patient develops a detectable condition and is considered to be in the ill state (the first black dot in Figure 1). We denote this time δ as the time-to-develop the condition. Following a period of time (delay-time), D (between the first and second black dots in Figure 1), the patient's condition worsens to the point where he/she is readmitted to the hospital. We denote this time-to-readmission as $\rho = \delta + D$ (the second black dot in Figure 1). Lastly, we let T denote the length of our model's planning horizon (e.g., $T = 30$ days). Clinical literature and policy both support a finite horizon model as the Centers for Medicare and Medicaid Services specify that hospital admissions only qualify as readmissions if they occur within 30 days of discharge.

At the point of a patient's discharge, the case manager needs to determine the post-discharge checkup plan for the patient for the next 30 days. Given n checkup opportunities, our goal is to place a checkup at each time t_i, $i \in \{1, \ldots, n\}$ (white circles in Figure 1), to maximize the probability of detecting the patient in the ill state. While there is a possibility of a competing risk of patient mortality, 30-day mortality rates post-discharge are very small relative to readmission rates.

In addition to choosing checkup times, decisions must be made regarding what type of checkup method (e.g., phone calls, home visits, doctors' office visits) to use at each checkup time, t_i. Given m different checkup methods, the indicator variable $y_{ij} \in \{0, 1\}$ denotes whether checkup method $j \in \{1, \ldots, m\}$ is used at time t_i. In Figure 1, $y_{1a} = y_{2b} = y_{3c} = y_{4d} = 1$. To model checkup method resource limitations, let w_j denote the maximum number of times checkup method $j \in \{1, \ldots, m\}$ can be used.

As mentioned in the contextual grounding of section 1, we are developing this research to help personalize monitoring plans for each patient at the provider/practice level. Thus, we allow these constraints to be tailored to what the clinician believes is an appropriate level of checkup intensity (i.e., how many office visits and phone calls they are able/willing to make). For example, our clinical collaborator indicates that most surgeons would typically be willing to do one office visit, two in cases where they are more concerned about the patient, and a maximum of three where the patient's condition indicates very high risk. These determinations, however, are typically made by the clinician based on a medical and historical knowledge of the patient and their condition and are difficult to quantify in a cost-based or constraint-based structure. Further, budgets for checkups are not typically considered when making individual checkup decisions for specific patients, hence the inclusion of costs does not fit the current practice and would provide barriers given that many clinicians are averse to such an approach in individual patient decision making. Hence, we allow the provider/practice to determine how many office visits and phone calls (i.e., w_j's) they believe to be appropriate on a patient-by-patient basis and enter this number as a model parameter. The model also allows for clinicians to perform sensitivity analysis to determine, for example, the marginal benefit of an extra phone call or office visit compared to their base resource allocation.

To account for the differences in checkup methods, that is, an office visit is more effective than a phone call, we let the detection rate $r_j \in [0, 1]$ denote the probability that method $j \in \{1, \ldots, m\}$ will detect a condition when the patient is in the ill state (i.e., true-positive). If $r = 1$, then we say that the checkup is a perfect checkup. If $r < 1$, we say that the checkup is an imperfect checkup. The detection rate accounts for the chance that a condition is present and yet is not detected. This could be due to an inability to detect illness based on the questions asked, poor patient responsiveness, or other reasons. Patients not answering the phone can also be considered, but based on discussions with a company that provides automated phone calls to detect readmittable conditions (www.cloud9hcs.com), they achieve full patient responses to their readmission detection scripts (questions) in greater than 85% of their phone calls. We do not consider false-positives in this model.

Each checkup policy is now defined as, $\Pi = (t_1, \ldots, t_n, y_{11}, \ldots, y_{nm})$. Further, let $N_i^{\Pi} \in \{0, 1\}$ be the indicator variable denoting whether or not the patient is detected in an ill state at time t_i, given policy Π. Our objective is to select the checkup policy that maximizes the probability of detecting the patient in an ill state (detection probability in shorthand):

$$\max_{\Pi} \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi}]. \tag{1}$$

3.2. Model Formulation and Solution Approach

The time-to-develop the condition, δ, is described by a differentiable probability density function $g_{\delta}(\cdot)$, which is assumed to be independent of delay-time, D. This assumption is necessary for the mathematical formulation and is present in all related machine maintenance literature. We also confirm statistical independence of these two random variables in Section 5.1, using historical data. D has PDF $f(\cdot)$, CDF $F(\cdot)$, and complementary cumulative distribution function (CCDF) $\bar{F}(\cdot)$. Furthermore, the time-to-readmission, ρ, has probability density function $g_\rho(\cdot)$, which is the convolution of δ and D. The checkup optimization can be expressed as follows:

$$\max_{\substack{t_1,\ldots,t_n\\ y_{11},\ldots,y_{nm}}} \quad \sum_{i=1}^{n}\sum_{\beta=1}^{m} y_{i\beta} r_\beta \sum_{s=1}^{i} \int_{t_{s-1}}^{t_s} g_\delta(k)\bar{F}(t_i-k)dk \prod_{q=s}^{i-1}\sum_{\alpha=1}^{m} y_{q\alpha}(1-r_\alpha) \tag{2}$$

$$\text{s.t.} \quad \sum_{l=1}^{m} y_{il} = 1, \quad \forall\, i \in \{1,\ldots,n\} \tag{3}$$

$$\sum_{i=1}^{n} y_{il} \le w_l, \quad \forall\, l \in \{1,\ldots,m\} \tag{4}$$

$$0 \le {}_i < t_{i+1} \le T, \quad \forall\, i \in \{1,\ldots,n-1\} \tag{5}$$

where $t_0 = 0$ and the empty product, Π, equals 1.

The first term in the objective, $y_{i\beta} r_\beta$, accounts for the detection rate of the method used for checkup i. The second term represents the probability that the patient developed the condition between checkups $(s-1)$ and s and is still not readmitted by checkup i. The last term (the product) represents the probability that checkups $s, \ldots, (i-1)$ all failed to properly detect the patient's existing condition. The constraint of Equation (3) ensures that only one checkup method is utilized at each checkup time, Equation (4) ensures that checkup method resource capacities are not violated, and Equation (5) ensures proper ordering of the checkups.

The goal is to design a complete post-discharge checkup plan at the time of the patient discharge. While the person doing the checkup could potentially learn new information with each phone call or office visit that could dynamically modify the time-to-develop the condition/time-to-readmission curve, this dynamic updating is out of scope of this project for several reasons. First, dynamic changes in schedule can be logistically difficult for both patients (having a constantly shifting schedule of appointments can interfere with their normal lives, lives of caregivers, and their sense of comfort/consistency) and for providers (changing their own checkup plans for a cohort of "in flight" patients and possibilities of conflicts between schedules). Second, this is not the way the system is currently designed and would likely provide significant barriers to adoption in clinical practice. Another consideration is that the presence of checkups themselves may reduce the likelihood of a patient becoming ill. Since, to our knowledge, there is no method based on the available data to account for improvement in the time-to-develop the condition density curve based on frequency and timing of checkups (and our goal is to develop a data-driven, practical approach), we omit this from our model and instead rely on our conservative estimate of the potential benefit of improved checkup schedules by disregarding additional educational benefits of checkups.

Remark 2. Note that our objective function only considers the probability of detection and does not account for how early the condition was detected. We chose this objective for several reasons. First, it is intuitive for the clinical audience and captures the essence of the post-discharge monitoring goal—to detect conditions and prevent readmissions. Second, there is no data, to our knowledge, that captures the benefits of capturing a condition earlier vs. later. Nevertheless, capturing conditions early would likely be beneficial. It is possible to modify our objective function to achieve this, given proper data on the benefits of early detection.

Solution Approach: We solve this program numerically by dividing it into subproblems and enumerating all feasible y vectors. For each subproblem, we implemented an algorithm that combines a genetic algorithm (GA) with an ascent algorithm in the following fashion. The GA is used to generate solutions through random initialization, mutation, and crossover (see Appendix C). In each generation, after the genetic operations, an ascent algorithm is applied to each of the solutions in the solution pool for no more than five iterations with decreasing step size. The master algorithm stops if the gradient is sufficiently small or the maximum number of iterations is reached. Note that the ascent algorithm alone is sufficient to find local optima if the distributions are differentiable with support on (0, +∞). The GA component is added to encourage escaping from local optima in the search for a global optimum and to handle distributions that are not differentiable and/or have finite support.

Remark 3. Note that the objective function is not necessarily concave. For example, when the delay-time is deterministic and we are optimizing for only one perfect checkup, the concavity of the objective function is equivalent to the concavity of the probability density function of the time-to-develop the condition. However, under reasonable parameterizations in our numerical analysis, we found that our problem tends to have a unique optimum near the mode of the time-to-readmission curve (see Figures D1 and D2 in Appendix D). Hence, the first-order necessary conditions we analyze below provide strong intuition regarding the region of interest for scheduling checkups.

3.3. Moving Parts and Assumptions

Our model consists of three moving parts that require estimation. The estimation of these moving parts is crucial and challenging due to data scarcity and censoring. In this section, we discuss each of the moving parts and modeling assumptions surrounding them. Later in section 5, we discuss the estimation in detail and conduct sensitivity analysis

- Detection rate of imperfect checkups (r)
  The detection rate of an imperfect checkup is defined as the probability of detecting an existing condition. In our numerical analyses, we consider $r = 0.6$ for phone calls as a baseline and conduct sensitivity analyses by varying r between 0.2 and 1. In section 4.3, we analyze the impact of detection rate (r) by studying gamma g_δ distributions.
- Time-to-develop the condition distribution (pdf: g_δ)
  The time-to-develop the condition distribution is the probability density of developing a readmission-causing condition after discharge. In order to establish the First-order Necessary Condition, we require g_δ to be continuously differentiable with support on $[0, T]$. In section 4.1, we analyze the structure of the checkup timing assuming g_δ is unimodal. However, in section 4.2, the unimodality assumption is relaxed. In Appendix E, we test our model for robustness using multi-modal g_δ distributions.
- Delay-time distribution (pdf: f)
  The delay-time distribution is the probability density of the time between condition onset and readmission. We assume that the delay-time is independent of the time-to-develop the condition. In order to establish the First-order Necessary Condition, we assume f to be continuously differentiable with support on $[0, T]$. In section 4.3, we analyzed the impact of detection rate (r) by studying exponentially distributed delay-time. Table 1 shows the results of the sensitivity analyses using different delay-time distributions.

Our model also assumes that (i) the 30-day post-discharge mortality rates are small relative to 30-day hospital readmission rates and therefore can be neglected; (ii) the post-discharge checkup plan is not dynamically modified or updated; and (iii) the planning horizon is finite (i.e., 30 days).

4. Structural Properties

In this section, we analyze special cases to develop structural insights, which are extended to more general cases through numerical analyses in section 5. We first focus on the timing of checkups. Then we examine how different features such as stochastic delay-time, D, and different detection rates imply small modifications to the general timing structure. The analysis in sections 4.1–4.3 serves to develop intuition into rules of thumb that are combined to design a practical, implementable policy for providers/practices described in section 4.4, with each section providing a key building block. The overarching goal is to provide guidance toward a practical policy that is effective based only on historical data without relying on the optimization itself.

4.1. General Checkup Timing in Optimal Policies

We later show through numerical analyses (section 5) that checkup timing has the highest impact on detecting an ill patient, so we begin our analysis with this feature. To understand the general structure of checkup timing, we analyze the case of current practice where standard protocol dictates a single doctor's

Table 1 Optimal 2-Checkup for Exponential/Gamma Delay-Time Distributions with Different Mean and Variance: The Optimal Policies Outperform Current Practice by 43.7% (μ = 2.35)

Distribution					Detection probability		
Delay-time distribution	E[D]	Var[D]	Time of first checkup	Time between checkups	Optimal 2-checkup	Current practice	Relative improvement
exponential (μ/2)	1.2	1.4	4.9	3.1	0.13	0.08	56.8%
exponential (μ)*	2.4	5.5	5.9	4.4	0.23	0.16	43.7%*
exponential (2μ)	4.7	22.1	7.4	6.1	0.35	0.29	23.9%
gamma (1/2, 2μ)	2.4	11.0	6.5	5.0	0.20	0.15	30.8%
gamma (2, μ/2)	2.4	2.8	5.5	3.9	0.25	0.16	56.4%
gamma (3, μ/3)	2.4	1.9	5.3	3.6	0.26	0.16	62.1%
gamma (4, μ/4)	2.4	1.4	5.2	3.5	0.26	0.16	65.3%

Notes. *Marks the estimated delay-time distribution using our chart review dataset. The timing of checkup (rounded to the first decimal place) is in days. In our numerical studies, we observed that the solutions are insensitive to rounding of the checkup timing.

office visit (n = 1). We begin by assuming a deterministic delay-time, $D = z \geq 0$, and a perfect detection rate. We later generalize these analytical results through numerical analyses in section 5. The objective function for this special case can be rewritten as follows:

$$\max_{t_1} \mathbb{E}[N_1^{\Pi}] = \max_{t_1} \int_0^{t_1} g_\delta(k)(1 - F(t_1 - k))dk$$
$$= \max_{t_1} \int_{t_1 - z}^{t_1} g_\delta(k)dk \tag{6}$$

The second equality follows from the fact that the deterministic delay-time, $D = z \geq 0$, implies $F(t_1 - k) = 1$, if $t_1 - k \geq z$, and $F(t_1 - k) = 0$ otherwise.

Differentiating the objective function with respect to t_1 yields the following First Order Necessary Condition (FONC) for optimality

$$0 = \frac{\partial}{\partial t_1} \int_{t_1 - z}^{t_1} g_\delta(k)dk \Rightarrow g_\delta(t_1 - z) = g_\delta(t_1) \tag{7}$$

Based on results from our data on readmitted cystectomy patients, we also leverage the fact that the time-to-develop the condition of readmitted patients, $g_\delta(k)$, is unimodal. By unimodality of $g_\delta(k)$, the condition $g_\delta(t_1 - z) = g_\delta(t_1)$ implies that $(t_1 - z)$ is before the mode of $g_\delta(k)$ and t_1 is after the mode of $g_\delta(k)$. Thus, the probability density of developing a condition at $t_1 - z$ must equal the probability density of a condition developing at t_1. In practical terms, this informs decision-makers that, given only one checkup opportunity, they should schedule the checkup a little bit ($<z$) after the time when conditions are most likely to develop.

Next, consider a more aggressive approach with n checkups. The following proposition shows that the general multivariate optimization can be transformed into a univariate optimization, focused only on the time of the first checkup. The proposition indicates the best way to achieve maximum coverage of high risk times in a patient's post-discharge recovery. Specifically, we want our checkups to cover as much of the period of time when the patient is at highest risk of having a readmission-causing condition as possible. This results in the following two insights. First, if checkups are too close (i.e., spaced closer than z time units), there is unnecessary overlap in the coverage (i.e., two checkups covering the same time period). Better coverage can be achieved by spacing them further apart without any loss in detection (since delay-time is deterministic). Second, we want the checkups to cover the high-risk period (i.e. the time window containing the highest time-to-develop the condition density), hence it is best to center all of the checkups around mode of the time-to-develop the condition distribution, since the density is decreasing monotonically on either side of the mode.

Proposition 1. *If the delay-time is deterministic* ($D = z$ *with probability* 1) *and the time-to-develop the condition* g_δ *is unimodal, then* (*i*) *it is sufficient to optimize* t_1 *only;* (*ii*) *the checkups are spaced* z *days apart equidistantly; and* (*iii*) *the densities of developing the condition are equal at* $t_1 - z$ *and* t_n

$$\max_{t_1,\ldots,t_n} \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi}] = \max_{t_1} \int_{t_1 - z}^{t_1 + (n-1)z} g_\delta(k)dk \tag{8}$$

$$\text{s.t. } g_\delta(t_1 - z) = g_\delta(t_1 + (n-1)z) \tag{9}$$

$$t_{i+1} = t_i + z, \quad \forall\, i \in \{1, \ldots, n-1\} \tag{10}$$

Proof. We first show that $(t_n - t_1) = (n - 1)z$. In other words, the time between the first and last checkups is exactly $(n - 1)z$.

The structure of our objective function appropriately avoids double counting the detection of conditions. To see how, notice that under the assumptions of deterministic delay-time ($D = z$) and perfect detection rates, the objective function in Equation (2) becomes

$$\sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} g_\delta(k)\bar{F}(t_i - k)dk = \sum_{i=1}^{n} \int_{\max(t_{i-1}, t_i - z)}^{t_i} g_\delta(k)dk \tag{11}$$

Thus, only the earliest successful checkup contributes a positive amount to the objective function. For example, if a condition was present during a time interval $(\delta, \delta + D)$ and three checkups were scheduled at some arbitrary times t_i, t_j, $t_k \in (\delta, \delta + D)$, then only the checkup at $\min\{t_i, t_j, t_k\}$ contributes a positive amount to the objective function.

This implies that an optimal solution must be such that the intervals $(t_i - z, t_i)$ are disjoint for all i. To see why, consider an arbitrary checkup schedule that has non-disjoint intervals. Suppose the smallest index corresponding to non-disjoint intervals is $j < n$ such that $(t_j - z, t_j)$ and $(t_{j+1} - z, t_{j+1})$ are non-disjoint. Then, $t_j = t_{j+1} - z + \gamma$ with $\gamma \in (0, z)$. We can construct another solution that is strictly better, by increasing t_{j+1} by $(z - \gamma)$. This increases the objective value by a non-negative amount:

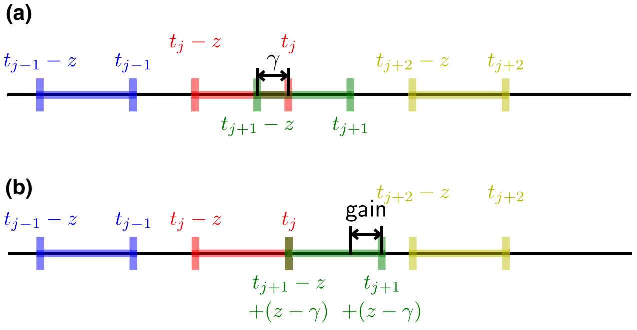
$$\begin{cases} \int_{t_{j+1}}^{\min(t_{j+2} - z, t_{j+1} + z - \gamma)} g_\delta(k)dk, & \text{if } j \leq n - 2 \\ \int_{t_{j+1}}^{t_{j+1} + z - \gamma} g_\delta(k)dk, & \text{if } j = n - 1. \end{cases} \tag{12}$$

If $j = n - 1$, then the change in the objective value is strictly positive. Similarly, if $j \leq n - 2$ and the upper limit of the integral in Equation (12) is $t_{j+1} + z - \gamma$, the change in the objective value is strictly positive and the adjustment of t_{j+1} leaves the intervals $(t_{j+1} - z, t_{j+1})$ and $(t_{j+2} - z, t_{j+2})$ disjoint. The last case we need to consider is if $j \leq n - 2$ and the upper limit of the integral in Equation (12) is $t_{j+2} - z$. In this case, there is a non-negative change in the objective value and the intervals $(t_{j+1} - z + (z - \gamma), t_{j+1} + (z - \gamma))$ and $(t_{j+2} - z, t_{j+2})$ become non-disjoint, so we can repeat the steps above. This process terminates in finite iterations and results in a strictly positive change in the objective value. Thus, we can conclude that an optimal solution must satisfy $(t_n - t_1) \geq (n - 1)z$. Figure 2 illustrates how Equation (12) is derived in the case of $j \leq n - 2$ and $t_{j+2} - z \geq t_{j+1} + z - \gamma$.

We will now argue that an optimal solution cannot have $(t_n - t_1) > (n - 1)z$. Combining this with our previous finding yields our desired result that an optimal solution must satisfy $(t_n - t_1) = (n - 1)z$. If $(t_n - t_1) > (n - 1)z$, then $\exists\ i \in \{1, \ldots, n - 1\}$ such that $t_{i+1} - t_i = z + \gamma$, with $\gamma > 0$. In other words, there is at least one pair of consecutive checkups that are spaced farther than z apart. A checkup schedule with this property is necessarily suboptimal because the objective value can be improved by adjusting either t_i or t_{i+1} (without changing any other checkup times), depending on their relative positions to the mode of $g_\delta(\cdot)$.

In particular, if $t_i < t_{i+1} \leq$ mode of $g_\delta(\cdot)$, we can increase the objective value by shifting the checkup i from t_i to $t_i + \epsilon$, where $\epsilon \in (0, \gamma]$. This increases the objective value by $\int_{t_i}^{t_i+\epsilon} g_\delta(k)dk - \int_{\max(t_i - z, t_{i-1})}^{\max(t_i+\epsilon-z, t_{i-1})} g_\delta(k)dk$. Observe that the second term is integrated over the interval $[\max(t_i - z, t_{i-1}), \max(t_i + \epsilon - z, t_{i-1})]$, which has length $\leq \epsilon$. Since the second integral interval is to the left of the first integral interval, and these two intervals are to the left of the mode, it follows that $\int_{t_i}^{t_i+\epsilon} g_\delta(k)dk - \int_{\max(t_i - z, t_{i-1})}^{\max(t_i+\epsilon-z, t_{i-1})} g_\delta(k)dk > 0$. By symmetry, if $t_i > t_{i+1} \geq$ mode of $g_\delta(\cdot)$, we can shift checkup $i + 1$ from t_{i+1} to $t_{i+1} - \epsilon$, where $\epsilon \in (0, \gamma]$, to achieve a non-negative improvement. If $t_i <$ mode of $g_\delta(\cdot) < t_{i+1}$, we can achieve a non-negative improvement by moving t_i to the right (if $g_\delta(t_i) \geq g_\delta(t_{i+1})$) or moving t_{i+1} to the left (if $g_\delta(t_i) < g_\delta(t_{i+1})$). The improvement is strictly positive if g_δ is strictly unimodal, that is, has a unique mode.

Figure 2 Schematic Sketch for Equation (12) [Color figure can be viewed at wileyonlinelibrary.com]



We can now conclude that an optimal solution must satisfy $(t_n - t_1) = (n - 1)z$. Given our previous result that an optimal solution must have checkup times such that the intervals $(t_i, t_i + z) \forall\ i$ are disjoint, this implies that an optimal solution must be of the form $t_i = t_{i-1} + z,\ \forall\ i \in \{2, \ldots, n\}$. This is equivalent to letting $t_i = t_1 + (i - 1)z,\ \forall\ i \in \{2, \ldots, n\}$. Note that this only holds assuming the delay-time is deterministic. This proves that $\max_{t_1} \int_{t_1 - z}^{t_1 + (n-1)z} g_\delta(k)dk$ is in fact optimal. □

Remark 4. If the distribution of the time-to-develop the condition is right/left skewed (yet still unimodal), this does not affect our optimality results at all, since our results assume nothing about the skewness of the curve. The checkups would still be centered around the mode, even though the mode will be later/sooner in the 30-day readmission window. If the distribution is not unimodal, then alternative optima might exist. Nonetheless, some of the properties from Proposition 1 still hold. For example, under the assumptions of bimodal distribution and deterministic delay-time, we know the following: (i) if there was only one checkup to place, Proposition 1 still holds; (ii) if there were multiple checkups, checkups are placed no closer than z days apart (might be farther than z days apart depending on the shape of the bimodal curve). For the general case with multiple modes, the First Order Necessary Conditions still hold and the problem can still be solved numerically.

From Proposition 1, we see that the problem effectively becomes the single checkup problem while letting $D = nz$. Thus, an optimal solution in the case of perfect inspection checkups and deterministic delay-times must satisfy the following conditions

$$g_\delta(t_1 - z) = g_\delta(t_1 + (n - 1)z) \tag{13}$$

$$t_{i+1} = t_i + z, \quad \forall\ i \in \{1, \ldots, n - 1\} \tag{14}$$

Reducing the n-dimensional optimization problem to a univariate optimization problem makes these conditions especially valuable because these univariate optimizations are easy to solve using ascent search or binary search even without specialized computer software. This can be achieved by solving the univariate FONC equation (which is in the form of $\psi(t_1) = 0$) using binary search since $g_\delta(t_1 - s) - g_\delta(t_1 +$

$(n-1)z$) is monotone increasing for a unimodal function), ascent search, or Newton's method. Furthermore, the conditions imply that an optimal policy schedules one contiguous block of checkups with the checkups collectively covering a time of length nz. Practically speaking, this informs decision-makers that if they have n perfect checkups (e.g., doctors' office visits), then the checkups should be scheduled surrounding the time when conditions develop most frequently such that there are z (delay-time) time units between each checkup.

4.2. Effect of Stochastic Delay-Time on Optimal Checkup Timing

Proposition 1 gives us the block structure of an optimal checkup policy with deterministic delay-time, D. In this section, we investigate how stochastic D affects the spacing of checkups within the block of checkups. First, relaxing the assumption that $D = z$, the objective function becomes

$$\max_{t_1,\ldots,t_n} \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} g_\delta(k)[1 - F(t_i - k)]dk \tag{15}$$

which for $n = 1$ equals $\int_0^{t_1} g_\delta(k)[1 - F(t_1 - k)]dk$, resulting in the following FONC:

$$\begin{aligned} 0 = \frac{\partial}{\partial t_1}\int_0^{t_1} g_\delta(k)[1 - F(t_1 - k)]dk \Rightarrow g_\delta(t_1) \\ = \int_0^{t_1} g_\delta(k) f(t_1 - k)dk = g_\rho(t_1) \end{aligned} \tag{16}$$

Notice that the RHS of Equation (16) is the formula for the probability density associated with a readmission occurring at t_1. This implies that at an optimal t_1, the marginal rate of developing a condition (i.e., the marginal increase in patients who could be detected if t_1 was increased) is equal to the marginal rate of a readmission occurring (i.e., the marginal lost patients that would be readmitted if t_1 was increased). Both results extend our intuition from section 4.1 to the case of stochastic delay-time.

Generalizing the FONC to an arbitrary number of checkups yields

$$\begin{aligned} \int_{t_{i-1}}^{t_i} g_\delta(k) f(t_i - k)dk = g_\delta(t_i)F(t_{i+1} - t_i), \\ \forall i \in \{1, \ldots, n-1\} \end{aligned} \tag{17}$$

$$\int_{t_{n-1}}^{t_n} g_\delta(k) f(t_n - k)dk = g_\delta(t_n) \tag{18}$$

The intuition behind these equations is similar to when $n = 1$ in that the optimal solution balances the marginal rate of catching a condition with the ith checkup with the marginal rate of missing a later condition. The LHS of Equation (17) is the probability of checkup i detecting a condition developed between t_{i-1} and t_i. Since the perfect checkup at t_{i-1} ensures t_i will only detect conditions between t_{i-1} and t_i, the LHS of Equation (17) can be thought of as the marginal benefit of moving inspection i slightly to the right from t_i to $t_i + \epsilon$ (as $\epsilon \to 0^+$), and therefore capturing more conditions that could have developed between t_i and $t_i + \epsilon$. This is essentially the marginal opportunity cost. The RHS of Equation (17) is the probability of t_{i+1} missing the condition developed after t_i. This is analogous to lost sales, in that it represents the marginal rate of patients developing a condition at t_i and being readmitted before the next inspection at t_{i+1}.

Rearranging the terms of Equation (17) implies the timing between inspections follows a newsvendor-type solution:

$$t_{i+1} - t_i = F^{-1}\left(\frac{\int_{t_{i-1}}^{t_i} g_\delta(k) f(t_i - k)dk}{g_\delta(t_i)}\right) \tag{19}$$

The structure of Equation (19) closely resembles the equation for the optimal stocking quantity in traditional newsvendor problems. This highlights the inherent tradeoff between (i) scheduling checkups closer together to increase the likelihood of detecting illnesses that develop between the checkups and (ii) scheduling checkups farther apart to have the opportunity to detect more illnesses by covering a wider span of time. Both of these tradeoffs are inherently linked to the density of the delay-time function, F. Thus, the distance between any two checkups is determined by a solution where the delay-time density functions as the demand function.

It is worth noting that one can construct a recursive algorithm to solve the optimization in light of Equation (19). For instance, given $t_0 = 0$ and t_1, one can determine $t_2 = t_1 + F^{-1}\left(\frac{\int_{t_0}^{t_1} g_\delta(k) f(t_1 - k)dk}{g_\delta(t_1)}\right)$. Recursively, one can determine $t_3, \ldots, t_n$. This reduces the problem to a univariate optimization where t_1 is the only decision variable. Moreover, an optimal solution must exist since we are maximizing a continuous function over a compact set. For the general case with stochastic delay-time and imperfect checkups, our solution procedure utilizes this recursive construction to generate the initial solution seeds (see Appendix C).

If the solution to the FONCs is not unique, then one can solve the following univariate maximization to generate the optimal checkup policy.

$$\max_{t_1} \quad \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi(t_1)}] \tag{20}$$

$$\text{s.t. } t_1 \in [0, T] \tag{21}$$

In this optimization problem, the checkup policy $\Pi(t_1)$ is drawn from a set of potential candidates based on the FONCs:

$$\Pi(t^1) = [t_1, t_2, \ldots, t_n]^T \tag{22}$$

$$= \begin{bmatrix} t_1 \\ t_1 + F^{-1}\left(\frac{\int_{t_0}^{t_1} g_\delta(k) f(t_1 - k)dk}{g_\delta(t_1)}\right) \\ t_2 + F^{-1}\left(\frac{\int_{t_1}^{t_2} g_\delta(k) f(t_2 - k)dk}{g_\delta(t_2)}\right) \\ \vdots \\ t_{n-1} + F^{-1}\left(\frac{\int_{t_{n-2}}^{t_{n-1}} g_\delta(k) f(t_{n-1} - k)dk}{g_\delta(t_{n-1})}\right) \end{bmatrix} \tag{23}$$

Remark 5. The analyses in this section are based on the KKT conditions, which assume (i) g_δ has support on [0, T]; (ii) f has support on [0, ∞); and (iii) g_δ and f are continuously differentiable. It is worth highlighting that these results do not require unimodality.

4.3. Effect of Imperfect Inspection Checkups on Optimal Checkup Timing

As previously mentioned, hospitals have various checkup methods available with differing detection rates. Hence, it is valuable from both a practical and a theoretical perspective to understand how the optimal timing of checkups is affected by the detection rates of the checkups. For the purpose of exposition, we let $r_{(i)}$ $\forall i \in \{1, \ldots, n\}$ denote the detection rate of the checkup method employed at time t_i. We begin by considering the case where $n = 2$ and $r_{(1)} = r_{(2)} = r$. This yields the following objective value

$$r\int_0^{t_1} g_\delta(k)[1 - F(t_1 - k)]dk + (1 - r)r \int_0^{t_1} g_\delta(k)[1 - F(t_2 - k)]dk + r\int_{t_1}^{t_2} g_\delta(k)[1 - F(t_2 - k)]dk \tag{24}$$

We can then derive FONCs as follows

$$\int_0^{t_1} g_\delta(k)f(t_1 - k)dk = g_\delta(t_1)(F(t_2 - t_1) + (1 - r)(1 - F(t_2 - t_1))) \tag{25}$$

$$(1 - r)\int_0^{t_1} g_\delta(k)f(t_2 - k)dk + \int_{t_1}^{t_2} g_\delta(k)f(t_2 - k)dk = g_\delta(t_2) \tag{26}$$

The intuition behind these equations is similar to the perfect checkup case in Equations (17) and (18). The LHS of Equation (25) is the probability density of detecting a condition that developed between 0 and t_1, that is, marginal rate of gain in terms of detection. The RHS of Equation (25) is the marginal density of missing a condition developed after t_1, that is, loss sales. To see this, note the term $g_\delta(t_1)F(t_2 - t_1)$ appears and has the same intuition as in Equation (17), that is, the condition developed after t_1 but the patient was readmitted before t_2. However, the inspection at t_2 could also miss an extant condition due to the imperfect detection. This event is captured by the term $g_\delta(t_1)(1 - r)(1 - F(t_2 - t_1))$, which implies the condition was detectable at time t_2 but failed to be detected. Equation (26) represents the tradeoff between lost sales (RHS) and marginal change in detection (LHS). The RHS of Equation (26) is the marginal density of a condition developing at time t_2, that is, lost sales as before since any conditions developing after t_2 will not be detected. The first term on the LHS is the density of a condition being detectable at time t_2 that developed on 0 to t_1 and was missed by the inspection at time t_1, that is, the marginal change in detection for conditions missed by the first inspection. The second term on the LHS is the probability density of detecting a condition that developed between t_1 and t_2.

Using the FONCs, we next show that as r increases, the two checkups move farther apart. Hence, by improving the detection rate of a particular method, the doctors should place inspections farther apart and can cover a larger time period in which to catch potentially developing conditions. The intuition behind this is that with a poor detection rate, a subsequent inspection can catch a condition that was previously missed if placed closer to the previous inspection. This comes at the expense of covering less overall timespan, as placing this inspection earlier will miss the opportunity to catch later developing conditions. As the detection rate increases, however, there is a smaller benefit of catching conditions missed by a previous inspection, since fewer patients are missed the first time.

For the analysis, let t_1^* and t_2^* be the optimal values of t_1 and t_2, respectively. To show this property analytically, we first introduce an inequality that relates the probability densities of developing the condition and readmission.

Definition 1. *Assuming g_ρ and g_δ are differentiable, the delayed readmission log-likelihood inequality at time t is defined as* $\frac{d}{dt}\log g_\delta(t) = \frac{g_\delta'(t)}{g_\delta(t)} \leq \frac{d}{dt}\log g_\rho(t) = \frac{g_\rho'(t)}{g_\rho(t)}$.

This inequality states that, at time t, the derivative of the log-likelihood of developing the condition is less than or equal to the derivative of the log-likelihood of readmission. This is similar to previous results we have seen relating the density functions of time-to-develop the condition, δ, and time-to-readmission, ρ. The following remark shows that this condition holds for Erlang and exponential distributions. The condition has been verified numerically for other distributions we use in our numerical studies (see Table 1 in section 5). As we shall see in our numerical analyses, the shape of Erlang distributions resembles the observed time-to-develop the condition, and an exponential distribution is actually the best fit distribution for the delay-time.

REMARK 6. If the time-to-develop the condition follows an Erlang distribution with scale μ and shape parameter k_δ (Erlang(k_δ, μ)), and the delay-time follows Erlang(k_D, μ), then the time-to-readmission follows an Erlang(k_ρ, μ) where $k_\rho = k_\delta + k_D$. The delayed readmission log-likelihood inequality becomes $(k_\delta - 1)t^{-1} \leq (k_\rho - 1)t^{-1}$, which holds $\forall t > 0$.

The following Lemma (proved in Appendix F) shows that, as the detection rate increases, the first inspection will be placed closer to the patient's time of discharge (i.e., moved earlier).

LEMMA 1. *If the delayed readmission log-likelihood inequality holds, then t_1^* decreases in r.*

Leveraging Lemma 1, we next show that the gap, $t_1^* - t_2^*$, widens as r increases. Notice that the optimal timing t_1^* and t_2^* is the solution to the FONCs, that is, Equations (25) and (26). For general delay-time and time-to-develop the condition distributions, the FONCs are essentially a set of integral equations without a closed form solution. In the following theorem, we consider the case where the delay-time is exponential and the time-to-develop the condition is Erlang so that the time-to-readmission is in closed form since the convolution of exponential and Erlang distributions is an Erlang distribution. The structure and shape of the Erlang and exponential distributions are close to what is observed in practice through our numerical analyses (see Figure 4 in section 5.1). With exponential-Erlang distributions, Equation (26) effectively becomes a polynomial where t_1^* can be directly expressed in closed form.

Theorem 1 now shows, for the case of Erlang and exponential densities for δ and D, that the two tests move farther apart as the detection rate increases. This result is later generalized in our numerical study.

THEOREM 1. *If the time-to-develop the condition follows Erlang(k, μ) and the delay-time follows exponential(μ), then $t_2^* - t_1^*$ strictly increases in r.*

PROOF. We begin with the following technical lemma, which is proved in Appendix F.

LEMMA 2. *If the delayed readmission log-likelihood inequality holds, then $\frac{g_\rho(t)}{g_\delta(t)}$ increases in t.*

Without loss of generality, assume $\mu = 1$. For $\mu \neq 1$, the problem can be scaled. We then rewrite Equation (26) as follows:

$$\begin{aligned} g_\rho(t_2^*) - g_\delta(t_2^*) &= r\int_0^{t_1^*} g_\delta(s)f(t_2^* - s)ds \\ \Leftrightarrow \frac{e^{-t_2^*}{t_2^*}^k}{k!} - \frac{e^{-t_2^*}{t_2^*}^{k-1}}{(k-1)!} &= r\int_0^{t_1^*}\frac{e^{-t_2^*}s^{k-1}}{(k-1)!}ds \end{aligned} \tag{27}$$

$$\frac{e^{-t_2^*}{t_2^*}^k}{k!} - \frac{e^{-t_2^*}{t_2^*}^{k-1}}{(k-1)!} = r\frac{e^{-t_2^*}{t_1^*}^k}{k!} \Leftrightarrow {t_2^*}^k - k{t_2^*}^{k-1} = r{t_1^*}^k \tag{28}$$

$$\Leftrightarrow t_1^* = \left(\frac{{t_2^*}^k - k{t_2^*}^{k-1}}{r}\right)^{\frac{1}{k}} \tag{29}$$

The first and second derivatives of t_1^* with respect to t_2^* are

$$\begin{aligned} \frac{\partial t_1^*(t_2^*)}{\partial t_2^*} &= \frac{(t_2^* - k + 1)\left(\frac{{t_2^*}^{k-1}(t_2^* - k)}{r}\right)^{\frac{1}{k}}}{t_2^*(t_2^* - k)} \text{ and} \\ \frac{\partial^2 t_1^*(t_2^*)}{\partial {t_2^*}^2} &= -\frac{(k-1)\left(\frac{{t_2^*}^{k-1}(t_2^* - k)}{r}\right)^{\frac{1}{k}}}{{t_2^*}^2(t_2^* - k)^2} \end{aligned} \tag{30}$$

Based on the first and second derivatives, we show that $t_1^*(t_2^*)$ has the following properties: (i) t_1^* strictly increases in t_2^*; (ii) $t_1^*(t_2^*)$ is concave; (iii) $\lim_{t_2^* \to +\infty} \frac{\partial t_1^*(t_2^*)}{\partial t_2^*} = (1/r)^{\frac{1}{k}} > 1$; and (iv) $\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} > 1$, $\forall t_1^*$, t_2^*.

For (i), notice that $(t_2^* - k)$ has to be strictly positive for $t_1^* \in \mathbb{R}^+$. Hence, $\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} > 0$, which implies t_1^* strictly increases in t_2^*. For (ii), since $(t_2^* - k) > 0$, it is clear that $\frac{\partial^2 t_1^*(t_2^*)}{\partial {t_2^*}^2} < 0$ for $k > 1$, integer. Hence $t_1^*(t_2^*)$ is concave. To see (iii), for $k > 1$ and $r \in (0, 1)$, we have

$$\begin{aligned} &\lim_{t_2^* \to +\infty} \frac{\partial t_1^*(t_2^*)}{\partial t_2^*} \\ &\quad = \lim_{t_2^* \to +\infty} \frac{(t_2^* - k + 1)(\frac{{t_2^*}^{k-1}(t_2^* - k)}{r})^{\frac{1}{k}}}{t_2^*(t_2^* - k)} \\ &\quad > \lim_{t_2^* \to +\infty} \frac{(t_2^* - k)({t_2^*}^{k-1}(t_2^* - k))^{1/k}}{t_2^*(t_2^* - k)}\left(\frac{1}{r}\right)^{1/k} \\ &\quad = \lim_{t_2^* \to +\infty}\left(\frac{{t_2^*}^k - k{t_2^*}^{k-1}}{{t_2^*}^k}\right)^{1/k}\left(\frac{1}{r}\right)^{1/k} \\ &\quad = \lim_{t_2^* \to +\infty}\left(1 - \frac{k}{t_2^*}\right)^{1/k}\left(\frac{1}{r}\right)^{1/k} = \left(\frac{1}{r}\right)^{1/k} > 1. \end{aligned}$$

(iv) follows from Properties 2 and 3. Given the four properties above, Figure 3 sketches $t_1^*(t_2^*)$ schematically.

Consider optimal t_1^* and t_2^* with detection rate r. As r increases, t_1^* decreases (Lemma 1). By property 1, t_2^* also decreases. Denote the new optimal solution as t_1^{**} and t_2^{**}. As shown in Figure 3, since the slope of $t_1^*(t_2^*)$ is always strictly greater than one, it follows that $t_2^* - t_2^{**} < t_1^* - t_1^{**}$. Therefore $t_2^{**} - t_1^{**} > t_2^* - t_1^*$ as desired, which completes our proof. □

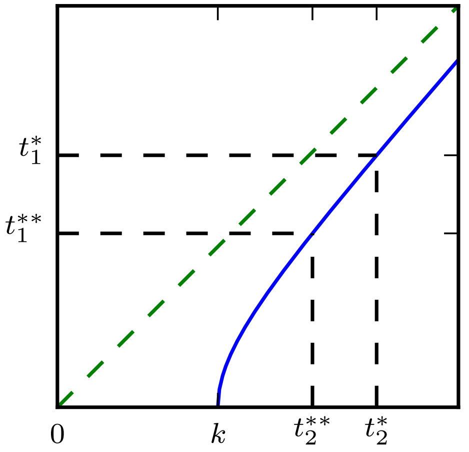
REMARK 7. Under the assumptions of Theorem 1, if the detection rate changes from r to $r + \epsilon$, ($\epsilon > 0$), then the increase in the gap between the two checkups is bounded above by $1 - r$ (if $k = 1$) or $2(r + \epsilon)k$ (if $k \geq 2$). Please see Appendix F for the proof.

In practical terms, checkups should be placed farther apart as the detection rates improve. This is because when the detection rate is relatively low, there is a benefit to scheduling checkups that "overlap" each other in case a checkup fails to detect an existing illness. However, this benefit diminishes as the detection rate improves, so the checkups spread farther apart from one another. This allows the checkup schedule to cover a wider range of potential readmissions without losing detection quality.

4.4. From Theory to Practice: Implementable Policies from Modeling Insights

Through the prior analysis, we have captured the key factors affecting the efficacy of post-discharge checkup policies. To summarize the analytical insights of the previous section into practical rules of thumb, we now illustrate how to design a simple checkup policy for doctors and discharge planners. Suppose a patient is to be discharged and a post-discharge follow-up plan needs to be determined by the case manager. The case manager first decides the aggressiveness of the follow-up plan, that is, how many office visits and phone calls to use. This can be done by evaluating the patient's readmission risk using existing risk calculators (Hu et al. 2014). Given the estimates of the time-to-develop the condition density curve and the delay-time D (later in section 5.1 we estimate the densities using historical data), the next step is to determine the timing of checkups.

Figure 3 Schematic Sketch of t_1^* as a Function of t_2^* [Color figure can be viewed at wileyonlinelibrary.com]



From the analyses in sections 4.1 and 4.2 and Proposition 1, the checkups should be placed approximately z days apart (z being the average delay-time) such that the first and the last checkups are at the same height on the time-to-develop the condition curve (one on either side of the mode). Finally, from Theorem 1, the case manager adjusts the spacing of checkups according to the detection rate of the checkups: higher detection rate spreads the checkups farther apart. For instance, the case manager should make less frequent contact with the patient if he/she believes that the patient was well educated for the diagnosis and understands what post-operative complications might happen (this translates to a higher detection rate); or the case manager may want to make frequent contact if he/she believes that the patient is less responsive to phone calls or is less adherent to the follow-up appointments (this translates to a lower detection rate). In the next section, we generalize the analytical insights using numerical studies to deepen the understanding of how to empirically estimate model parameters, of the impact of office visit and phone call sequencing, and of quantity vs. quality of checkups.

5. Numerical Analyses

In this section, we conduct extensive numerical analyses on cystectomy readmissions from a regional hospital as well as the national SID to address the key questions that arise in post-discharge checkup policies: when to schedule checkups, how many checkups to schedule, and what types of checkups to schedule. First, we study two-checkup policies with one phone call and one office visit, which are consistent with current practice at our partner hospitals. We show that our approach improves the detection probability upon current practice by up to 43.7% when applied to readmitted patients. We test the robustness of our model with different exponential and gamma delay-time distributions. We also verify the delayed readmission log-likelihood inequality defined in section 4.3. Next, we examine more aggressive checkup plans with more checkups to develop insights into: (i) optimal checkup timing and sequencing, (ii) effects of varying the detection rate, and (iii) checkup quantity vs. quality. We then validate our work by applying the optimal policies found to a different subset of patients and show that our results continue to hold. We conclude this section by summarizing rules of thumb that can be easily implemented by healthcare professionals to develop post-discharge checkup policies that have the potential to improve detection of readmission causing conditions.

5.1. Data and Model Parametrization

The numerical analyses in this section are based on two datasets. The first dataset contains delay-time information of 327 cystectomy patients discharged from our partner hospital between 2007 and 2012. The information in the dataset includes the following: date of discharge from the hospital, date of first contact with the healthcare provider after discharge, who initiated the contact, what the chief complaint was, date of readmission, what condition caused the readmission, and when the condition was first experienced. By computing the difference between the date of readmission and date of condition onset, we obtain the delay-time for each patient in this cohort. The data were manually collected by a medical student and a medical fellow at our partner hospital by going over medical charts and reviewing each patient's triage notes upon readmission. This patient cohort consisted of 79 female and 248 male patients between 37 and 91 years old (mean = 65.9, standard deviation = 11.2). Among the 327 patients, 63 patients (19%) were readmitted within 30 days of discharge. We used this database to obtain data on the delay-time random variable and the time-to-develop the condition random variable. Note that we focus on the readmitted patients only and exclude the patients who were not readmitted from our analysis. We also ignore the intervention and prevention effect of the checkups a patient received, which, at our partner hospital, typically included a phone call and a follow-up office visit on the 2nd and 12th day after discharge respectively.

We acknowledge there are many empirical challenges with this type of data and we do not address them all in this study. One of the key challenges is the estimation of the distributions. Since we only used readmitted patients in our estimation, it is likely that the estimated distributions differ from the ones parameterized using all patients, including readmitted and non-readmitted patients. In addition, since we ignored the intervention and prevention effect of existing checkups, our estimated distributions could be biased. In Appendix G, we provide an initial approach addressing how incorporating both readmitted and non-readmitted patients might affect our model's performance. Notice that results presented in that appendix are obtained from a limited case study on a very specific dataset. Nevertheless, empirical estimation is not the primary focus of our study and the remaining empirical challenges are left to future work. Notice that, to the best of our knowledge, this is the first study in the clinical or operational literature to attempt to characterize these two variables using actual data. This is because existing available datasets do not capture delay-time or time when a readmission-causing condition developed. Due to data scarcity, we conducted our numerical analysis using population-based distribution curves. Given sufficient delay-time data, our approach can be tailored to individual patients by applying transfer learning techniques for personalized readmission forecasting (Helm et al. 2016). We demonstrate robustness of our optimal policies to distribution in Table 1 and the analytical results from section 4 are not dependent on the form of the delay-time distribution. Further, the mean of the delay-time distribution observed in the data (2.35 days) is very close to delay-time estimates for common readmission-causing conditions in a survey given to an independent group of five practicing surgeons (average of 2 days). These cross-checks should help mitigate some concerns about the accuracy of the estimation. We also tested the dependency between delay-time and the time-to-develop the condition using the 63 readmitted patients from this new dataset. The correlation between the two variables is 0.14, and they are independent ($p < 0.05$) using the Hilbert–Schmidt independence criterion (Gretton et al. 2008). While data for this study was collected manually as a proof of concept, this process could be appropriately scaled with IT support due to the proliferation of electronic health records. This type of analysis, however, is left to future work.

The second dataset comes from the SID, which was gathered as part of the Healthcare Cost and Utilization Project sponsored by the Agency for Healthcare Research and Quality. From the SID dataset, we identified 717 cystectomy patients (ICD-9 code 577, 5771, and 5779) from the states of Florida, Iowa, North Carolina, New York, and Washington that were readmitted within 30 days of discharge in 2009 and 2010. As mentioned in section 1, we choose cystectomy patients as a proof of concept given that our clinical collaborator is an expert in this type of surgery and that it has one of the highest readmission rates in the U.S. Note that subsequent work by our collaborator's surgical research group indicates the dynamics of cystectomy are similar to many other surgeries, particularly lower torso/abdomen surgeries (Jacobs et al. 2017), and our clinical collaborator believes this approach would be broadly applicable in the surgery domain; this includes surgeries targeted for inclusion in Medicare's readmission penalty program (HRRP). To further verify that the unimodality assumption holds for other surgery cohorts, we extracted the readmission records of patients who had some of the most common abdominal and chest surgeries in 2009 and 2010: Abdominal Aortic Aneurysm Repair (AAA), Esophagectomy, Pancreatectomy, Aortic Calve Replacement (AVR), Coronary Artery Bypass Grafting (CABG), and Lung Resection. In all six cases, the time-to-readmission and the estimated time-to-develop the condition curves (estimated using readmitted patients) appeared to be unimodal (see Appendix H).

We excluded patients who had ICD-9 code 4411, 4412, 4413, 4415 or 4416, patients who were 18 years old or younger, and patients who died during cystectomy or during their inpatient stay. The SID database captures the length of time between each patient's initial discharge and his/her subsequent readmission. Among the 717 patients, 385 patients from 2010 were used for parametrization and optimization of the models, and 332 patients from 2009 were used to test the optimal policies. We used the first dataset to estimate the delay-time distribution and to validate the efficacy of recovering the time-to-develop the condition distribution. To do that, we started by fitting distributions to the observed time-to-readmission (shown in Figure 4a) and to the observed delay-time (shown in Figure 4b). Gamma and exponential distributions worked well to model the time-to-readmission and the delay-time, respectively.

Given the time-to-readmission and the delay-time distributions, we recovered the time-to-develop the condition distribution through a numerical inverse Laplace transform (see Appendix I). The numerical Laplace inversion fitted the true time-to-develop the condition well with a Pearson χ^2 p-value $= 0.36$. This validates the efficacy of recovering the distribution of the time-to-develop the condition using inverse Laplace transform.

With an effective approach to recover the time-to-develop the condition, we expanded our analysis to the SID database (which includes patients from many hospitals across five states). Using the 2010 SID patients, we fitted a gamma distribution to the time-to-readmission as shown in Figure 5. Since the delay-time information was not recorded on the SID database, we assumed that the delay-time for the SID patients followed the same distribution as the delay-time observed on patients at our partner hospital (exponential(2.35)). We used the inverse Laplace transform to estimate the time-to-develop the condition distribution (see Figure 5).

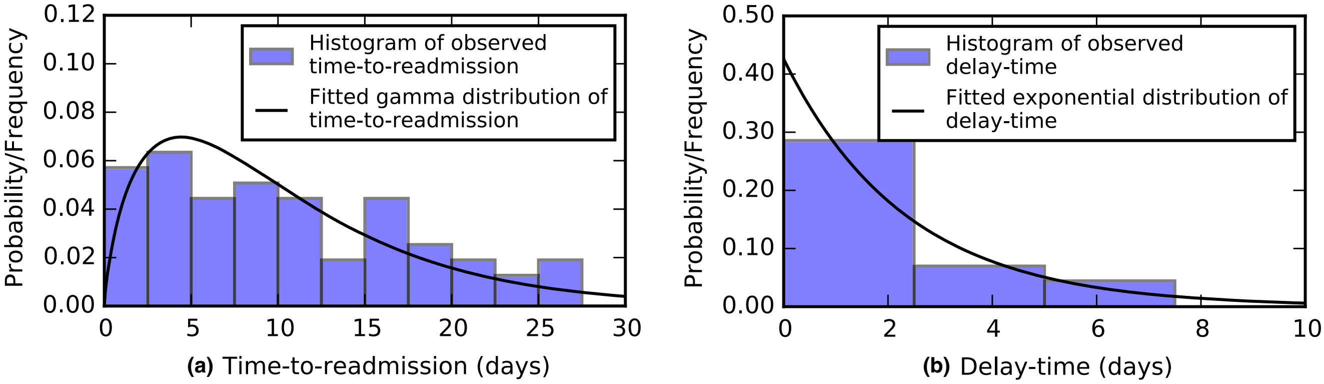
5.2. Comparison of Policies Against Current Practice

With the model parameterized on the 2010 SID patients, we evaluated how our policy improves upon the current practice at our partner hospital. We also examined the robustness of our model by fitting various exponential and gamma delay-time distributions (see Table 1). The distributions tested in Table 1 satisfy the delayed readmission log-likelihood inequality defined in section 4.3.

The current practice for post-discharge monitoring at our partner hospitals is to place a phone call on the 2nd day after discharge and an office visit on the 12th day after discharge. Throughout our numerical analyses, we assume that an office visit is a perfect inspection with detection rate $r = 1$; and a phone call is an imperfect inspection with detection rate $r = 0.6$ (given the patient has developed a condition, a phone call will detect the condition successfully with probability 0.6). These values were estimated by our clinical collaborators. In section 5.4, we perform a sensitivity analysis on the detection rate.

Applying the algorithm described in section 3.2, we solve for the optimal 2-checkup policies with one phone call and one office visit (for fair comparison with current practice) using the 2010 SID patients. We

Figure 4 Time-to-Readmission and Delay-Time Distribution Fitted from Medical Charts [Color figure can be viewed at wileyonlinelibrary.com]

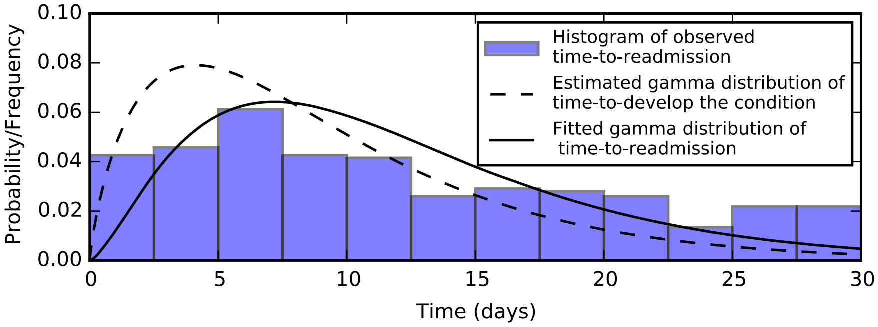


(a) Time-to-readmission (days)

(b) Delay-time (days)

Note. Time-to-readmission $\rho \sim$ gamma(1.74, 5.98), delay-time $D \sim$ exponential(2.35).

Figure 5 Fitted Time-to-Readmission and Recovered Time-to-Develop the Condition for 2010 SID Patients [Color figure can be viewed at wileyonlinelibrary.com]



Note. Time-to-readmission $\rho \sim$ gamma(2.50, 4.80), time-to-develop the condition $\delta \sim$ gamma(1.81, 5.08).

tested seven delay-time distributions (see Table 1) with the same mean and different variance as a sensitivity analysis, since the delay-time distribution is estimated based on a small sample of 63 patients and no other publicly available dataset captured delay-time information. Table 1 shows how our policy outperforms current practice by significantly increasing the probability that ill patients are detected before readmission (defined as the detection probability). The relative improvement of the detection probability ranges from 23.9% to 65.3% (average = 49%) for the exponential and gamma delay-time distributions tested. This improvement is achieved solely by optimizing the timing and sequencing of the two checkups. As we shall see in the following sections, the detection probability further increases if we adopt more aggressive post-discharge monitoring policies by increasing the number of checkups. However, we would like to point out that the improvement is computed, using readmitted patients only, which represent 19% of the entire cohort. Hence, when taking both readmitted and non-readmitted patients into account, the improvement might be smaller. As a sanity check, we conducted simulations and verified that, under current practice, the simulated readmission rates predicted by our model were very close to the readmission rates that were actually observed in the data (both around 20%).

In Table 1, where the mean of the gamma distribution is held constant and the variance is increased, we see that increased (gamma-distributed) delay-time variance leads to greater spacing between checkups. The performance of the optimal policy also degrades as the (gamma-distributed) delay-time variance increases. This implies that efforts at standardizing patients' behavior at home could have benefits for readmission reduction because it reduces the delay-time variance. This variance effect is offset in the exponential case by the concurrent increase in mean delay-time, which indicates that efforts to keep patient conditions from degrading too fast (e.g., compliance with physician orders and adherence to medication), can also provide significant benefit by allowing the healthcare provider time to detect the condition before it becomes too severe. Note that our approach can be tailored to each patient's time to readmission characteristics, but because of data scarcity, it is difficult tailor the delay-time. If there were sufficient data, the delay-time could also be personalized using the same method used to personalize time to readmission predictions (Helm et al. 2016).

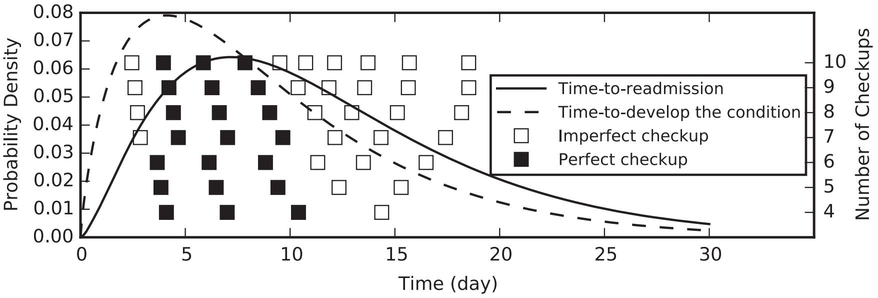
5.3. Optimal Timing and Sequencing of Checkups: Timing Outweighs Sequencing

Next, we explore the delay-time-spaced block structure shown by Proposition 1 and the optimal sequencing of checkups in a more generalized scheme involving 4–10 checkups in total with three office visits. Though conducting ten checkups within a 30-day period could be burdensome for both clinicians and patients, the purpose here is to study 10-checkup policies as the extreme upper bound for the sake of comparison and completeness, and further investigate the structure of checkup policies and their timing and sequencing.

From Figure 6, we draw the following insights: (i) checkups are scheduled in a contiguous block surrounding the mode of the time-to-develop the condition distribution with spacing approximately equal to the mean delay time. Slightly wider spacing is observed around the perfect checkups and the spacing increases as the probability of developing the condition decreases; and (ii) consecutive perfect checkups are placed surrounding the mode of the time-to-develop the condition curve (i.e., put the best checkups in the most hazardous period).

Although optimal policies favor consecutive perfect checkups around the mode, it is sometimes impractical to schedule them consecutively in a short period of time; particularly because many patients may live far from the hospital where their initial treatment occurred, making frequent travel to the

Figure 6 Optimal *n*-Checkup Sequencing and Timing, $n \in \{4, \ldots, 10\}$: Consecutive Perfect Checkups Appear Around the Mode of $g_\delta(\cdot)$



Notes. Assumptions: $D \sim$ exponential(2.35); r of perfect checkups = 1, r of imperfect checkups = 0.6; the left axis denotes the probability density; the right axis denotes the number of checkups. The detection probabilities are 0.40, 0.43, 0.46, 0.48, 0.50, 0.52, and 0.54, respectively (from bottom to top).

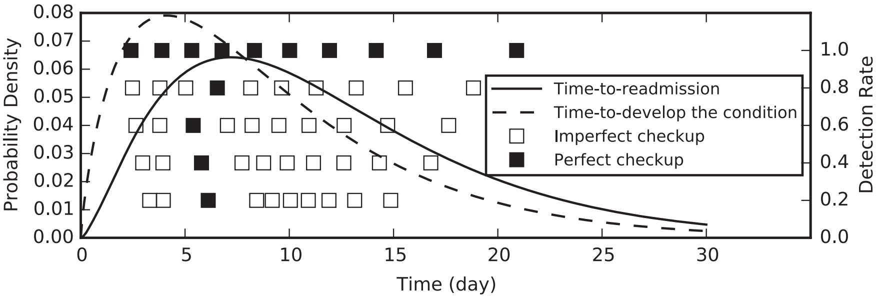
hospital difficult or impossible. Fortunately, we find that, as long as the timing is optimal, the policies are robust to sequencing; that is, the gaps between the worst-case and the best-case sequences for all policies in our test suite (1–10 checkups consisting of 0–3 office visits and phone calls) ranged between 0.2% and 0.5%, indicating that the timing of checkups is much more important than the sequencing. One way to explain why sequencing is less important is that the optimization will mimic a perfect checkup by scheduling multiple imperfect checkups closer together. For example, three phone calls of detection rate 0.6 (made at once) have an equivalent detection rate of $1 - 0.4^3 = 0.94$. We conjecture that, by striking a balance between the spacing of checkups and the effective detection rate, the sub-optimal sequencing can mimic the behavior of the optimal sequencing. The robustness to sequencing is a valuable property: as the number of checkups increases, the number of permutations of checkup sequences becomes large (e.g., the 10-checkup policy in Figure 6 has $\binom{10}{3} = 120$ sequences), requiring a significant amount of computational power to obtain an optimal solution. Results from the sequencing analysis, however, generate near-optimal policies by fixing the checkup sequence that is convenient for the physician and the patient and then optimizing the timing of checkups. This also allows for accommodating physician and patient preferences with little degradation in performance.

REMARK 8. (MULTI-MODAL TIME-TO-DEVELOP THE CONDITION DISTRIBUTIONS). In Appendix E, we test our model numerically using a multi-modal time-to-develop the condition distribution estimated using a Gaussian Kernel Density Estimator. We show that checkup policies can still be solved numerically to optimality and the differences in optimal detection probabilities are within 2%.

5.4. Impact of Detection Rate on Timing: Greater Coverage with Better Checkups

In this section, we study the impact of varying the detection probability of an imperfect checkup, r, and extend the insight drawn from Theorem 1 using a realistic potential monitoring schedule (according to

Figure 7 Optimal Checkup Timings under Different Detection Rates: Checkups are Placed in a Contiguous Block and Move Farther Apart as the Detection Rate Increases



Notes. Assumptions: $D \sim$ exponential(2.35), number of checkups = 10. The detection probabilities are 0.31, 0.41, 0.50, 0.57, and 0.64, respectively (from bottom to top).

our clinical collaborator) of one office visit and nine phone calls. While this scenario is more aggressive than current practice, it is still reasonable because phone calls can be done cost-effectively using nurses, trained technicians, or even automated call systems (see www.cloud9hcs.com and Tagliente et al. 2016). Results are presented in Figure 7.

In Figure 7, the spacing between checkups increases as the detection rate improves. This aligns with Theorem 1 and our intuition: more accurate checkups can be spread farther apart; whereas less accurate checkups should be placed closer together to account for the higher probability that the condition is missed by previous checkups. With more accurate checkups, the associated larger spacings will cover a longer time period. Since checkups are scheduled less frequently, patients and family members are less likely to be inconvenienced. For example, too much contact may lead patients to become irritated, ignore phone calls, or not consider questions as attentively. Another benefit is that by covering a longer time period, there is increased ability to detect potentially developing conditions. Finally, the extended monitoring period may help patients feel that they are receiving better attention/care, which can build trust between the patient and clinician, thereby improving patient satisfaction.

5.5. Marginal Benefits of Increasing Checkup Quantity vs. Improving Checkup Quality: Quantity Outweighs Quality

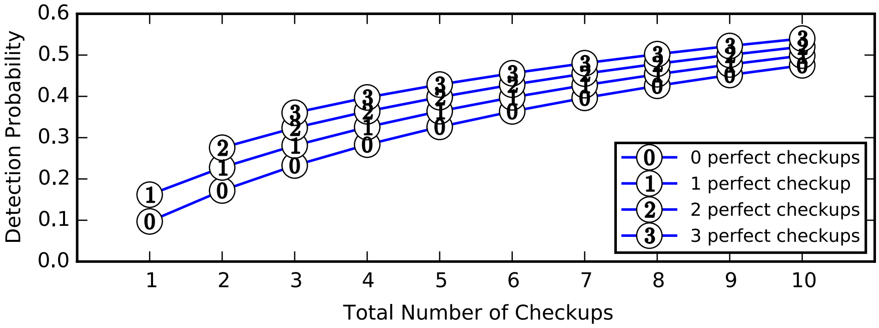
Since scheduling frequent follow-up office visits will increase the burden on frequently heavily loaded clinician schedules (Baron 2010), in this section, we consider the value of doing more phone calls as a substitute for office visits. Importantly for the clinical community, we find that checkup quantity is more important than quality; that is, multiple phone calls function as a good substitute for office visits.

In our first experiment, we study optimal checkup policies that have a total number of checkups between one and ten with zero to three office visits.

As shown in Figure 8, both increasing the number of checkups and increasing the number of perfect checkups improves the detection probability. However, we find that adding one additional phone call is nearly as effective as switching one phone call to an office visit. In our test suite where 1–10 checkups consisting of 0–3 office visits and phone calls were optimized, scheduling one additional phone call increases the detection probability by an average of 3.35% whereas replacing a phone call with an office visit (and rerunning the optimization) increases the detection probability by 3.45%. We also calculated the minimum number of additional phone calls needed to outperform replacing a phone call with an office visit. Across our test suite, on average, an office visit (r = 1) can be replaced with 2.57 phone calls (r = 0.6). Further, when the total number of checkups is less than five, an office visit can be replaced with two phone calls.

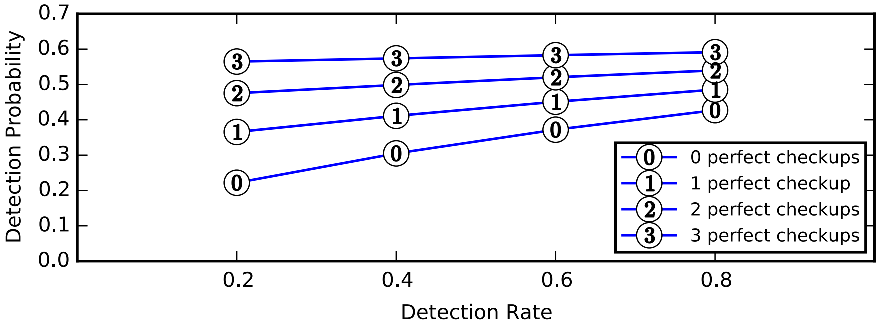
This result is highly valuable from the practical perspective, as phone calls are significantly less resource-intensive than office visits for both patients and physicians. Notice that phone calls have numerous benefits over office visits: (i) patients may be located far from the clinic and may have limited mobility and transportation options; (ii) making an office visit is burdensome as the capacity of the clinic and physicians' time are limited; and (iii) making phone calls can be done efficiently through specialized call centers or physicians' nursing or auxiliary staff in their spare time. The key finding is that an effective checkup policy can leverage these inexpensive phone calls to achieve similar results as those obtained with the more expensive and inconvenient office visits.

Figure 8 Detection Probability of *n*-Checkup Policies with 0–3 Perfect Checkups: Conducting More Imperfect Checkups is Better Than Replacing 1 Imperfect Checkup with 1 Perfect Checkup; a Perfect Checkup Can Be Replaced with 2.57 Imperfect Checkups with 0.6 Detection Rate [Color figure can be viewed at wileyonlinelibrary.com]



Note. Assumptions: $D \sim$ exponential(2.35), r of perfect checkups = 1, r of imperfect checkups = 0.6.

Figure 9 Detection Probability as a Function of Detection Rate: 20% Absolute Improvement in Detection Rate Achieves 29%–70% (average = 47%) of the Benefit Achieved by Upgrading an Imperfect Checkups to a Perfect Checkup [Color figure can be viewed at wileyonlinelibrary.com]



Note. Assumptions: $D \sim$ exponential(2.35) with 10 checkups.

5.6. The Benefit of Improving the Efficacy of Phone Calls

One strong interest in efforts at readmission reduction lies in designing effective questionnaires for phone and telemedicine checkups (see, e.g., readmission reduction startup company Cloud9, which has developed detailed questionnaires for many conditions, www.cloud9hcs.com) and testing predictive models based on historical responses to survey questions. Design of such questionnaires to effectively target the main causes of readmission (as an example for cystectomy, five main conditions account for almost all of the readmissions) can increase the detection probability of a phone call. These questionnaires are particularly easy to implement if the call is being conducted by someone who is not the physician or, or if it is conducted by an automated call system. To determine the importance of such improvements and subsequently the amount of effort that should be expended to perfect such surveys, we analyzed the impact of the detection probability, r, on the efficacy of a monitoring schedule.

Figure 9 shows that, as might be expected, the benefit of replacing a phone call with an office visit diminishes as the detection rate improves. To analyze the overall impact, we developed a test suite, where policies consisting of 10 checkups with 0–3 office visits were optimized. We incremented the detection rate from 0.2 to 0.8 (with a step size of 0.2), with 0.2 and 0.8 functioning as extreme lower/upper bounds for the sake of comparison and completeness. We started by computing the detection probability as a function of the detection rate of the phone calls. We then estimated (i) the improvement in detection probability achieved by upgrading an existing phone call to an office visit; and (ii) the improvement in detection probability achieved by increasing the detection rate of the phone calls. Finally, we computed the relative effectiveness of increasing the phone call detection rate by 20% (compared to upgrading an existing phone call to an office visit). A relative effectiveness of 100% means that increasing the phone call detection rate by 20% is as effective as upgrading a phone call to an office visit.

Across this test suite, on average, increasing the detection rate by 20% absolutely (e.g., 0.2 → 0.4) achieves 29%–70% (average = 47%) of the benefit achieved by replacing a phone call with an office visit (see detailed computation in Appendix J). The following table shows the relative effectives.

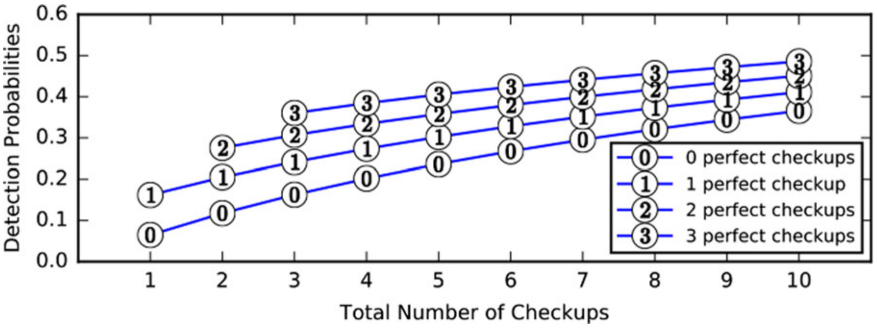
The relative marginal benefit of increasing the detection rate is greater when the detection rate is low and the number of office visits is few (see Table 2). Notice that the relationship is not linear (plausibly concave as shown in Figure 9). The intuition is that the effort required to improve checkup policies increases as the policies get more aggressive and effective.

The practical implication suggests that physicians can benefit significantly by designing more effective phone call questionnaires, which may be used to help replace excessive or burdensome office visits. Increasing the detection rate of phone calls may be achieved by providing patient education upon hospital discharge (e.g., informing patients of symptoms that indicate worsening conditions), ensuring that the content of post-discharge questionnaires are tailored as much possible to individual patients and their personal characteristics (which can be identified with readmission risk models at the time of discharge), and targeting high risk conditions (e.g., infection, dehydration, kidney failure, failure to thrive) with focused questions.

Table 2 Relative Effectiveness of Increasing Phone Call Detection Rate with Respect to Replacement of a Phone Call with an Office Visit

No. of office visits\Phone call detection rate	0.2 → 0.4	0.4 → 0.6	0.6 → 0.8
0 → 1	58%	63%	70%
1 → 2	41%	46%	50%
2 → 3	35%	29%	31%

Figure 10 Detection Probability of Checkup Policies with 0–3 Perfect Checkups Tested on 2009 SID Patients (Method 1) [Color figure can be viewed at wileyonlinelibrary.com]

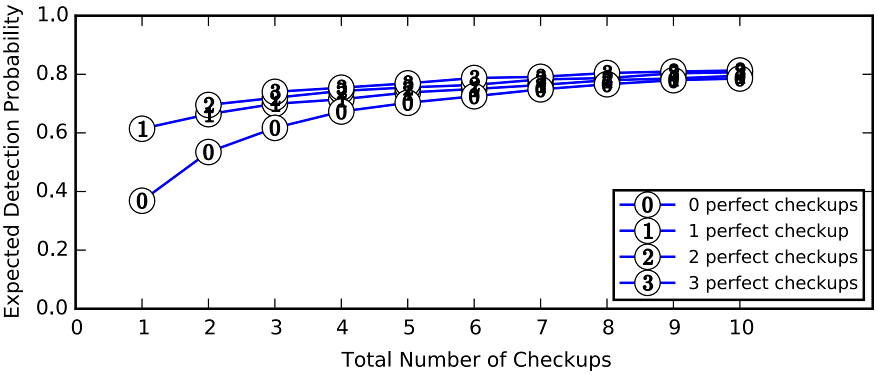


5.7. Out-of-Sample Testing on a Separate Dataset and Solution Robustness

To validate and test our models, we parameterized our delay-time random variable using the first dataset from our partner hospital for radical cystectomy patients. We then estimated the time-to-readmission by fitting a gamma distribution (best fit) to the 2010 SID dataset, also for radical cystectomy patients. Using our inverse Laplace transform method, we were able to recover the distribution on the time-to-develop the condition. Finally, we generated an optimal monitoring schedule based on the dynamics obtained from combining our partner hospital delay-time data with the 2010 SID readmission data. We then tested this policy on a new dataset, 2009 SID data, comparing our checkup times to the readmissions for cystectomy patients across five states in 2009. To do so, we consider two methods. In both methods, we begin by determining the optimal policy with parameters estimated from 2010 SID data. *Method 1:* We compare the performance of the optimal policy from 2010 data when applied to a time-to-readmission curve estimated from the 2009 data vs. the policy that optimizes according to the true 2009 time-to-readmission curve. We can then compare the optimality gap caused by errors in estimation of the time-to-readmission curve. *Method 2:* We apply the 2010 optimal policy to all the cystectomy patients from 2009 SID data and estimate the performance using each patient's actual readmission time and calculating the probability that his/her delay-time was long enough such that one of the inspections from our optimal policy would have caught the condition before it caused a readmission (see Appendix K for details).

Method 1 is shown in Figure 10. The detection probabilities range from 0.1 to 0.5 and are very close to the detection probabilities using a time-to-readmission curve estimated with the 2009 data itself (in-sample). The absolute optimality gaps were less than 5% (see Appendix L). We also calculated the relative optimality gaps and switched the testing and training sets to further validate the findings (see Appendix M. The largest relative optimality gaps were observed in one-checkup models, which are not advisable in practice. As the number of checkups increases, the relative optimality gap diminishes. This indicates that the model becomes more robust as the number of checkups increases, providing more support for the idea that quantity of checkups is highly important. Practically speaking, not only does larger quantity eliminate the need for excessive office visits, it also

Figure 11 Proportion of Conditions Captured by the Optimal Policy with 0–3 Perfect Checkups Obtained, Using 2010 SID Patients and Tested on 2009 SID Patients (Method 2) [Color figure can be viewed at wileyonlinelibrary.com]



increases the robustness of the solution to errors in estimation.

Method 2 evaluates how well the optimal policy would have performed in practice if implemented on the radical cystectomy patients from the 2009 SID data. Figure 11 presents the results of this study, indicating that the optimal policies estimated from 2010 SID data would in fact have been highly effective if put into practice on the patients of the out-of-sample dataset. In particular, the estimated detection probabilities (based on actual readmission times) are greater than 60% using one or more perfect checkups on the 2009 SID patients. It is worth highlighting the difference between Methods 1 and 2 (i.e., Figure 8 vs. Figure 11): in Figure 8, we were plotting the objective function, which is parameterized with gamma and exponential distribution curves fitted from the training dataset. However, in Figure 11, we were plotting a different objective, which uses the actual time to readmission observations combined with the delay-time distribution function plus the discrete observations (see Appendix K for more details).

This improved performance seems to stem from the fact that the true time-to-readmission for cystectomy patients tends to be more heavily front-loaded in the first 7–8 days than the fitted gamma distribution. Another fact that contributed to this higher performance is that the time to readmissions we used are in days (discrete) rather than time (continuous). Using discrete data created a lumping effect and lead to improved performance. We are unable to use the exact time of readmission (continuous data) to validate our model as it is protected information and could be used to identify patients. Moreover, since the optimal policies tend to also bunch a number of checkups soon after patient discharge, this policy ends up actually detecting more conditions in practice than would have been estimated by the fitted gamma distribution for time-to-readmission.

As a further benefit revealed by this study (seen in Figure 11), it appears that one office visit along with a few phone calls is sufficient to capture much of the value of post-discharge checkups. This is good news for busy clinicians concerned about the added burden of increased checkups.

5.8. Design of Practical Post-Discharge Checkup Policy

Combining the insights drawn from our analytical and numerical analyses, we provide the following rules of thumb to facilitate the design of post-discharge checkup policies.

- **Timing of checkups outweighs sequencing:** (i) schedule checkups in a block surrounding the most-likely time (mode) of developing a condition; (ii) keep the time between checkups close to the expected delay-time; (iii) office visits should be scheduled near the time of highest risk of readmission for the patient.
- **Cover a longer time period and reduce office visits with better checkups:** Improving the quality of phone call checkups (e.g., better questionnaires, patient education) allows the checkup team to (i) cover a longer time period with less frequent calls (better for patients and detects more potential conditions), (ii) reduce the number of office visits without reducing readmission detection (better for patients, clinicians, and healthcare organizations). Further, helping to standardize patient behavior at home, thereby reducing delay-time variance, has added detection benefits.
- **Quantity of checkups outweighs quality:** Multiple imperfect checkups serve as a good substitute for office visits; that is, making more phone calls can be nearly as effective as replacing a few phone calls with office visits. Further, the larger the quantity of checkups, the more robust the solution is to errors in estimation/optimization.

In practice, hospitals could use the following steps to design better post-discharge monitoring policies: (i) estimate the time-to-develop the condition and the delay-time; (ii) design an effective phone call questionnaire; (iii) schedule checkups in a block with spacings approximately equal to the mean delay-time; (iv) schedule office visits (perfect checkups) close to the time at which patients are at the highest risk of readmission; (v) spread phone calls farther apart from each other to cover a longer time period with improvements on the phone call questionnaires.

6. Discussion and Conclusions

In this study, we address the prevalent issue of hospital readmissions that concerns healthcare professionals, hospital patients, and policy makers. We propose an analytical model based on delay-time analysis to design more effective post-discharge checkup policies for individual patients. Key results from our model not only provide theoretical extensions of the traditional delay-time analysis framework, but also important insights for healthcare decision makers designing post-discharge checkup policies. By simultaneously optimizing with respect to multiple factors such as the number of checkups, the timing of checkups, and the types of checkup methods used, our model demonstrates significant improvements over current practice. Using the same number of checkups, current practice (which detects only 16%

of the conditions experienced by readmission-bound patients) can be improved up to 23%, a relative improvement of 43.7%.

Future extensions upon this research may involve examining the benefit of detecting illnesses as early as possible. The current model assumes equal benefit from all early illness detections, however, it may be valuable to assign more benefit to earlier detections as they may result in less burden on the patients. Similarly, the current model also assumes that checkups have constant detection rate over the duration of a patient's readmission-causing condition. It may be valuable to examine the effect of time-dependent detection rate of phone calls, for example, the detection rate becomes higher as the patient has had the condition for a longer time. Another extension is to jointly optimize discharge (inpatient) and post-discharge (outpatient) decisions as the timing of discharge can affect readmission risk (Kelly et al. 2015, Rosen et al. 2017). While parameterizing our model with real data, we realized that empirical estimation could be challenging as our model requires two distributions (time-to-develop the condition and delay-time) as the input. One of the key empirical challenges is the issue of censoring, as we only utilized data within the finite 30-day readmission penalty window. In addition, patients have different intrinsic readmission risk: while some patients would not be readmitted, other patients would be readmitted regardless of post-discharge monitoring and interventions. Though the two distributions (and data beyond 30-day follow-up) are not widely available currently, we believe that our analysis will motivate the documentation and utilization of the delay-time and time-to-develop the condition information. We leave to future work the personalized delay-time and time-to-develop the condition forecast as well as more robust empirical estimation that considers the censorship of data.

The application of our model and findings has the potential for broad impacts including reduced hospital readmissions, improved quality of patient care, improved patient satisfaction, and reduced healthcare costs, all without overburdening clinicians (as clinician burden is often a major barrier to implementation of new healthcare practices). This is achievable by aligning checkup policy design with a number of key insights, namely: timing of checkups is the most important factor, checkup timing should be adjusted according to checkup detection rates, and checkup quantity is more important than checkup quality. At the same time, our model presents unique extensions to the traditional delay-time analysis framework by allowing for a time-varying failure rate and inhomogeneous detection rate. Thus, our model extends the scope of delay-time modeling and provides new insights into the structure of these types of problems. This ultimately broadens the scope of problems in which delay-time analysis can be applied.

Tested on an out-of-sample dataset containing 332 patients from the states of Florida, Iowa, North Carolina, New York, and Washington, our results demonstrate robustness, with absolute optimality gaps within 5%. As the number of checkups increases, the robustness further increases as the optimality gaps diminish. Our clinical collaborators have shown great interest in implementing our models and look forward to putting them through clinical testing. Going beyond cystectomy patients, the new framework developed has the potential to significantly reduce readmissions from a variety of surgical procedures, thereby improving the quality of patient care and decreasing healthcare costs.

Acknowledgments

This research was supported in part by NSF CMMI-1552545. The computational resources and services were provided by Advanced Research Computing at the University of Michigan, Ann Arbor. We thank Naveen Krishnan and Benjamin Li for their efforts in the medical chart review project. We also thank our clinical collaborators: Alexander Helfand, Chang He, Brent Hollenbeck, and Bruce Jacobs for their support and valuable feedback toward completion of the research presented in this study. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. We thank Naveen Krishnan and Benjamin Li for their efforts in the medical chart review project. We also thank Alexander Helfand, Chang He, Brent Hollenbeck, and Bruce Jacobs for their support and valuable feedback toward completion of the research presented in this study.

References

Ayer, T., O. Alagoz, N. K. Stout. 2012. OR forum-a POMDP approach to personalize mammography screening decisions. *Oper. Res.* **60**(5): 1019–1034.

Ayer, T., O. Alagoz, N. K. Stout, E. S. Burnside. 2015. Heterogeneity in womens adherence and its role in optimal breast cancer screening policies. *Management Sci.* **62**(5): 1339–1362.

Barlow, R. E., F. Proschan. 1996. *Mathematical Theory of Reliability*, vol. 17. SIAM, Philadelphia, PA.

Baron, R. J. 2010. What's keeping us so busy in primary care? A snapshot from one practice. *New Engl. J. Med.* **362**(17): 1632–1636.

Bartel, A. P., C. W. Chan, S.-H. Kim. 2016. Should hospitals keep their patients longer? The role of inpatient care in reducing post-discharge mortality. Technical report, National Bureau of Economic Research.

Bavafa, H., S. Savin, C. Terwiesch. 2013. Managing office revisit intervals and patient panel sizes in primary care. *Available at SSRN 2363685*.

Bayati, M., M. Braverman, M. Gillam, K. M. Mack, G. Ruiz, M. S. Smith, E. Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE* **9**(10): e109264.

Bellone, J. M., J. C. Barner, D. A. Lopez. 2012. Postdischarge interventions by pharmacists and impact on hospital readmission rates. *J. Am. Pharm. Assoc.* **52**(3): 358.

Benbassat, J., M. Taragin. 2000. Hospital readmissions as a measure of quality of health care: Advantages and limitations. *Arch. Intern. Med.* **160**(8): 1074.

Brailsford, S. C., P. R. Harper, J. Sykes. 2012. Incorporating human behaviour in simulation models of screening for breast cancer. *Eur. J. Oper. Res.* **219**(3): 491–507.

Brandeau, M. L., F. Sainfort, W. P. Pierskalla. 2004. *Operations Research and Health Care: A Handbook of Methods and Applications*, vol. 70. Springer, Boston, MA.

Chan, C. W., V. F. Farias, N. Bambos, G. J. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* **60**(6): 1323–1341.

Christer, A. H. 1999. Developments in delay time analysis for modelling plant maintenance. *J. Oper. Res. Soc.* **50**(11): 1120–1137.

Christer, A. H., N. Jack. 1991. An integral-equation approach for replacement modelling over finite time horizons. *IMA J. Manage. Math.* **3**(1): 31–44.

Costantino, M. E., B. Frey, B. Hall, P. Painter. 2013. The influence of a postdischarge intervention on reducing hospital readmissions in a medicare population. *Popul. Health Manag.* **16**(5): 310–316.

D'Amore, J., J. Murray, H. Powers, C. Johnson. 2011. Does telephone follow-up predict patient satisfaction and readmission? *Popul. Health Manag.* **14**(5): 249–255.

Dagpunar, J. S. 1994. Some necessary and sufficient conditions for age replacement with non-zero downtimes. *J. Oper. Res. Soc.* **45**(2): 225–229.

Dartmouth Atlas Project. 2013. The revolving door: A report on U.S. hospital readmissions. Technical report.

Deo, S., J. Gallien, J. O. Jónasson. 2014. Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: Models and application to mozambique. *Available at SSRN 2511549*.

Dudas, V., T. Bookwalter, K. M. Kerr, S. Z. Pantilat. 2001. The impact of follow-up telephone calls to patients after hospitalization. *Am. J. Med.* **111**(9): 26–30.

Erenay, F. S., O. Alagoz, A. Said. 2014. Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manuf. Serv. Oper. Manag.* **16**(3): 381–400.

Fu, B., W. Wang, X. Shi. 2012. A risk analysis based on a two-stage delayed diagnosis regression model with application to chronic disease progression. *Eur. J. Oper. Res.* **218**(3): 847–855.

Green, L. V., S. Savin, Y. Lu. 2013. Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication. *Health Aff.* **32**(1): 11–19.

Gretton, A., K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola. 2008. A kernel statistical test of independence. *Adv. Neural Inf. Process. Syst.* **20**: 585–592.

Güneş, E. D., E. L. Örmeci, D. Kunduzcu. 2015. Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. *Prod. Oper. Manag.* **24**(1): 1–20.

Harper, P. R., S. K. Jones. 2005. Mathematical models for the early detection and treatment of colorectal cancer. *Health Care Manage. Sci.* **8**(2): 101–109.

Helm, J. E., M. S. Lavieri, M. P. Van OyenJ. D. Stein, D. C. Musch. 2015. Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Oper. Res.* **63**(5): 979–999.

Helm, J. E., A. Alaeddini, J. M. Stauffer, K. M. Bretthauer, T. A. Skolarus. 2016. Reducing hospital readmissions by integrating empirical prediction with resource optimization. *Prod. Oper. Manag.* **25**(2): 233–257.

Holland, R., J. Battersby, I. Harvey, E. Lenaghan, J. Smith, L. Hay. 2005. Systematic review of multidisciplinary interventions in heart failure. *Heart* **91**(7): 899–906.

Hu, M., B. Jacobs, J. Montgomery, C. He, J. Ye, Y. Zhang, J. Brathwaite, T. Morgan, K. Hafez, A. Weizer, S. Gilbert, C. Lee, M. Lavieri, J. Helm, B. Hollenbeck, T. Skolarus. 2014. Sharpening the focus on causes and timing of readmission after radical cystectomy for bladder cancer. *Cancer* **120**(9): 1409–1416.

Jacobs, B. L., Y. Zhang, J. Tan, H. Z. Ye, T. A. Skolarus, B. K. Hollenbeck. 2013. Hospitalization trends after prostate and bladder surgery: Implications of potential payment reforms. *J. Urol.* **189**(1): 59–65.

Jacobs, B. L., C. He, B. Y. Li, A. Helfand, N. Krishnan, T. Borza, A. A. Ghaferi, B. K. Hollenbeck, J. E. Helm, M. S. Lavieri, T. A. Skolarus. 2017. Variation in readmission expenditures after high-risk surgery. *J. Surg. Res.* **213**: 60–68.

James, J. 2013. Medicare hospital readmissions reduction program. *Health Affairs Health Policy Brief* **34**(2): 1–5.

Jardine, A. K. S., A. H. C. Tsang. 2005. *Maintenance, Replacement, and Reliability: Theory and Applications*. Dekker Mechanical Engineering, Taylor & Francis, Boca Raton, FL.

Joynt, K. E., A. K. Jha. 2012. Thirty-day readmissions-truth and consequences. *N. Engl. J. Med.* **366**(15): 1366–1369.

Kelly, K. N., J. C. Iannuzzi, C. T. Aquina, C. P. Probst, K. Noyes, J. R. T. Monson, F. J. Fleming. 2015. Timing of discharge: A key to understanding the reason for readmission after colorectal surgery. *J. Gastrointest. Surg.* **19**(3): 418–428.

Kent, D. L., R. Shachter, H. C. Sox, N. S. Hui, L. D. Shortliffe, S. Moynihan, F. M. Torti. 1989. Efficient scheduling of cystoscopies in monitoring for recurrent bladder cancer. *Med. Decis. Making* **9**(1): 26–37.

Kim, S.-H., C. W. Chan, M. Olivares, G. Escobar. 2014. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Sci.* **61**(1): 19–38.

Kocher, K. E., B. K. Nallamothu, J. D. Birkmeyer, J. B. Dimick. 2013. Emergency department visits after surgery are common for medicare patients, suggesting opportunities to improve care. *Health Aff.* **32**(9): 1600–1607.

Koh, H. K., K. G. Sebelius. 2010. Promoting prevention through the affordable care act. *New Engl. J. Med.* **363**(14): 1296–1299.

Leeds, I. L., V. Sadiraj, J. C. Cox, S. X. Gao, T. M. Pawlik, K. E. Schnier, J. F. Sweeney. 2015. Discharge decision-making after complex surgery: Surgeon behavior compared to predictive modeling to reduce surgical readmissions. *J. Am. Coll. Surg.* **221**(4): S123–S124.

Maillart, L. M., J. S. Ivy, S. Ransom, K. Diehl. 2008. Assessing dynamic breast cancer screening policies. *Oper. Res.* **56**(6): 1411–1427.

Milioni, A. Z., S. R. Pliska. 1988. Optimal inspection under semi-markovian deterioration: Basic results. *Nav. Res. Log.* **35**(5): 373–392.

Myers, E. R., D. C. McCrory, K. Nanda, L. Bastian, D. B. Matchar. 2000. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. *Am. J. Epidemiol.* **151**(12): 1158–1171.

Pinsky, P. F. 2004. An early-and late-stage convolution model for disease natural history. *Biometrics* **60**(1): 191–198.

PwC Health Research Institute. 2010. The price of excess: Identifying waste in healthcare spending 6.

Rosen, J. E., M. C. Salazar, K. Dharmarajan, A. W. Kim, F. C. Detterbeck, D. J. Boffa. 2017. Length of stay from the hospital perspective: Practice of early discharge is not associated with increased readmission risk after lung cancer surgery. *Ann. Surg.* **266**(2): 383–388.

Sanders, G. D., A. M. Bayoumi, V. Sundaram, S. P. Bilir, C. P. Neukermans, C. E. Rydzak, L. R. Douglass, L. C. Lazzeroni, M. Holodniy, D. K. Owens. 2005. Cost-effectiveness of screening for hiv in the era of highly active antiretroviral therapy. *New Engl. J. Med.* **352**(6): 570–585.

Sim, S. H., J. Endrenyi. 1993. A failure-repair model with minimal and major maintenance. *IEEE Trans. Rel.* **42**(1): 134–140.

Skolarus, T. A., B. L. Jacobs, F. R. Schroeck, C. He, A. M. Helfand, J. E. Helm, M. Hu, M. S. Lavieri, B. K. Hollenbeck. 2015. Understanding hospital readmission intensity after radical cystectomy. *J. Urol.* **193**(5): 1500–1506.

Tagliente, I., L. Trieste, T. Solvoll, F. Murgia, S. Bella. 2016. Telemonitoring in cystic fibrosis: A 4-year assessment and simulation for the next 6 years. *Interact. J. Med. Res.* **5**(2): e11.

Teng, Y., L. Han, W. Tu, N. Kong. 2011. Optimizing coverage for a chlamydia trachomatis screening program. Proceedings of the 2011 IEEE International Conference on Automation Science and Engineering, August 24–27, Trieste, Italy, pp. 531–536.

Tsodikov, A., A. Szabo, J. Wegelin. 2006. A population model of prostate cancer incidence. *Stat. Med.* **25**(16): 2846–2866.

Wang, H. 2002. A survey of maintenance policies of deteriorating systems. *Eur. J. Oper. Res.* **139**(3): 469–489.

Wang, W. 2012. An overview of the recent advances in delay-time-based maintenance modelling. *Reliab. Eng. Syst. Safe.* **106**: 165–178.

Weinberger, M., E. Z. Oddone, W. G. Henderson. 1996. Does increased access to primary care reduce hospital readmissions? *New Engl. J. Med.* **334**(22): 1441–1447.

White, C. C. 1977. A markov quality control process subject to partial observation. *Management Sci.* **23**(8): 843–852.

Wong, F. K. Y., S. K. Y. Chow, T. M. F. Chan, S. K. F. Tam. 2013. Comparison of effects between home visits with telephone calls and telephone calls only for transitional discharge support: A randomised controlled trial. *Age Ageing* **43**(1): 91–97.

Yeh, R. H. 1997. Optimal inspection and replacement policies for multi-state deteriorating systems. *Eur. J. Oper. Res.* **96**(2): 248–259.

Zhang, J., B. T. Denton, H. Balasubramanian, N. D. Shah, B. A. Inman. 2012a. Optimization of prostate biopsy referral decisions. *Manuf. Serv. Oper. Manag.* **14**(4): 529–547.

Zhang, S., P. Hanagal, P. Frazier, A. J. Meltzer, D. B. Schneider. 2012b. Optimal patient-specific post-operative surveillance for vascular surgery. Working paper, Cornell University.

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article:

Appendix A: Personalization of Readmission Risk.
Appendix B: Model Notation and Parameters.
Appendix C: Solution Approach.
Appendix D: Example of Non-Concave Objective Function.
Appendix E: Multi-modal Distribution.
Appendix F: Proof of Lemmas and Remarks.
Appendix G: Different Patient Types.
Appendix H: Unimodality Assumption in Six Other Major Surgeries.
Appendix I: Inverse Laplace Transform.
Appendix J: Checkup Quantity vs. Quality.
Appendix K: Description of Validation Method 2.
Appendix L: Validation Results Using Method 1.
Appendix M: Switching Training and Testing Set.