6-2021

Essays on predictive and prescriptive analytics for risk monitoring and population screening in healthcare management

Yan HE

Singapore Management University

Citation

HE, Yan. Essays on predictive and prescriptive analytics for risk monitoring and population screening in healthcare management. (2021). 1-197.

Available at: https://ink.library.smu.edu.sg/etd_coll/359

ESSAYS ON PREDICTIVE AND PRESCRIPTIVE ANALYTICS FOR RISK MONITORING AND POPULATION SCREENING IN HEALTHCARE MANAGEMENT

YAN HE

SINGAPORE MANAGEMENT UNIVERSITY

2021

Essays on Predictive and Prescriptive Analytics for Risk Monitoring and Population Screening in Healthcare Management

by

Yan He

Submitted to Lee Kong Chian School of Business
in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Business

Dissertation Committee:

Zhichao Zheng (Daniel) (Chair)
Associate Professor of Operations Management
Lee Kong Chian School of Business, Singapore Management University

Yini Gao (Sarah)
Assistant Professor of Operations Management
Lee Kong Chian School of Business, Singapore Management University

Guiyun Feng
Assistant Professor of Operations Management
Lee Kong Chian School of Business, Singapore Management University

Hai Wang
Assistant Professor of Information Systems
School of Computing and Information Systems, Singapore Management University

SINGAPORE MANAGEMENT UNIVERSITY

I hereby declare that this dissertation is my original work
and it has been written by me in its entirety.
I have duly acknowledged all the sources of information
which have been used in this dissertation.

This dissertation has also not been submitted for any degree
in any university previously.

He Yan

Yan He
19 May 2021

Abstract

Due to increased aging populations and changes in lifestyles, we have witnessed an increased prevalence of various chronic and acute diseases and a drastic rise in healthcare expenditures in recent years. It is of paramount importance for public health to promote regular screening and close monitoring to detect the early onset of diseases. On the other hand, the increasing availability of healthcare data and advancement in data analytics offer a huge potential to facilitate this goal. We can analyze the vast amount of data and recommend more personalized diagnostic tests after receiving results and signals from screening tests and monitoring systems, which are critical decisions for the effective and efficient implementation of such screening programs and monitoring systems. Meanwhile, it is also necessary to consider human behavioral issues and their impact in making the recommendations. In particular, individual adherence to the recommended diagnostic tests can significantly affect the effectiveness and efficiency of the programs. This dissertation aims to integrate predictive analytics, optimization techniques, and behavioral models to improve risk monitoring and decision-making in patient monitoring systems and population screening programs.

This dissertation first studies the real-time risk monitoring problem for patients in intensive care units (ICUs). We identify a critical lag in the provision of information due to the long lead time to measure some laboratory test variables (e.g., creatinine, platelets, and bilirubin) used in calculating the Sequential Organ Failure Assessment (SOFA) score, a well-established and important risk measure for patients in ICUs. We develop machine learning models to estimate such variables using easily mea-

sured bedside variables, the rate of changes in bedside variables, and time lag from the previous laboratory test, which mimics how physicians assess patient conditions in practice. Then the predicted laboratory test variables can be used to calculate an estimate of the real-time SOFA score. We further take advantage of the estimated standard deviations from these models to construct intervals of the real-time SOFA scores. We hypothesize that the estimated score intervals could capture the uncertainty in patient condition since the previous test and provide valuable information in a new dimension that complements the nominal SOFA scores. Using a dataset collected from an ICU in a tertiary hospital in Singapore, we calibrate our model and validate the hypothesis by comparing the prognostic accuracy of the proposed approach on patients' 24-hour mortality and 30-day readmission with those from the SOFA score calculated using the conventional approaches. The proposed methodology could be applied to other risk measures to improve their prognostic accuracy and provide more reliable early warning for timely intervention.

The methodologies developed in the previous chapter can help raise a warning of potential deterioration in a patient's health condition, but the exact problem still has to be confirmed through follow-up diagnostic tests, which are typically more invasive and expensive. Medical resource overuse has become increasingly common in recent years and caused diverse problems, including unnecessary and risky diagnostic tests and overly intensive or expensive treatments. There is a growing call for more evidence-based decisions to reduce unnecessary diagnostic tests. The next part of the thesis dives into this problem to optimize the prescription of diagnostic tests during the health monitoring process, leveraging the improved risk monitoring tools developed in the previous chapter. In particular, we develop a finite-horizon, partially observable Markov decision process model to optimize the time to initiate a diagnostic test. Our model captures both measured and estimated clinical variables (including estimated intervals) in real-time to update the belief on a patient's underlying health condition. We apply the model to monitor patients' blood glucose levels to detect hyperglycemia, a common complication of critical illness. We calibrate the model using the same ICU dataset as in the previous chapter and demonstrate that the new approach can advance the detection time with fewer diagnostic tests. The methodology can also be applied to many other health monitoring systems, especially those powered by smart wearable health devices for chronic diseases. However, to optimally design the warning signals and recommend the diagnostic tests for such a monitoring system, one must consider the impact of human behavioral issues, especially individuals' perception of the warning signals and adherence to the recommendations. We address this challenge in the next chapter in the optimal design of population screening programs for cancer surveillance and screening.

Cancer remains one of the leading causes of human death, while early detection enables timely intervention and reduction in mortality rate. Two-stage screening programs are broadly implemented in practice among large average-risk populations to effectively and efficiently detect cancer in the early stages. Individuals receiving positive results in first-stage (initial) tests are recommended to undergo second-stage tests for further diagnosis. Notably, individuals' adherence to the second-stage tests, which is closely associated with the initial test design (sensitivity and specificity) and personal characteristics, varies considerably across individuals and leads to different cancer detection rates and demands for second-stage tests. We adopt a Bayesian persuasion framework to model the optimal initial test design problem in the context of colorectal cancer screening. Our goal is to balance the trade-off between test effectiveness (i.e., detection rates of cancer incidences) and test efficiency (i.e., demands for second-stage tests), considering individuals' adherence behavior. We conduct a nationwide survey in Singapore to calibrate the individual's response to changes in the test design. With the embedded behavioral model, we next optimize the threshold selection in the initial test design (which decides the test sensitivity and specificity).

We characterized the structural properties of an optimal initial test design. Using various data and information collected locally in Singapore and from the literature, we demonstrate that a well-designed initial test can detect more cancer incidences with fewer second-stage tests than the current practice. We further explore the benefits of using heterogeneous initial tests for different sub-populations and use the interpretable clustering technique to search for implementable rules to partition the population. We find that customized tests with simply an age-gender partition rule could bring significant extra benefits.

To conclude, this thesis studies the optimal design of real-time patient monitoring systems and population screening programs, using a combination of techniques from machine learning, optimization, game theory and survey design. By analyzing the comprehensive datasets collected from various sources, we showcase that well-designed monitoring systems and screening programs can benefit individuals, healthcare service providers, and health systems through improved effectiveness and efficiency in healthcare service delivery.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Prof. Zhichao Zheng, for giving me the opportunity to participate in wonderful projects in healthcare management and providing me constant support and invaluable guidance throughout the research. Prof. Zhichao Zheng is a creative, wisdom, meticulous and gentle scholar with academic aspirations and pursuits. His precious qualities have influenced me and shaped me into a more rigorous, responsible and genuine person. I can't imagine how I could survive the Ph.D. program without his mentorship.

I would also like to extend my special appreciation to my co-author, Prof. Yini Gao. Both she and Prof. Zhichao Zheng spent considerable time leading me to read papers, improving my mathematical skills, training my critical thinking, and guiding me in writing papers. As I look back on the past five years, these are the most memorable and cherished experiences. It has been my great privilege to have such a brilliant, intelligent, obliging and easy-going collaborator in my life.

I am also deeply indebted to my committee members, Prof. Hai Wang and Prof. Guiyun Feng, for their valuable suggestions on improving my dissertation and for raising questions that inspired me to broaden my research from various aspects. I would also like to extend my gratitude to the rest of the faculty members in my department for their encouragement and instruction during my study.

My special thanks also go to the staff in our Ph.D. office, especially Wah Eng Loh, for her prompt assistance and unlimited patience in reminding us to complete various tasks during the past five years.

I am particularly grateful to my peers, Qian Luo, Peng Wang and Nicholas Yeo. Qian Luo is one of the most sincere people that I have ever met. He is always passionate about life and is always willing to help whenever others have any difficulties. Peng Wang is a good listener. Whenever I have any negative emotions, he always listens to me and encourages me to cheer up. Nicholas Yeo is very kind and friendly. As a local, he shows us around and takes us to eat tasty food in our leisure time. I am so blessed to have such incredible friends in my life.

I would like to thank my seniors, juniors, and friends I have met during my Ph.D. study, Mengyu Ji, Jiahui Xu, Xueying Bian, Guang Cheng, Yuchong Zhang, Zilin Chen and Wenjia Zhang, for all the fun we had together and the support they provided. I am also very grateful to all the friends who have ever come into my life, Ying Wang, Mengjie Wang, Shen-Jingru Fang, Xiuqi Wang, Shanglu Li, Lu Chen, Bingyan Wang, Siyu Fang, Jia Xu, Ru Wei, Yizhu Wu, Wanyan Wang, Weishan Lyu, Chengzi Huang, Chenxi Cai, Ting Jiang, Dan Chen, Rui Sun, Yahui Yang, Yan Dai, Zichao Yang, Qinyi Ma, etc. I sincerely appreciate their presence and accompanying.

Last but not least, I would like to express my deepest gratitude to my family, especially my parents, Zhongjian He and Lanping Ouyang, and my brother Tianjun He, for their unconditional love and tremendous support. Special thanks also go to my beloved, Haomin Xu, for his company in all the happy and dark moments.

Thank you all for brightening up my world!

Yan He

Chapter 1

Introduction

Over the last decades, the increase in aging populations and the adoption of unhealthy lifestyles (e.g., physical inactivity, unhealthy diets and excessive alcohol consumption) have contributed to a rapid rise in the prevalence of numerous diseases. For example, the number of adults with diabetes has risen from 108 million in 1980 to 422 million in 2014; and the instances of cancer have reached 18.1 million in 2018 and are projected to rise to 29.5 million in 2040 (Zhou et al. 2016, Are et al. 2020). The increasing prevalence of diseases has resulted in a drastic surge in healthcare expenditures. According to the World Health Organization, the worldwide healthcare expenditure has continually risen between 2000 and 2018 and reached US $8.3 trillion in 2018, accounting for 10% of global GDP (Vrijburg and Hernández-Peña 2020). This unsustainable phenomenon has raised global attention to promote regular health screening and close monitoring to detect the early onset of diseases effectively and efficiently. In recent years, the expanding volume of healthcare data and advances in data analytics provides tremendous potential to facilitate this goal. We can explore the vast amount of information hidden in the data to provide individuals with more personalized diagnostic tests after receiving risk alerts and signals from screening programs and monitoring systems. Meanwhile, it is also crucial to consider human behavioral factors, as individual compliance with recommended diagnostic tests can greatly affect the effectiveness and efficiency of the programs. This dissertation aims to improve risk monitoring and decision-making in patient risk monitoring systems and population screening programs using a combination of techniques from predictive analytics, optimization, and behavioral models.

Chapter 2 looks into the decade-long challenges in the real-time risk monitoring and limitations of the existing early warning systems. It has long been recognized as challenging and essential to providing early identification of evolving illness and timely life-saving interventions prior to the occurrence of unexpected adverse events for patients, especially for critically ill patients in the intensive care unit (ICU) (Shickel et al., 2019). With the widespread use of electronic health record systems, patients' dynamic physiological data can be easily tracked and recorded. Moreover, the rapid development of machine learning and predictive modeling techniques has provided researchers with technical supports to delve into this problem using the recorded data (Solares et al., 2020). A number of early warning systems have already been developed to track patients' real-time health conditions in ICUs, such as Sequential Organ Failure Assessment (SOFA) score, Multiple Organ Dysfunction Score (MODS) and Logistic Organ Dysfunction Score (LODS) (Rapsang and Shyam, 2014). However, most existing scoring systems are calculated using long lead-time physiological variables; for example, the three laboratory test variables used to calculate the SOFA score (e.g., creatinine, platelets, and bilirubin) are not updated frequently, making the SOFA score unable to describe patients' real-time conditions and trigger early warnings sufficiently ahead of time for physicians to initiate effective and timely intervention. In addition, the existing scoring systems only give a score that typically fails to capture the uncertainty and ambiguity in the risk assessment. To address these issues, we develop a new framework to enhance any existing real-time early warning scoring systems based on the physicians' practice. The key component of the new framework is to predict the uncertainty in those long lead-time variables using variables and information easily obtained at the bedside. The framework then derives the real-time risk score and calculates an interval for the risk score that captures the uncertainty in patient condition, which provides another dimension of information for real-time patient risk monitoring. We validate the association of the estimated real-time scores and uncertainty intervals on 24-hour mortality and 30-day readmission, and we demonstrate that the new framework is able to improve the prognostic accuracy of the nominal scores significantly. Moreover, we develop a refined patient risk classification based on real-time estimated risk levels and the new dimension of uncertainty in risk assessment and further propose general guidelines for patient risk monitoring under the new classification.

Chapter 3 further studies how to leverage the methodologies developed in Chapter 2 to prescribe more personalized and evidence-based diagnostic tests. Diagnostic test results account for 60% to 70% of all critical decisions, including medications, further testing and discharges (Forsman, 1996). According to a recent report, the global clinical diagnostic service market size reached $200.3 billion in 2020, representing a significant component of healthcare spending (Grand View Research, 2021). However, about 20% to 30% of diagnostic tests have been identified as inappropriate (Zhi et al., 2013a). In addition, diagnostic tests may lead to a number of negative consequences, such as unnecessary patient discomfort, excessive utilization of phlebotomy and additional blood transfusions (Eaton et al., 2017). It has drawn increasing interest amid a growing emphasis on improving the effectiveness and efficiency of diagnostic tests. Despite enormous attention and efforts that have been invested in the medical field, there are no studies in operations management that address the overutilization problem of diagnostic tests using real-time predictive analytics. In this chapter, we propose a finite-horizon, partially observable Markov decision process (POMDP) model to optimize the prescription of diagnostic tests in the detection of acute diseases. Specifically, we build a real-time predictive model that incorporates frequently updated clinical information to estimate individuals' disease progressions, and we further employ the uncertainty interval to capture the ambiguity of the risk assessment. We then embed the predictive model in the POMDP framework and use the predictions and uncertainty intervals as observations for health belief updates and decision-making support. We evaluate the performance of our model through simulations and case studies in the ICU setting. Our results show that the proposed framework is able to advance the detection time with fewer diagnostic tests. Our analysis provides a new framework for hospitals, smart healthcare service providers and governments to design optimal diagnostic testing policies, which plays a vital role in managing medical resources and improving patients' health outcomes.

Chapter 4 further considers human behavioral factors. Heterogeneity in individuals' adherence to medications, medical interventions, and disease screening programs has long been noted in healthcare, and numerous works have identified that individual behavioral factors contribute substantially to this phenomenon (e.g., Osterberg and Blaschke, 2005, Morgan et al., 2015, Robiner, 2005). For example, Golman et al. (2017) discuss several behavior factors that lead to information avoidance behavior, including “optimism maintenance” (i.e., people are optimistic about their health states and tend to dismiss the unwarranted information) and “risk, loss, and disappointment aversion” (i.e., aversion of possible perceived disappointment or loss). In this regard, it is crucial to incorporate individuals' endogenous behavioral responses when designing screening tests. In this chapter, we study the optimal initial test design problem in the context of CRC screening. To detect cancer in the early stages, two-stage screening programs are widely implemented in practice, where individuals receiving positive outcomes in first-stage (initial) tests are recommended to undergoing second-stage tests for further diagnosis. The design of the initial test, i.e., selecting cut-off points for generating test outcomes, is crucial for screening effectiveness and efficiency. In addition, it is observed that not all individuals receiving positive outcomes would follow up with the second-stage test; and the adherence behavior is closely associated with the initial test cut-off point selection as it may influence individuals' trust on the test (Plumb et al., 2017, Lee and Lee, 2018). We aim to balance the trade-off between test effectiveness (i.e., cancer detections) and test efficiency (i.e., economic costs), considering individuals' guideline adherence behavior. We adopt a Bayesian persuasion framework with information avoidance to characterize the initial test design and the response from individuals to the screening guideline and then leverage a nationwide survey conducted in Singapore to calibrate the individual's behavior. We show that under certain conditions, an initial test with a binary outcome (i.e., a dichotomous test) is optimal for screening effective maximization, and a continuous test is optimal for screening compliance maximization. We further explore the benefits of using heterogeneous initial tests to different sub-populations and apply the interpretable clustering technique to search for implementable rules in partitioning the population. We demonstrate that a well-designed initial test is able to detect more cancer incidences with fewer second-stage tests compared to current practice, and customized tests with an age-gender partition rule would bring substantial benefits.

Chapter 2

Real-time Estimated SOFA Score with Intervals: Improved Risk Monitoring with Estimated Uncertainty in Health Condition for Patients in ICU

2.1 Introduction

Sepsis, an inflammatory response to infection, is the leading cause of hospital deaths (Hall et al. 2011, Liu et al. 2014, Paoli et al. 2018) and is a major component of worldwide healthcare expenditures (Adhikari et al. 2010, Vincent et al. 2014, Rudd et al. 2018). Sepsis accounted for more than \$24 billion (13%) of total US hospital expenses in 2013, and the number of sepsis cases is still on the rise (Paoli et al. 2018). SOFA score was designed to quantify organ dysfunction in sepsis (Vincent et al. 1996, 1998), and it was later validated as a predictive marker of patients' mortality in the ICU (Ferreira et al. 2001, Holder et al. 2017). SOFA score measures the level of dysfunction for six organ systems: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems, and it contains 13 variables, including vital signs, laboratory results, and medications. According to the new definition of Sepsis 3 (Singer et al. 2016), organ dysfunction can be identified as an acute change in the SOFA score of 2 or more points, which is associated with an in-hospital mortality greater than 10%. The definition emphasizes the paramount importance and substantial benefits of timely updating the SOFA score.

However, since three components of the SOFA score—creatinine, platelets, and bilirubin—require laboratory tests that are less frequently conducted than vital signs (Singer et al. 2016, Kumwilaisak et al. 2008), it would be difficult to capture patients' organ dysfunction in time. Although increasing the frequency of laboratory tests may allow relatively early detection of deterioration, it could impose heavy financial burdens on both patients and the health systems. It might also lead to over-utilization of phlebotomy, decreased hemoglobin value, and, consequently, hospital-acquired anemia (Raith et al. 2017, Finkelsztein et al. 2017). On the contrary, numerous interventions have been implemented across multiple institutions to reduce laboratory tests (Keehan et al. 2015, Tsujita et al. 2010). For example, an initiative of the ABIM Foundation, Choosing Wisely, aims to decrease unnecessary medical tests, treatments, and procedures (Salisbury et al. 2011). Furthermore, the turnaround time of a laboratory test—the time between submitting the laboratory specimen and receiving the results—could range from several hours to days (May et al. 2006, Eaton et al. 2017), further increasing the difficulty of updating patients' SOFA score in time. As a result, SOFA scores calculated using lagged information are subject to the uncertainties in the changes of patient conditions since the previous test. This partially contributed to the criticism on the effectiveness of the SOFA score as a risk monitoring tool (Marik and Taeb 2017, Freund et al. 2017).

To address this issue, a new measure named quick SOFA (qSOFA)—which only uses bedside measurable variables including respiratory rate, mental status and systolic blood pressure—was proposed to act as a proxy of SOFA for early detection of suspected sepsis (Singer et al. 2016). However, it was noted that qSOFA was less robust than SOFA due to its simplicity (Singer et al. 2016), and there were also controversial views on the effectiveness of the qSOFA score (Marik and Taeb 2017, Hwang et al. 2018). This chapter tries to develop an alternative approach to estimate SOFA scores in real-time by leveraging machine learning techniques to mimic physicians' practice. The idea is to construct machine learning models to estimate the real-time values of less frequently updated laboratory test variables—in particular, creatinine, platelets, and bilirubin—(for simplicity, we would refer to these variables as "test variables") using easily measured bedside variables (for simplicity, we would refer to these variables as "bedside variables"). Then the predicted test variables would be used to calculate an estimate of the real-time SOFA score. The proposed approach can be viewed as an enhanced version of qSOFA that used machine learning models to quantify the link between bedside variables and test variables. We also incorporate the rate of changes in bedside variables and time lag from the previous test into these machine learning models. This mimics how physicians assessed patient conditions in practice and quantified the practice in a modeling framework to improve existing risk monitoring systems.

The new approach not only provides a point estimate of the real-time SOFA score, but we also take advantage of the estimated standard errors from these models to construct intervals of the real-time test variables and SOFA scores. We hypothesize that such intervals could capture the uncertainty in patients' health conditions since the previous test and provide valuable information from a new dimension that complements the point estimates of the SOFA scores. Based on the estimated uncertainty in health conditions—which would be a new dimension of information that complements the nominal risk level measured by the score—we could develop a more refined risk classification for patients and provide more precise recommendations for decision-making. Specifically, higher levels of estimated uncertainty could serve as a piece of evidence to administer necessary diagnostic tests, which would help address the issue of overtesting widely reported in practice (Clouzeau et al. 2019, Zhi et al.

2013b). Using a dataset collected from an ICU from a tertiary hospital in Singapore, we calibrate our model and validate the hypothesis by comparing the prognostic accuracy of the proposed approach on patients' 24-hour mortality and 30-day readmission with those from the SOFA score calculated using the conventional approach as well as qSOFA. The proposed methodology could be applied to other risk scores to improve their prognostic accuracy and improve their effectiveness in patient risk monitoring. We also calibrate our models to improve MODS and LODS.

2.2 Materials and Methods

2.2.1 Data

Patients' records are collected from the Cardiothoracic Intensive Care Unit (CTICU) in the National University Hospital (Singapore) from March 2010 to October 2016. The CTICU used the IntelliSpace Critical Care and Anesthesia (ICCA, a digital tracking system provided by Philips) system to track patients' clinical records in real-time automatically. We collect patients' demographic and clinical records throughout the ICU stay, including vital signs, laboratory tests, medications, electrocardiography (ECG), and nursing notes. This study is approved by the National Healthcare Group (NHG) Domain Specific Review Board (DSRB) (Reference number: 2016/00062).

2.2.2 Outcomes and Measures

The primary outcome of the study is 24-hour mortality, and the 30-day readmission is also measured. The prognostic accuracy of the scores is assessed from two perspectives: discriminative power and reclassification improvement. Discriminative power refers to the model's capability of differentiating patients with different risk categories. We adopt Nagelkerke's R Square and the area under the receiver operating characteristic (ROC) curve (AUC) to assess a model's discriminative power. Reclassification measures how well a new model reclassifies patients with different outcomes into correct risk categories than the benchmark model. We evaluate the reclassification improvement by constructing a reclassification table and deriving the Net Reclassification Improvement (NRI).

2.2.3 Models

We construct three ordinary least squares (OLS) models, using changes in bedside variables, rates of changes in bedside variables, and the time lag since the last laboratory test to estimate real-time values of test variables—i.e., creatinine, platelets and bilirubin—used in calculating SOFA scores. Let X_t denote bedside variables measured at time t, and Y_t denote a target test variable measured at time t. We use Figure 1 to illustrate the process of the variable updating and the scheme of our models. Figure 1 depicts a timeline with five time points in consideration, $t = t_1, t_2, t_3, t_4$ and t_5. Note that these time points are not necessary separated by equal intervals. Since X_t are easily measurable bedside variables, they could be updated at a high frequency (at any time point with real-time monitoring systems), which is represented in Figure 1 that we could observe X_t at $t = t_1$, t_2, t_3, t_4 and t_5. However, Y_t is a laboratory test variable that could only be updated when a test is conducted, so it is usually updated at a lower frequency compared to X_t. For example, in Figure 1, we could only observe Y_t at $t = t_1$ and t_4, but we don't observe the values of Y_t at $t = t_2, t_3$, and t_4. We use Δt to denote the time interval between the current time point t and the last time when the test variable was updated, e.g., $\Delta t_2 = t_2 - t_1$, $\Delta t_3 = t_3 - t_1$ (because the latest update in the test variable before t_3 happens at t_1). ΔX_t denotes the changes in bedside variables from the last time when the test variable is updated to the current time point t, e.g., $\Delta X_{t_2} = X_{t_2} - X_{t_1}$, $\Delta X_{t_3} = X_{t_3} - X_{t_1}$. We define ΔY_t similarly, but note that we do not observe ΔY_t for all t.

Our models aims to predict ΔY_t using changes in bedside variables ΔX_t, rates of changes in bedside variables $\frac{\Delta X_t}{\Delta t}$, and the time lag since the last laboratory test Δt

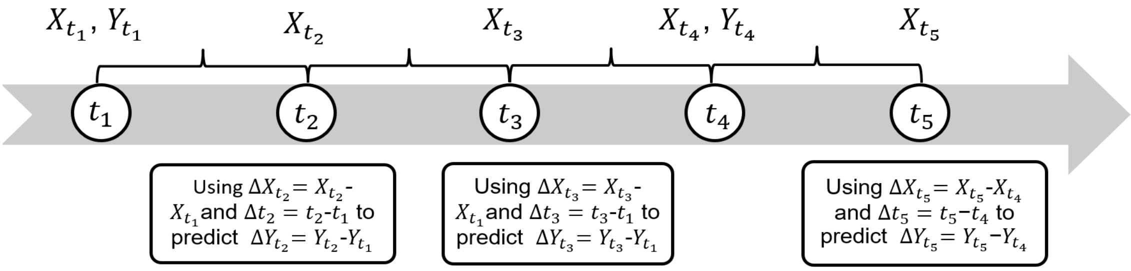

X_{t_1}, Y_{t_1}
X_{t_2}
X_{t_3}
X_{t_4}, Y_{t_4}
X_{t_5}
t_1
t_2
t_3
t_4
t_5
Using $\Delta X_{t_2}= X_{t_2}$-$X_{t_1}$and $\Delta t_2 = t_2$-t_1 to predict $\Delta Y_{t_2}= Y_{t_2}$-Y_{t_1}
Using $\Delta X_{t_3}= X_{t_3}$-$X_{t_1}$and $\Delta t_3 = t_3$-t_1 to predict $\Delta Y_{t_3}= Y_{t_3}$-Y_{t_1}
Using $\Delta X_{t_5}= X_{t_5}$-X_{t_4} and $\Delta t_5 = t_5 - t_4$ to predict $\Delta Y_{t_5}= Y_{t_5} - Y_{t_4}$

Figure 2.1: Process of variable updating and scheme of predictive models
Note: In this figure, X updates at every checking point. Y is updated only at time points t_1 and t_4. The model presented is built to predict Y at other time points.

to predict the change in the test variable ΔY_t, which could be used to estimate the current value of Y_t and then calculate an estimate of the current SOFA score for the patient. Our models are summarized in the following equation:

$$\Delta Y_t = \alpha + \beta \Delta X_t + \gamma \frac{\Delta X_t}{\Delta t} + \tau \Delta t + \epsilon,$$

where α, β, γ and τ are regression coefficients, and ϵ is the random error term that captured the unobservable factors that affect the changes in the test variables. The models also provided confidence intervals (CIs) for estimated ΔY_t, from which we could calculate the CIs and lengths of CIs (LoCIs) for estimated Y_t.

We further derive CIs and LoCIs for estimated SOFA scores. If multiple test variables are estimated at the same time point, we take the conservative approach when calculating the LoCIs for the estimated SOFA score, i.e., the worst values from the estimated CI of each test variable are taken to compute the worst value of the estimated SOFA score, and the best values from the estimated CI of each test variable are taken to compute the best value of the estimated SOFA score. In this approach, we estimate the longest possible LoCI of the estimated SOFA score. Specifically, we ignore correlations among the test variables[1] and use 98.3% CI of the three test variables and the conservative approach above to construct the 95% CI of the estimated SOFA score ($98.3\%^3 = 95\%$). All the reported CIs in this paper are

[1] The correlation coefficient of creatinine and bilirubin, of creatinine and platelets and of bilirubin and platelets are -0.06, 0.01 and 0.2, respectively.

95% CI of the estimated variables or scores.

2.3 Results

A total of 5,351 patients are enrolled in our study. Among these patients, the mean age is 59.9 years, and 3,960 (74.0%) are male. Of the study cohort, 263 (4.9%) died in the hospital, and 197 (3.7%) were readmitted within 30 days. Table 2.1 summarizes the demographics of the study population.

Characteristics	Values
No. of patients	5,351
Age, mean (SD) [range], year	59.9 (13.6) [13, 99]
Male, n (%)	3,960 (74.0)
Race, n (%)	
Chinese	3,504 (65.5)
Malay	807 (15.1)
Indian	434 (8.1)
Others	606 (11.3)
Died in hospital, n (%)	263 (4.9)
Readmitted in 30 days, n (%)	197 (3.7)

SD: standard deviation

Table 2.1: Demographics of the study population

We use a wide range of variables that could be easily measured or obtained at the bedside to predict test variables, including vital signs, results from bedside arterial blood gas (ABG) tests, bedside electrocardiograms, medication, and other readily available variables in real-time. In the study ICU, the turnaround time of ABG using point-of-care testing is less than 5 minutes. Note that we don't require all of these variables to be measured in real-time. Instead, our methods could be applied to update the real-time SOFA score and estimate intervals as long as one or more of these are updated. The list of complete beside variables is provided in the appendix.

We extract 746,357 time points of data when the beside variables are updated. Among these time points, creatinine is updated 27,872 times (5.11%), platelets 28,936 times (5.31%), and bilirubin 6,049 times (1.11%). The average intervals between consecutive measurements are 22.12 hours, 21.28 hours, 40.83 hours for creatinine, platelets, and bilirubin, respectively. More summary statistics of the test variables were provided in Appendix A.1. For each test variable, we carry out 5-fold cross-validation to select the combination of bedside variables that produces the lowest root mean squared error in predicting the test variable. Next, we retrain the predictive model for each test variable using the complete dataset. The final predictive models for all the test variables are provided in Appendix A.2.

When testing the performance of our proposed approach to construct point estimates and LoCI for real-time SOFA scores, we select the 445,753 time points of data when all the test variables are estimated as the test set. These data points represent the most challenging cases for the predictive models as only bedside variables are available, and the real-time SOFA scores have to rely on a lot of estimated values of the test variables. Among these test data points, 8,698 experienced in-hospital mortality within 24 hours, and 197 were readmitted to the ICU within 30 days.

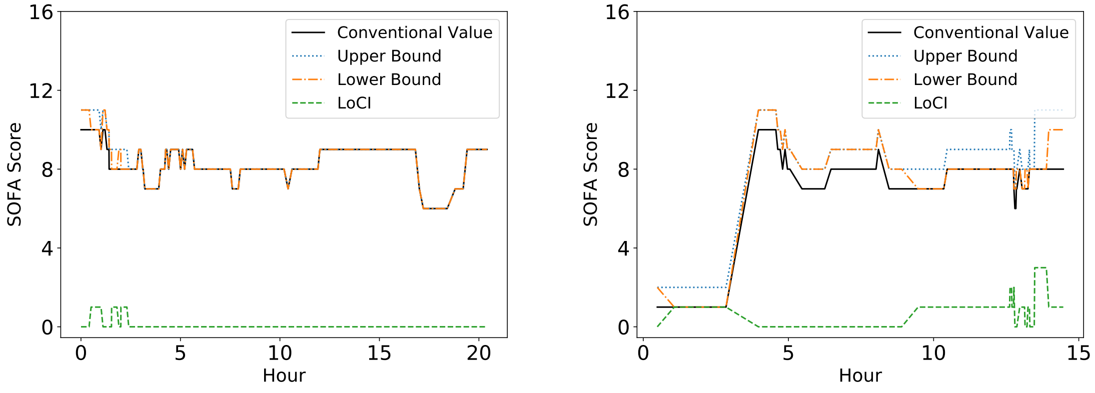
We first check whether the estimated LoCIs are predictive of 24-hour in-hospital mortality. Table 2.2 summarizes the estimated LoCIs for all the test variables and the estimated real-time SOFA score among different patient groups. The p-values are derived from Welch's t-tests. The estimated LoCIs of creatinine, platelets and SOFA score are significantly larger ($p < 0.001$) among patients who died within 24 hours than those who survived after 24 hours. These results show that the estimated LoCIs are indicative of patient conditions, and larger intervals are associated with worse outcomes in terms of 24-hour mortality. Before presenting the multivariate analyses to confirm the benefits of using estimated LoCI to monitor patient conditions, we plot the score trajectories of two typical patients during their ICU stays in Figure 2.2—one survived (Figure 2.2a), but the other died in the hospital (Figure 2.2b)—to illustrate the value of constructed score intervals.

Figure 2.2a shows that the patient's SOFA score calculated from the conventional

	All Patients (n = 445,753)	Survived $\geq$ 24 h (n = 437,055)	Died < 24 h Died (n = 8,698)	p-value
LoCI of estimated Creatinine	3.193	3.155	5.102	$< 0.001^{***}$
LoCI of estimated Platelets	3.306	3.268	5.249	$< 0.001^{***}$
LoCI of estimated Bilirubin	5.232	5.218	5.935	$< 0.01^{**}$
LoCI of estimated SOFA score	0.342	0.341	0.408	$< 0.001^{***}$

LoCI: length of confidence interval; SOFA: sequential organ failure assessment
(Significance Level: 0 ’***’; 0.001 ‘**’; 0.01 ‘*’)

Table 2.2: Estimated LoCIs of test variables and SOFA score among all patients, patients survived $\geq$ 24 hours, and patients died < 24 hours



(a) Survivor

(b) Death

Figure 2.2: Trajectories of SOFA scores and estimated real-time SOFA scores with intervals during two patients' ICU stays
Conventional Value: SOFA score calculated from the conventional approach without using estimated test variables
Upper Bound: constructed upper bound of the estimated real-time SOFA score.
Lower Bound: constructed lower bound of the estimated real-time SOFA score.
LoCI: length of confidence interval (LoCI) of the estimated real-time SOFA score, i.e., Upper Bound – Lower Bound.

approach is quite stable towards the end of the ICU stay, but the patient eventually died, which indicates that the SOFA score fails to provide a warning before the adverse event. However, the constructed intervals for the estimated SOFA score from our approach fluctuate during the patient's ICU stay and increases significantly towards the end of the stay. This demonstrates how the constructed intervals could identify the uncertainty in patient conditions when the laboratory tests are not conducted and provide early warning when the conventional SOFA calculation fails. For another patient showed in Figure 2.2b, the estimated intervals stay zero throughout the patient's ICU stay, which indicates our confidence in the real-time SOFA score, and the patient condition was truly stable.

To demonstrate the values of our proposed approach to estimate real-time SOFA scores and construct the intervals, we compare five logistic regression models in predicting patients' 24-hour mortality with different sets of predictors: (1) SOFA score calculated with the conventional method (for simplicity, we would refer to this as SOFA score below); (2) SOFA score and LoCI of estimated real-time SOFA score; (3) estimated real-time SOFA score and its LoCI; (4) estimated real-time SOFA score and estimated test variables' LoCIs; (5) qSOFA score. We refer to these models as Model 1, Model 2, Model 3, Model 4, and Model 5, respectively, and Model 1 is the base model. Table 2.3 summarizes the regression coefficients for these models.

We observe that in Models 2-4, the odds ratio of the LoCIs for estimated SOFA scores and test variables are all significantly greater than 1 ($p < 0.05$), implying that larger LoCIs were associated with a higher risk of 24-hour mortality. In Models 2 and 3, even after controlling for the SOFA score and estimated SOFA score, the LoCI of the estimated SOFA score is still statistically significant in predicting 24-hour mortality. This result confirms that the constructed intervals provide additional value in capturing patient health conditions, as hypothesized before. The LoCIs for the test variables also provide such additional values.

	Coefficients (n = 445,753)	Std. Error (n = 437,055)	Odds ratios Died (n = 8,698)	p-value
Model 1				
SOFA score	0.3860	0.0033	1.4712	< 0.001***
Model 2				
SOFA score	0.3935	0.0034	1.4822	< 0.001***
LoCI of estimated SOFA score	0.3714	0.0185	1.4497	< 0.001***
Model 3				
Estimated SOFA score	0.3983	0.0034	1.4894	< 0.001***
LoCI of estimated SOFA score	0.2711	0.0184	1.3115	< 0.001***
Model 4				
Estimated SOFA score	0.3854	0.0034	1.4703	< 0.001***
LoCI of estimated Creatinine	0.0954	0.0028	1.1001	< 0.001***
LoCI of estimated Platelet	0.0997	0.0023	1.1049	< 0.001***
LoCI of estimated Bilirubin	0.0014	0.0005	1.0014	< 0.05*
Model 5				
qSOFA score	1.2465	0.0132	3.4782	< 0.001***

LoCI: length of confidence interval; SOFA: sequential organ failure assessment; qSOFA: quick SOFA
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*')

Table 2.3: Regression coefficients of models on predicting patients' 24-hour mortality

To assess the first four models' predictive power, we first conduct the likelihood ratio tests (Steyerberg et al. 2012, Deeks and Altman 2004) to determine if the observed difference in the model fit is statistically significant. We find that Models 2, 3, and 4 exhibit significant deviance reductions ($p < 0.001$) compared to the base model (i.e., Model 1), in which the SOFA score is the only predictor. Next, we calculate these five model's Nagelkerke's R-squareds, which measure the goodness of fit of the logistic regression model to the data. We observe that Nagelkerke's R-squareds increase in the first four models (Table 3), indicating that using the estimated real-time SOFA score and adding the constructed intervals for the score and test variables improve the models' fit to the data. Model 5's Nagelkerke's R-squared is significantly worse ($p < 0.001$) than the other four models, indicating that the qSOFA score does not fit well in our study population and is inferior to the proposed approach leveraging bedside variables to estimate the real-time SOFA score. To assess these models' discriminative power, we generate their ROC curves (Figure 2.3) and compare their AUCs (Table 2.4). Model 4—which uses estimated real-time SOFA score and test variables' LoCIs—has the highest AUC. Again, qSOFA has the worst discriminative power in our study population.

	Model 1	Model 2	Model 3	Model 4	Model 5
Nagelkerke's R square	0.185	0.189	0.197	0.224	0.125
AUC	0.843	0.844	0.844	0.870	0.778

Model 1: using SOFA (sequential organ failure assessment) score only; Model 2: using SOFA score and LoCI (length of confidence interval) of estimated real-time SOFA score; Model 3: using estimated real-time SOFA score and its LoCI; Model 4: using estimated real-time SOFA score and estimated test variables' LoCIs; Model 5: using qSOFA (quick SOFA) score; AUC: area under the receiver operating characteristic curve

Table 2.4: Performance comparison between models on predicting patients' 24-hour mortality

Next, we focus on Model 4 and investigate its reclassification improvement over the base model. NRI is a popular measure to evaluate improved discrimination by a new model versus the benchmark model (Pencina et al., 2011, Steyerberg et al., 2012, Leening et al., 2014, Alba et al., 2017). We consider three risk levels for 24-hour mortality, 0–5%, 5–10%, and $> 10\%$, and develop the reclassification table as shown in Table 4. Among 8,698 data points in which patients died within 24 hours, Model 4 correctly reclassifies 2,101 (= 755 + 396 + 950) cases from lower risk levels to higher risk levels compared to the predictions from the base model; while it only makes 501 (= 301 + 0 + 200) worse predictions. Among 437,055 data points in which patients survived 24 hours, Model 4 makes 10,080 improved reclassifications and 9,965 worsened reclassifications. Overall, 18.42% cases are correctly reclassified by Model 4, i.e., NRI = 0.1842 ($p < 0.001$). We also calculate continuous NRI, which does not depend on the choice of risk-level cut-offs but captured the proportion of changes in predicted risk in the correct direction versus the proportion of changes in the wrong direction. The continuous NRI is 0.6960 (95% CI, 0.6751–0.7169, ($p < 0.001$)),

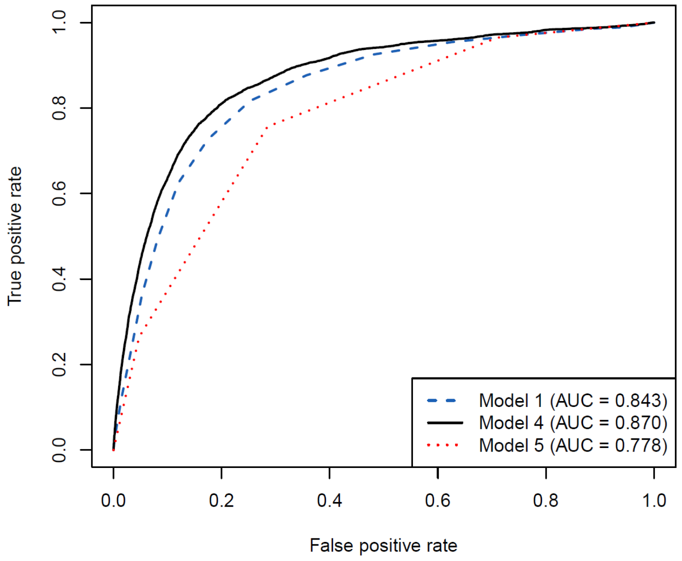


Figure 2.3: ROC curves of models on predicting patients' 24-hour mortality
Model 1: using SOFA (sequential organ failure assessment) score only; Model 2: using SOFA score and LoCI length of confidence interval) of estimated real-time SOFA score; Model 3: using estimated real-time SOFA score and its LoCI; Model 4: using estimated real-time SOFA score and estimated test variables' LoCIs; Model 5: using qSOFA (quick SOFA) score; AUC: area under the receiver operating characteristic curve

which indicates the significantly improved discriminative capability of our model. We further compute the integrated discrimination improvement (IDI), which integrates the NRI over all possible risk-level cut-offs and is mathematically equivalent to the difference in discrimination slopes of the two models in comparison (identical to the difference in Pearson R-squared) (Pencina et al., 2012). The IDI is 0.0344 (95% CI, 0.0325–0.0363, ($p < 0.001$)), which further confirms the superiority of Model 4 over the base model.

We carry out the same comparison between the models for 30-day readmission, which is known to be a tough outcome to predict for ICU patients (Desautels et al. 2017). From Table 2.6, it is observed that 30-day readmission is indeed challenging to predict, but our method, especially Model 4, significantly improves the model fit and discriminative power over the base model. The results confirm the importance of capturing the uncertainty in patient conditions through the estimated intervals for

	Predicted risk from Model 4, %			
Predicted risk from Model 1, %	0-5	5-10	>10	Total
Died < 24 h				
0-5	3,267	755	396	4418
5-10	301	1,094	950	2345
>10	0	200	1,735	1935
Total	3,568	2,049	3,081	8698
Survived ≥ 24 h				
0-5	394,258	5,749	1,496	387898
5-10	6,362	12,978	2,720	18023
>10	1	3,717	9,774	39832
Total	400,621	22,444	13,990	437055

Model 1: using SOFA (sequential organ failure assessment) score only; Model 4: using estimated real-time SOFA score and estimated test variables' LoCIs (lengths of confidence intervals); Dark grey shaded numbers indicated improved reclassification by Model 4 over Model 1; Light grey shaded numbers indicated worsened reclassification by Model 4 over Model 1.

Table 2.5: Reclassification table (risk categories: $0-5\%$, $5-10\%$, $>10\%$)

lagged test variables. We also consider other outcome measures, including 12- and 36-hour mortality. The findings are consistent with other outcomes and confirm the superior performance of Model 4. Detailed results are presented in Table 2.7.

To validate that our method can improve other risk scores, we consider another two commonly used scores in ICUs, MODS and LODS. The test variables used in MODS are the same as SOFA, but LODS uses three more test variables, urea, white blood cell and prothrombin time. We carry out the same procedure to train the models to predict the new test variables, and the final models are presented in the appendix. Next, we conduct the same comparison between four logistic regression models to predict the outcomes using different sets of predictors: (1) the original score (either MODS or LODS) calculated with the conventional method (for simplicity, we would refer to this as the original score below); (2) the original score and LoCI of the estimated real-time score; (3) estimated real-time score and its LoCI; (4) estimated

	Model 1	Model 2	Model 3	Model 4	Model 5
Nagelkerke's R Square	0.002	0.002	0.002	0.041	0.001
AUC	0.550	0.550	0.550	0.663	0.518

Model 1: using SOFA (sequential organ failure assessment) score only; Model 2: using SOFA score and LoCI (length of confidence interval) of estimated real-time SOFA score; Model 3: using estimated real-time SOFA score and its LoCI; Model 4: using estimated real-time SOFA score and estimated test variables' LoCIs; Model 5: using qSOFA (quick SOFA) score; AUC: area under the receiver operating characteristic curve.

Table 2.6: Performance comparison between models on predicting 30-day readmission

Prediction	Evaluation	Model 1	Model 2	Model 3	Model 4	Model 5
12-hour mortality	Nagelkerke's R Square	0.190	0.199	0.211	0.255	0.129
	AUC	0.866	0.866	0.872	0.898	0.801
36-hour mortality	Nagelkerke's R Square	0.188	0.192	0.197	0.228	0.124
	AUC	0.833	0.833	0.833	0.856	0.766

Model 1: using SOFA (sequential organ failure assessment) score only; Model 2: using SOFA score and LoCI (length of confidence interval) of estimated real-time SOFA score; Model 3: using estimated real-time SOFA score and its LoCI; Model 4: using estimated real-time SOFA score and estimated test variables' LoCIs; Model 5: using qSOFA (quick SOFA) score; AUC: area under the receiver operating characteristic curve

Table 2.7: Performance comparison between models on predicting patients' 12-hour, 36-hour mortality

real-time score and estimated test variables' LoCIs. The detailed results are presented in Table 2.8. We observe the same findings—using the estimated real-time score with constructed intervals for test variables (i.e., Model 4) performed the best in all outcomes and significantly improved the base model (i.e., Model 1) for both MODS and LODS.

2.4 Discussion

The onset of organ dysfunction is associated with higher mortality and complication rates, making timely detection of deterioration and intervention crucial to critically ill patients. The SOFA score has been validated as an effective risk measure of patient conditions and is widely adopted in clinical practice for risk monitoring. Nevertheless,

Score	Prediction	Evaluation	Model 1	Model 2	Model 3	Model 4
MODS	12-hour mortality	Nagelkerke's R Square	0.210	0.210	0.218	0.238
		AUC	0.872	0.873	0.878	0.897
	24-hour mortality	Nagelkerke's R Square	0.205	0.206	0.209	0.231
		AUC	0.852	0.852	0.854	0.874
	36-hour mortality	Nagelkerke's R Square	0.211	0.212	0.214	0.234
		AUC	0.845	0.845	0.846	0.964
	30-day readmission	Nagelkerke's R Square	0.003	0.003	0.003	0.012
		AUC	0.552	0.551	0.549	0.647
LODS	12-hour mortality	Nagelkerke's R Square	0.258	0.261	0.267	0.339
		AUC	0.894	0.895	0.919	0.920
	24-hour mortality	Nagelkerke's R Square	0.242	0.244	0.249	0.297
		AUC	0.870	0.871	0.874	0.892
	36-hour mortality	Nagelkerke's R Square	0.235	0.237	0.243	0.280
		AUC	0.856	0.857	0.860	0.876
	30-day readmission	Nagelkerke's R Square	0.005	0.004	0.005	0.121
		AUC	0.578	0.581	0.584	0.666

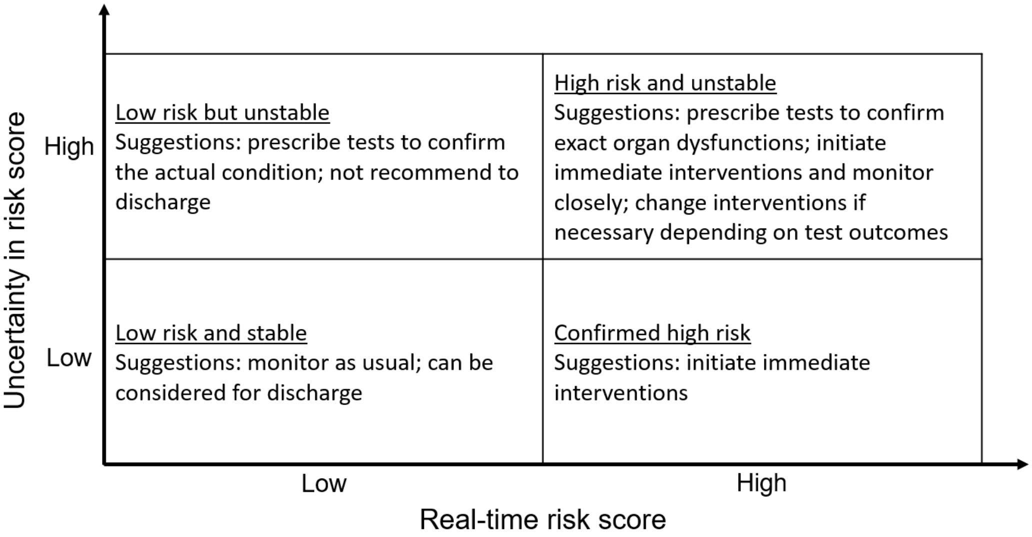
Model 1: using the original score score only; Model 2: using the original score and LoCI (length of confidence interval) of estimated real-time score; Model 3: using estimated real-time score and its LoCI; Model 4: using estimated real-time score and estimated test variables' LoCIs; AUC: area under the receiver operating characteristic curve

Table 2.8: Performance comparison between models on predicting different outcomes for MODS and LODS

the effectiveness of the SOFA score as a risk monitoring tool has been questioned, partially because some key clinical variables (creatinine, bilirubin and platelets) used in calculating the SOFA score are typically not frequently tested in clinical practice. Hence, it is essential to establish a risk monitoring system that captures patient conditions in real-time for organ function surveillance (Ferreira et al. 2001). qSOFA was proposed to address this issue using a handful of bedside variables and easy calculation. Its simplistic structure, however, rendered its effectiveness in capturing patient conditions. We identifies this critical time lag in calculating SOFA scores and potentially many other risk scores. To address this issue, we develop a framework to mimic how physicians assessed patient conditions in practice and develop machine learning models to estimate the real-time values of the test variables using easily obtained bedside variables and information. We further leverage these models to construct intervals for test variables to quantify the uncertainty in conditions a patient might have developed since the previous test.

Shickel et al. (2019)) and Aşuroğlu and Oğul 2021 recently demonstrated that using latent and frequently updated clinical information to predict the real-time SOFA score could improve the effectiveness of the SOFA score under deep learning frameworks. Our study goes a step further, showing that capturing the uncertainty in patient conditions can also be critical for risk monitoring systems. Without real-time laboratory tests, a patient's true condition is uncertain and could only be partially reflected by other observable information. In practice, physicians attend to the patients and assess their conditions using information available at the bedside as well as lagged laboratory test reports. The physicians might form some estimates about the real-time value of those unobservable variables—which, even if tested, could only be available after some time—and make decisions based on their estimates. During such a process, the physicians are aware of the uncertainty in their estimates and would certainly take such uncertainty or confidence of their estimates into their decision-making process. Our proposed modeling framework mimick such processes and quantify both the estimates and the uncertainties in the estimates of test variables through constructed confidence intervals. We demonstrate that these intervals could accurately reflect the uncertainties in patient conditions and provide additional value in real-time patient risk monitoring.

With enhanced patient monitoring systems that quantify the uncertainties in risk assessment, we can refine the guidelines for decision-making interventions. Figure 2.4 proposes a refined patient classification based on real-time estimated risk score (e.g., SOFA, MODS, LODS) and estimates uncertainties in risk assessment, which could be obtained from constructed intervals from our approach. This is an exemplary scheme, and one could certainly have more granular stratification in each dimension. For patients in different risk groups (Figure 2.4), we propose recommended decisions correspondingly. If a patient is assessed with low risk and low uncertainty, the patient is in relatively good and stable condition; physicians can continue to monitor the patient as usual, and such a patient could be considered for discharge from ICU. If a patient is assessed with low risk but high uncertainty, the patient's condition is not stable, and there is a nonnegligible chance that the patient's actual condition is worse; physicians are advised to prescribe relevant tests to find out the actual condition based on the variables which have large estimated intervals; the patient should not be considered for discharge. If a patient is assessed with high risk and low uncertainty, physicians can confirm that the patient is in poor condition with known issues revealed by the score, and necessary interventions should be taken immediately. If a patient is assessed with high risk and high uncertainty, the actual condition could be even worse, and there could be unknown organ dysfunctions; the immediate recommendation is to prescribe relevant tests to confirm the exact organ dysfunctions and initiate some interventions; at the same time, the patient should be monitored closely, and appropriate adjustments to the interventions should be taken if necessary depending on the test outcomes.



	Low	High
High	Low risk but unstable Suggestions: prescribe tests to confirm the actual condition; not recommend to discharge	High risk and unstable Suggestions: prescribe tests to confirm exact organ dysfunctions; initiate immediate interventions and monitor closely; change interventions if necessary depending on test outcomes
Low	Low risk and stable Suggestions: monitor as usual; can be considered for discharge	Confirmed high risk Suggestions: initiate immediate interventions

Figure 2.4: Intervene instructions

Leveraging the estimated uncertainty in patient conditions, this chapter sheds light on the clinical practice regarding when to order laboratory tests. In practice, many laboratory tests are conducted routinely in ICU to monitor patient conditions. For example, in our study ICU, a renal panel is usually performed once a day. It could be performed more frequently (e.g., twice a day) for relatively high-risk patients but with fixed and equally spread intervals throughout the day. Moreover, it is widely reported that many laboratory tests are inappropriately used in practice (Clouzeau et al. 2019). For example, 20%-30% of diagnostic tests are identified as inappropriate in a meta-analysis (Zhi et al. 2013b). Our methods could be viewed as a tool to trigger on-demand laboratory tests for patients with estimated high uncertainties in their conditions. With the integration of real-time clinical information, the enhanced risk monitoring system could provide validated real-time risk assessment with validated uncertainties in such assessment to the physicians. The need for any (additional) laboratory tests would then be based on quantified and validated evidence, which can reduce unnecessary and repeated testing.

Our study has the following limitations. First, our results are based on analyzing the data from a single center. Although the methodology is general enough to enhance any risk scores, the improvement might differ in other units. Second, we only test the enhancement for a few risk scores, i.e., SOFA, MODS and LODS. There are many other risk scores developed in the literature and implemented in practice, which we could not exhaust in this study. Finally, the study is based on a retrospective patient cohort. A prospective randomized clinical trial would be required to confirm and validate the benefits of enhanced risk monitoring systems.

2.5 Conclusion

This chapter proposes a new framework to estimate the real-time values of risk assessment scores for patient monitoring in ICUs. We develop machine learning models to estimate the real-time values of the laboratory test variables used to calculate the scores and then compute estimates of real-time score values. We leverage the estimated confidence intervals in the test variables to quantify the uncertainty in patients' health conditions from the previous tests. We validate that such intervals provide significant additional values for risk monitoring in real-time. The proposed method provides a foundation for finer patient risk classification and decision recommendation. In particular, the estimated uncertainty in patients' health conditions could be used to trigger on-demand laboratory tests. External validation and clinical tries are warranted to confirm the benefits of enhanced risk monitoring systems.

Chapter 3

A Predictive and Prescriptive Method to Reduce Repetitive Tests

3.1 Introduction

The global healthcare expenditure has risen drastically in recent years, yielding a higher annual growth rate relative to Gross National Product (Hernández-Peña, 2019). Due to this unsustainable phenomenon, there is a growing interest in promoting high-value medical care initiatives, in which the overuse of diagnostic tests has attracted significant attention (Eaton et al. 2017). Notably, the global clinical diagnostic service market size has reached \$200.3 billion in 2020 and is expected to expand at a compound annual growth rate of 4.7% from 2021 to 2028 (Grand View Research 2021), while 20% to 30% of all diagnostic tests are identified as inappropriate (Zhi et al. 2013a). In addition to the heavy financial burden that the overuse problem imposes on patients and the healthcare system, it also results in many negative consequences such as unnecessary patient discomfort, excessive utilization of phlebotomy, decreased hemoglobin values, and hospital-acquired anemia, which further leads to additional blood transfusions, increased mortality, and prolonged length of stay (LOS) for hospitalized patients (Eaton et al. 2017, Cheng et al. 2019). Therefore, it is crucial to address the diagnostic overutilization issue from both economic and health considerations.

The overutilization of diagnostic tests, most of which are in the form of unnecessary repetitive tests, is particularly prevalent in intensive care units (ICUs) (Beriault et al. 2021). ICUs cater to critically ill patients who are in need of close monitoring, imminent intervention, and life-sustaining treatment to restore or maintain organ function. In ICUs, timely tracking of patient physiologic indicators is essential for physicians to identify potential organ dysfunction and acute deterioration of patients. For example, systolic blood pressure at or below 90 $mmHg$ or diastolic blood pressure at or below 60 $mmHg$ is typically considered hypotension, which leads to inadequate blood flow to body organs and may contribute to stroke, heart attack, kidney failure and shock if not promptly intervened (Chalmers et al., 2008, Feldstein and Weder, 2012). Glucose level serves as a key indicator of hyperglycemia, that if a patient's glucose level is above 10 $mmol/L$, insulin therapy should be initiated immediately in case of increased mortality and hyperglycemia-associated complication rates (Ichai and Preiser, 2010). Although patients' basic physiological parameters (e.g., blood pressure, respiratory rate) can be tracked bedside with the real-time monitoring system, there are still various critical indicators that require invasive or expensive diagnostic tests (e.g., white blood cell counts, glucose level). In fact, diagnostic test results leverage 60% to 70% of all critical decisions, including medications, further testing and discharges (Forsman, 1996). Repetitive tests enable physicians to identify patients' physical problems and confirm their condition, while failure in the timely update of key indicators may adversely contribute to delayed treatments, prolonged LOS and higher mortality rate (Ong et al., 2018).

In addition, the overutilization of diagnostic tests also stems from the smart healthcare system. With the advancement of information technology, the smart healthcare market has been rapidly emerging in recent years. Smart healthcare, which aims to provide more convenient and personalized healthcare services as well as achieve efficient and effective utilization of healthcare resources, has gained extensive attention owing to the growing demand for early detection of diseases and preventive care (Tian et al. 2019). For example, the Singapore government contracted with Fitbit, Inc., which is a smart healthcare provider, to provide health trackers and services to up to one million Singapore citizens as part of the development of the smart city ecosystem (Reuters, 2019). Serving as an essential pillar in the modern concept of smart city, smart wearable health devices (WHDs) have evolved to respond to the demands of early detection and preventive care by providing real-time health status monitoring services (Aziz et al., 2016). For instance, Apple Watch provides early notification for atrial fibrillation by real-time tracking the user's heart rate and heart rate variability using photoplethysmography; HeartGuide, which is also a smart WHD, provides early notification of hypertension by real-time tracking users' blood pressure. Individuals who receive early warnings of deterioration may subsequently seek further medical assistance. However, as pointed out by Baig and Gholamhosseini (2013) and Vogt et al. (2019), the false positive (FP) rates of smart WHDs are very high, which could consequently give rise to hospital congestion and waste of medical resources.

Thus motivated, we aim to balance the trade-off between the overutilization of diagnostic tests and the delayed identification of individuals' deterioration. In this chapter, we adopt a POMDP framework to characterize the optimal time to prescribe a diagnostic test. Specifically, we focus on acute diseases and put our sights on real-time health monitoring contexts, where inpatients or users continuously or frequently receive routine, harmless, and cheap (or free) medical tests (e.g., heart rate and blood pressure tests). Given that the individual's critical clinical indicators may not be frequently updated, the true progression of the patient's status cannot be fully observed in real-time. Utilizing all frequently updated physiological parameters as predictors, we propose a real-time prediction model to estimate key indicators for some acute diseases (e.g., glucose levels for hyperglycemia) and calculate an interval to capture the uncertainty of the prediction. We further embed the predictive model into the modeling framework to determine whether to order a diagnostic test in each decision epoch.

We illustrate the performance of our framework through comprehensive simulations and case studies. Specifically, we study the optimal testing policy for detecting hyperglycemia in an ICU with real data. We demonstrate the benefits of the proposed model compared to practical benchmark policies across multiple metrics (e.g., the number of diagnostic tests, detection time).

This chapter makes the following key contributions:

(1) To the best of our knowledge, this is the first work in operations management to address the overutilization problem of diagnostic tests in healthcare services using real-time predictive analytics. We propose a POMDP framework that incorporates a prediction model to characterize the optimal prescription diagnostic tests.

(2) In contrast to the static or age-specific information matrices (also known as observation matrices) that are extensively applied in the disease screening literature, we develop a prediction model that integrates the time lag of key indicators (i.e., the time elapsed from the last testing epoch) and the dynamic changes in patient physiological parameters to construct the observation consisting of a predicted outcome and an uncertainty interval. We then estimate the information matrices with respect to the time lag to capture the time effect of the prediction. In addition, we consider a maximum allowed time lag in the modeling framework (i.e., patients are required to receive at least one diagnostic test within a given time period). All these features differentiate our study from conventional POMDP models and render it more in line with real practices.

(3) Based on results from simulation and the case study, we demonstrate the significant benefits of the optimal testing policy generated by our modeling framework

compared with some practical benchmark policies. Specifically, our framework can advance the detection time with few tests in hyperglycemia detection.

3.2 Literature Review

Our work, which adopts predictive and prescriptive methods to reduce repetitive diagnostic tests, is closely related to the literature on optimizing diagnostic testing policy in both medical and management areas.

There are extensive researches on reducing unnecessary diagnostic tests for acute diseases in the medical area. Most of them focus on deriving heuristic policies that are easy to implement to tackle the overuse problem. For example, Le Maguet et al. (2015), Iturrate et al. (2016) and Dhanani et al. (2018) address the overutilization of diagnostic testing problem by changing the culture of tests ordering and/or the features of electronic ordering systems, and they illustrate the effectiveness of the proposed methods by experimental studies. Medical researchers also adopt simulation methods to derive the diagnostic testing policy by evaluating the performance of various routine and heuristic testing policies. However, they focus on the utilization of diagnostic testing in the context of chronic disease. For example, to balance the mortality rate, life-years gained, and relevant cost, Boer et al. (1998) propose a computer simulation model to evaluate the performances of different breast screening policies. Michaelson et al. (1999) develop a simulation model to determine the optimal screening policy by comparing the cost-benefit consequences of different screening intervals. McLay et al. (2010) develop a simulation-optimization model which suggests age-based screening optimal policies for cervical cancer under a cost-effectiveness framework. They state that their proposed dynamic policy can approximately achieve the same health benefit with fewer scheduled screenings when compared with the recommended policy. For a comprehensive review of medical papers on the proper use of diagnostic tests, please refer to Bindraban et al. (2018) and Koleva-Kolarova et al.

(2015). Our work is greatly different from these researches as we focus on optimizing diagnostic testing strategies with analytical optimization models in acute disease contexts.

Operations research and optimization techniques have also been widely applied in developing diagnostic testing strategies with analytical models. To the best of our knowledge, most prior works on diagnostic testing in the operations research literature focus on optimizing screening tests for chronic disease. For example, Lee and Pierskalla (1988) investigate the diagnostic screening problem for contagious diseases by developing a screening optimization model with little or no latent periods. They derive the optimal screening policy by minimizing the average number of infected people in the population. Özekici and Pliska (1991) study the optimal inspection problem with a delayed Markov process. Their objective is to minimize the total expected cost associated with the sensitivity and specificity of the test. Assuming stationary disease aggression, they apply their model to a cancer screening setting and derive the optimal screening policy. Parmigiani (1993) propose a continuous-time non-Markovian stochastic model to analyze the diagnostic testing problem with error-free tests and derive the optimal inspection policy from the cost minimization perspective. For breast cancer biopsy based on mammography observations, Chhatwal et al. (2010) use a Markov decision process to model the problem and compute the optimal screening policy by maximizing the expected quality-adjusted life years (QALYs). However, this chapter focuses on the overuse of repetitive diagnostic tests in acute disease settings where patients are in emergent status and usually require lots of diagnostic tests for assessment. Our work further differs by modeling the diagnostic testing problem with a predictive model embedded POMDP framework.

A substream of the literature that is closely related to our work is about the use of POMDP in developing optimal screening policies for various cancers. Utilizing the sensitivity and specificity of mammography as observation probabilities, Maillart et al. (2008) evaluate some age-dependent screening policies using a POMDP model with sample-path enumeration in a breast cancer screening problem. They try to identify efficient screening policies for different patient groups to balance the associated mortality risk of breast cancer and the implementation effort of screening. Ayer et al. (2012) propose a POMDP approach to personalize mammography screening decisions based on patients' historical screening results and risk characteristics. Given that mammography screening has high potential risks (e.g., high FP rate), in their paper, they try to maximize the total expected QALY for each patient by considering the costs associated with the QALY losses of underuse and overuse of mammography screening. Under a similar setting, Ayer et al. (2016) further incorporate imperfect and heterogeneous adherence behaviors and breast cancer risk in their model to derive the optimal screening policy. Similarly, Zhang et al. (2012) and Erenay et al. (2014) develop POMDP models to address the screening problem for prostate cancer and colorectal cancer, respectively. They derive the optimal screening policies by maximizing the expected QALYs by considering the QALYs loss associated with one-time screening and long-term benefits. We refer the readers to Alagoz (2011) for a more comprehensive review of the application of POMDP in screening problems.

In line with this stream of literature, we also develop a POMDP framework to address the overuse of medical tests problem. However, our work differs from these works in one or more critical ways. First, instead of studying the screening problems under chronic disease settings, we focus on acute disease contexts. Second, we consider a maximum allowed test interval in the testing policy, which renders it more in line with real practices and greatly differentiates it from conventional POMDP models. Third, we embed a predictive model with real-time patients' risk data in the POMDP framework and construct the information matrices based on the predictions. However, the mentioned POMDP application literature usually constructs the information matrix considering the specificity and sensitivity of diagnostic tests. Moreover, as pointed out by Jenkins et al. (2018), clinical prediction models are quickly becoming outdated and less accurate over time due to accumulated uncertainty. We consider this new feature in our predictive model. As a result, our information matrices are also time-dependent, which further distinguishes our model from other conventional POMDP models.

Our work is also related to the literature on developing predictive and prescriptive methods for optimal diagnostic testing control. For example, To address the proper use of lab testing problems in the hospital, Cismondi et al. (2013) develop a binary classifier using fuzzy modeling to determine whether a lab test should be administered or not based on the information gain from the predicted lab results. However, they do not optimize the decision based on the predictive information. Cheng et al. (2019) consider the optimal diagnostic testing problem with an MDP—Reinforcement Learning (RL) framework. They first adapt the multi-output Gaussian process to derive the hourly predictions of non-frequent update clinical variables. Second, they use MDP to model patient trajectories by using the predictions as the state. In their model, they construct a vector-valued reward function to address the trade-off among different objectives. Finally, they use the reinforcement learning method to derive the lab testing policy. Our work critically differs from Cheng et al. (2019) in the following aspects. First, we use a POMDP analytical framework to model the problem with considering patients' states are partially observable. Second, instead of only considering the time lag in the state space, we further consider its impact on prediction accuracy. Last but not the least, they apply a vector-valued reward function to address the trade-off between the over-ordering and under-ordering of any given lab test; however, we focus on a one dimension reward function. This chapter also broadly relates to literature on developing predictive and prescriptive framework for healthcare services, e.g., Bertsimas et al. (2016), Spencer et al. (2014), Xu (2015), and Xu and Chan (2016) use predictive information to optimize healthcare decisions in different contexts. Similar to these works, we aim to incorporate real-time predictive information into dynamic prescriptive analytics to provide real-time decision support on recommending diagnostic tests. Our model features the dynamic change in the accuracy of predictive information and captures the practice of maximum recommended test interval, and thus differs from the methodologies developed in the above papers.

Outline of this chapter: The rest of the paper is organized as follows. Relevant literature is reviewed in Section 3.2. We describe the problem setting and framework with the POMDP model in Section 3.3. In Section 3.4, we convert the optimality equations of the POMDP model into an alternative form and propose a modified pruning algorithm based on the new representation. In Section 3.5, we present extensive numerical experiments to evaluate the performance of the proposed model and policy. Section 3.6 applies our framework to blood glucose monitoring in an ICU and evaluates the model's performance using real data. We conclude our work with a discussion of limitations in section 3.7.

3.3 Model

In ICUs, while patients' vital signs can be monitored in real-time at the bedside via electronic monitoring systems, indicators that require diagnostic tests (e.g., glucose level, creatinine level) cannot be directly observed, that physicians need to order diagnostic tests to learn the exact health status of patients (e.g., have hyperglycemia or not). However, some of these indicators serve as key signs for some acute diseases that require timely initiation of interventions (e.g., glucose levels for hyperglycemia, creatinine levels for acute kidney disease). As noted earlier, frequent diagnostic testing may increase patients' financial burden and clinical risk, while the underuse of diagnostic tests may delay the detection of disease progression and subsequently lead to delayed treatment. To balance the trade-off between test overutilization and delayed treatment, we formulate it as a sequential decision model based on a discrete-time

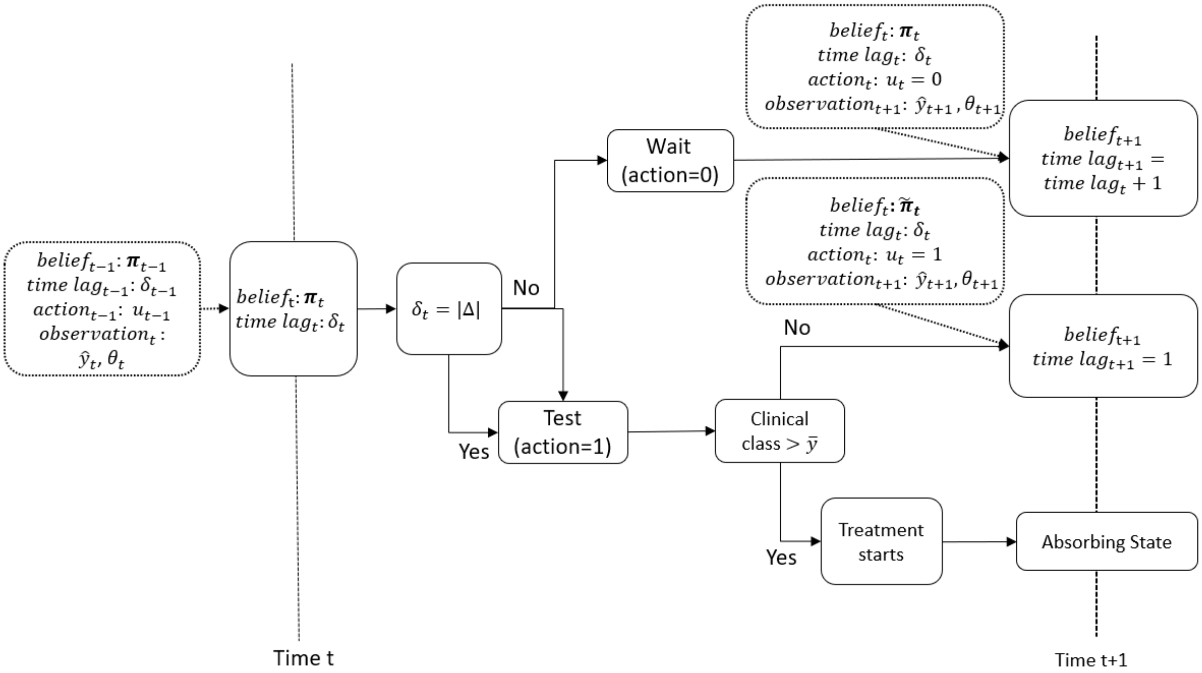

$belief_{t-1}$: $\boldsymbol{\pi}_{t-1}$
$time\ lag_{t-1}$: δ_{t-1}
$action_{t-1}$: u_{t-1}
$observation_t$: $\hat{y}_t, \theta_t$
$belief_t$: $\boldsymbol{\pi}_t$
$time\ lag_t$: δ_t
$\delta_t = |\Delta|$
No
Yes
Wait (action=0)
Test (action=1)
$belief_t$: $\boldsymbol{\pi}_t$
$time\ lag_t$: δ_t
$action_t$: $u_t = 0$
$observation_{t+1}$: $\hat{y}_{t+1}, \theta_{t+1}$
$belief_t$: $\tilde{\boldsymbol{\pi}}_t$
$time\ lag_t$: δ_t
$action_t$: $u_t = 1$
$observation_{t+1}$: $\hat{y}_{t+1}, \theta_{t+1}$
$belief_{t+1}$
$time\ lag_{t+1} = time\ lag_t + 1$
$belief_{t+1}$
$time\ lag_{t+1} = 1$
Clinical class > $\bar{y}$
No
Yes
Treatment starts
Absorbing State
Time t
Time t+1

Figure 3.1: Event Flow

finite-horizon POMDP framework.

Suppose the true value of the key indicator, which reflects a patient's health status, is known before the process starts. Physicians hold prior beliefs about patients' health status. In the following epochs, physicians have to determine whether to order a diagnostic test to update the key indicator based on frequently updated information and patients' historical information. Specifically, we leverage such information to predict the key indicator and introduce an interval to capture the uncertainty of the prediction. Notably, the time lag from the last testing epoch could possibly influence the prediction accuracy; therefore, we also incorporate this feature in our framework, and we further introduce a maximum allowed time lag (i.e., patients are required to receive at least one diagnostic test within a given time period). After observing the predictive information, physicians will update their beliefs about patients' health conditions and subsequently decide whether to prescribe a diagnostic test. If a test is ordered and the result suggests the necessity of treatment, the patient will receive the treatment and the system transits to the absorbing state. Otherwise, the system state and information will be updated, and the system will progress to the next epoch until the end of the time horizon. Concretely, we illustrate the brief event flow in Figure 3.1 and define the notations of the POMDP model as follows.

Time horizon: $t \in \mathcal{T}$, where $\mathcal{T} = \{0, 1, 2, 3, ..., T\}$.

Action space: $u_t \in \mathcal{U}$, where $\mathcal{U} = \{0, 1\}$. $u_t = 1$ denotes the decision to order a diagnostic test in epoch t, and $u_t = 0$ denotes the decision to wait until the next decision epoch. Specifically, we assume that all patients receive a diagnostic test in epoch 0 prior to their admissions into ICUs, i.e., $u_0 = 1$.

State space: Let r_t denote the true value of the key indicator of patients in epoch t, and we classify patients into one of Y classes based on r_t. If a patient is under treatment, we use A to denote the clinical class. Without loss of generality, let $y_t \in \mathcal{Y} = \{1, 2, ..., Y, A\}$ denote patients' clinical class in epoch t, where class 1 represents the best health condition while class Y corresponds to the worst health condition.

We define a patient's state as $\boldsymbol{s}_t$, where $\boldsymbol{s}_t \in \mathcal{S} \equiv \{(y_t, \delta_t) | y_t \in \mathcal{Y},\ \delta_t \in \Delta\}$. $\delta_t \in \Delta = \{1, 2, ..., |\Delta|\}$ represents the time lag from the last testing epoch, i.e., $\delta_t = t - t^*$, where $t^* = \max\limits_{k=0,1,2,...,t-1} \{k | u_k = 1\}$. Specifically, we assume that when $\delta_t = |\Delta|$, patient will receive a mandatory diagnostic test, which is consistent with current practice that patients are recommended to receive at least one diagnostic test in a given time period (e.g., annual physical examination, routine blood tests in ICUs). Note that δ and A are completely observable, whereas $y_t \in \mathcal{Y} \backslash \{A\}$ is not completely observable.

Suppose $\bar{y}$ is the predefined intervention threshold, that if a diagnostic test is conducted and finds that the patient's true clinical class y_t is not better than $\bar{y}$ (i.e., $y_t \geq \bar{y}$), physicians will intervene immediately with appropriate treatment, and the patient state will move to the absorbing state $(A, 1)$ in the next epoch. We assume that at the last epoch $t = T$, patients will undergo a mandatory diagnostic test if they are not in the absorbing state, i.e., $u_T = 1$ if $s_T \neq (A, 1)$. We conclude that the set of possible actions in core state $\boldsymbol{s}_t \in \mathcal{S}$ $(t < T)$ is as follows:

$$\mathcal{U}_t = \begin{cases} \{0\} & \text{if } s_t = (A, 1) \\ \{1\} & \text{if } s_t \neq (A, 1), \delta_t = |\Delta| \\ \{0, 1\} & \text{otherwise.} \end{cases}$$

Belief state: $\boldsymbol{\pi}_t \in \Pi \equiv \{\boldsymbol{\pi} | \sum_i \pi(i) = 1, \pi(i) \geq 0, i \in \mathcal{Y}\}$. $\boldsymbol{\pi}_t$ denotes the probability distribution over the space of the patient's clinical class. For $i \in \mathcal{Y}$, $\pi_t(i)$ represents the probability that the patient's clinical class is A in epoch t. Because the absorbing state is observable, we have either $\pi_t(A) = 1$ or $\pi_t(A) = 0$.

Observation space: As mentioned earlier, we can not know a patient's key indicator r_t without performing a diagnostic test. However, the fluctuations of clinical results could potentially be associated with changes in some intensively monitored variables, such as heart rate, blood pressure and respiratory rate. Thus, we use the changes in the frequently updated variables and rates of changes in these variables, which reflect or partially reflect the disease progression, to predict the change in the key indicator. Moreover, to capture the time effect of information, we also incorporate δ_t into the prediction model. The prediction is based on an ordinary least squares (OLS) model:

$$\hat{r}_t = f(r_{t^*}, \delta_t, \boldsymbol{x}_{t^*}, \boldsymbol{x}_t) = r_{t^*} + \boldsymbol{a}^{'}(\boldsymbol{x}_t - \boldsymbol{x}_{t^*}) + \boldsymbol{b}^{'}\frac{(\boldsymbol{x}_t - \boldsymbol{x}_{t^*})}{\delta_t} + c\delta_t + d,$$

where r_{t^*} denotes the latest measurement of the key indicator prior to epoch t and $\boldsymbol{x}_{t^*}$ denotes the latest measurement of predictive variables. Based on the predicted result $\hat{r}_t$ and the predefined classification of the key indicator, we can obtain $\hat{y}_t$, which serves as an observation in our model. In addition, we introduce the confidence interval of the OLS model, denoted by ci_t, to measure the uncertainty of the predicted result. We classify the uncertainty interval into $|\Theta|$ levels, and use $\theta_t \in \Theta = \{1, 2, ..., |\Theta|\}$

to denote the uncertainty level of the predictive result in epoch t, where level 1 corresponds to the minimum uncertainty level while level $|\Theta|$ represents the maximum uncertainty level.

With all the information presented above, the observation space is then denoted by $\mathcal{O}=\{\hat{y}, \theta | \hat{y} \in \mathcal{Y}, \theta_t \in \Theta\}$. In summary, our approach captures the essential change in patients' health status by incorporating not only the changes in the frequently updated information but also the time effect of such information. Moreover, we include an interval in the observation space to capture the predictive ambiguity. All these features distinguish our model from conventional POMDP models.

Transition probability: Let $p(y_{t+1} = j | y_t = i, u_t = u)$ denote the transition from clinical class y_t to y_{t+1} with action u_t, and let $p^{'}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, u_t)$ represent the transition from state $\boldsymbol{s}_t$ to $\boldsymbol{s}_{t+1}$ with action u_t. Concretely, we have

$$p^{'}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, u_t = 0) = \begin{cases} p(y_{t+1} | y_t) & \text{if } y_t \neq A, \delta_{t+1} = \delta_t + 1 \\ 1 & \text{if } y_t = A, y_{t+1} = A, \delta_{t+1} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$p^{'}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, u_t = 1) = \begin{cases} p(y_{t+1} | y_t) & \text{if } y_t \neq A, y_t < \bar{y} \\ 1 & \text{if } (y_t = A \text{ or } y_t \geq \bar{y}) \text{ and } y_{t+1} = A, \delta_{t+1} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

We use $P_{i,j}(u) = \{p(y_{t+1} = j | y_t = i, u_t = u)\}$ to denote the transition matrix of clinical class.

Information matrix: Let $q(\hat{y}_t, \theta_t | \boldsymbol{s}_t)$ denote the probability of observing $\hat{y}_t$ and

θ_t at the beginning of epoch t given that the patient's state is $\boldsymbol{s}_t$. Concretely,

$$q(\hat{y}_t, \theta_t|\boldsymbol{s}_t) = \begin{cases} 1 & \text{if } y_t = A, \hat{y}_t = A \\ q(\hat{y}_t, \theta_t|y_t, \delta_t) & \text{if } y_t \neq A, \hat{y}_t \neq A \\ 0 & \text{otherwise.} \end{cases}$$

Given a specific $\delta = \delta^{'}$, we use $B_{i,(j,k)}(\delta^{'}) = \{q(\hat{y} = j, \theta = k|y = i, \delta = \delta^{'})\}$ to denote the information matrix.

Update matrix: Assume diagnostic tests are error-free and let $\tilde{y}_t$ denote a patient's true clinical class based on the test result. Therefore, given the clinical class is y_t, the probability of observing $\tilde{y}_t$ after performing a test is

$$l_t(\tilde{y}_t|y_t) = \begin{cases} 1 & \text{if } \tilde{y}_t = y_t \\ 0 & \text{otherwise.} \end{cases}$$

We use $\mathcal{L}_{i,j} = \{l(\tilde{y} = j|y = i)\}$ to denote the update matrix.

Cost function: Suppose a patient is in state s_t in epoch t, and an action u_t is taken. Then an immediate cost $c_t(y_t, u_t)$ will be incurred, which takes the following form if $t \neq T$:

$$c_t(y_t, u_t) = \begin{cases} d(y_t) & \text{if } u_t = 0, y_t \neq A \\ k + h(y_t) & \text{if } u_t = 1, y_t \neq A \\ 0 & \text{if } y_t = A, \end{cases}$$

and the terminal cost $c_T(y_T)$ is defined as:

$$c_T(y_T) = \begin{cases} k + h(y_T) & \text{if } y_T \neq A \\ 0 & \text{if } y_T = A. \end{cases}$$

Here, k represents the cost of receiving a diagnostic test; $d(y_t)$ denotes the cost of delaying a diagnotic test to the next epoch, and we assume $d(y_t)$ is nondecreasing in y_t, i.e., a worse clinical class is associated with a higher delay cost; $h(y_t)$ represents the intervention cost, which is nondecreasing in y_t. Since treatment is initiated only if a diagnostic test is ordered and the test result indicates $y_t \geq \bar{y}$, so $h(y_t) = 0$ if $y_t < \bar{y}$.

In the last epoch T, if the patient's state is A—which indicates that he/she is already under treatment—no additional costs will be incurred; otherwise, a diagnostic test with cost k will be ordered and the corresponding intervention cost $h(y_T)$ will be incurred.

We use $\boldsymbol{c}_t(u_t)$ to denote the cost vector over the belief state $\boldsymbol{\pi}_t \in \Pi$ in epoch t if $t < T$, with the *i-th* element equals to $c_t(y_t = i, u_t = u)$, and $\boldsymbol{c}_T$ to denote the terminal cost vector with the *i-th* element equals to $c_T(y_T = i)$.

Belief update:

If $u_t = 0$, in epoch $t+1$, given the new observation $\hat{y}_{t+1}$, θ_{t+1} and time lag δ_{t+1}, the Bayesian update of the belief state $\boldsymbol{\pi}_t$ is computed as follows:

$$\boldsymbol{\pi}_{t+1} = T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}) = \frac{\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})P^{'}(u_t)\boldsymbol{\pi_t}}{f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)},$$

where $\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1}) = diag\big(B_{1,(\hat{y}_{t+1},\theta_{t+1})}(\delta_{t+1}), ..., B_{Y,(\hat{y}_{t+1},\theta_{t+1})}(\delta_{t+1}), B_{A,(\hat{y}_{t+1},\theta_{t+1})}(\delta_{t+1})\big)$ and

$$\begin{aligned} f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t) &= \mathbf{1}^{'}_{Y+1}\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})P^{'}(u_t)\boldsymbol{\pi}_t \\ &= \sum_{y_{t+1}\in\mathcal{Y}} q(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1}, \delta_{t+1}) \sum_{y_t\in\mathcal{Y}} p(y_{t+1}|y_t, u_t)\pi_t(y_t), \end{aligned}$$

which is the conditional probability of observing $\hat{y}_{t+1}$ and θ_{t+1} given $\boldsymbol{\pi}_t, \delta_{t+1}$ and u_t. Note that the *i-th* element of $\boldsymbol{\pi}_{t+1}$ equals to

$$\frac{q(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1} = i, \delta_{t+1}) \sum_{y_t\in\mathcal{Y}} p(y_{t+1} = i|y_t, u_t = 0)\pi_t(y_t)}{f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}.$$

If $u_t = 1$, which indicates a diagnostic test is ordered in epoch t, we can observe the patient's true clinical class $\tilde{y}_t$ and update the belief state to $\tilde{\boldsymbol{\pi}}_t$,

$$\tilde{\boldsymbol{\pi}}_t = \tilde{T}(\boldsymbol{\pi}_t, \tilde{y}_t) = \frac{\tilde{L}_{\tilde{y}_t}\boldsymbol{\pi}_t}{\tilde{f}_t(\tilde{y}_t|\boldsymbol{\pi}_t)},$$

where $\tilde{L}_{\tilde{y}_t} = diag(L_{1,\tilde{y}_t}, ..., L_{Y,\tilde{y}_t}, L_{A,\tilde{y}_t})$ and $\tilde{f}_t(\tilde{y}_t|\boldsymbol{\pi}_t) = \mathbf{1}^{'}_{Y+1}\tilde{L}_{\tilde{y}_t}\boldsymbol{\pi}_t = \sum_{y_t\in\mathcal{Y}} l_t(\tilde{y}_t|y_t)\pi_t(y_t) = \pi_t(\tilde{y}_t)$, which is the conditional probability of observing $\tilde{y}_t$ if a diagnostic test is ordered and finds the patient clinical class is $\tilde{y}_t$. Note that $\tilde{\boldsymbol{\pi}}_t$ is a unit vector with $\tilde{\pi}_t(\tilde{y}_t)$ equals 1.

Then, in epoch $t+1$, the Bayesian update of the belief state is computed by

$$\boldsymbol{\pi}_{t+1} = T(\tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t = 1, \hat{y}_{t+1}, \theta_{t+1}).$$

Discount factor: We use $\rho \in (0, 1]$ to denote the discount factor.

Optimality Equations: If a patient is under treatment (i.e., $\pi(A) = 1$), no extra costs will be incurred; otherwise, let $Q_t(\boldsymbol{\pi}_t, \delta_t, u_t)$ denote the total expected cost and $v_t(\boldsymbol{\pi}_t, \delta_t)$ represent the minimum total expected cost of a patient. Then

$$\begin{aligned}Q_t(\boldsymbol{\pi}_t, \delta_t, u_t = 0) &= \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1}\in\mathcal{O}} \sum_{\theta_{t+1}\in\Theta} v_{t+1}(\boldsymbol{\pi}_{t+1}, \delta_{t+1}) \cdot f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)\\ &= \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1}\in\mathcal{O}} \sum_{\theta_{t+1}\in\Theta} v_{t+1}(T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}), \delta_t + 1)\\ &\cdot f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t),\end{aligned}$$

$$\begin{aligned}Q_t(\boldsymbol{\pi}_t, \delta_t, u_t = 1) =&\boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\tilde{y}_t\in\mathcal{O}} \tilde{f}_t(\tilde{y}_t|\boldsymbol{\pi}_t) \sum_{\hat{y}_{t+1}\in\mathcal{O}} \sum_{\theta_{t+1}\in\Theta} v_{t+1}(T(\tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}), \delta_{t+1} = 1)\cdot\\ &f_t(\hat{y}_{t+1}, \theta_{t+1}|\tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t)\end{aligned}$$

and

$$v_t(\boldsymbol{\pi}_t, \delta_t) = \min_{u_t\in\mathcal{U}_t} Q_t(\boldsymbol{\pi}_t, \delta_t, u_t).$$

To summarize, for $t < T$,

$$v_t(\boldsymbol{\pi}_t, \delta_t) = \begin{cases} 0 & \text{if } \pi_t(A) = 1\\ \min\limits_{u_t\in\mathcal{U}_t} Q_t(\boldsymbol{\pi}_t, \delta_t, u_t) & \text{if } \pi_t(A) \neq 1, \delta_t < |\Delta|\\ Q_t(\boldsymbol{\pi}_t, \delta_t, u_t = 0) & \text{otherwise.}\end{cases}$$

The terminal cost $v(\boldsymbol{\pi}_T)$ is represented by

$$v(\pi_T) = \boldsymbol{c}_T^{'}\boldsymbol{\pi}_T.$$

3.4 Equivalent Optimality Equations and Modified Algorithm

In this section, we first present an alternative equivalent representation for the optimality equations in Section 3.4.1, and discuss a modified incremental pruning algorithm in Section 3.4.2.

3.4.1 An Alternative Representation for the Optimality Equations

POMDP is a continuous state space Markov decision process (MDP) due to infinitely many states, and classical backward induction in MDP cannot be directly applied to solve the problem. Pioneering work by Smallwood and Sondik (1973) show that the value function of a POMDP is piecewise linear and concave (*pwlc*) with respect to the belief space.

However, our proposed POMDP model differs from the conventional POMDP models in two aspects. First, if a diagnostic test is performed, we need to "renew" our belief vector based on the test result. Concretely, if a diagnostic test is conducted in epoch t and the test result suggests a clinical class $\tilde{y}_t$, belief $\boldsymbol{\pi}_t$ should be updated by $\tilde{\boldsymbol{\pi}}_t = \frac{\tilde{L}_{\tilde{y}_t}\boldsymbol{\pi}_t}{\tilde{f}_t(\tilde{y}_t|\boldsymbol{\pi}_t)}$; and in epoch $t+1$, $\boldsymbol{\pi}_{t+1}$ will be updated based on $\tilde{\boldsymbol{\pi}}_t$ rather than $\boldsymbol{\pi}_t$. Moreover, our model considers a maximum allowed time lag $|\Delta|$ such that patients will receive a mandatory test when $\delta = |\Delta|$. These new features make the preservation of *pwlc* property under our model remains unclear. In the following theorem, we prove that the optimal value function in our model still preserves the *pwlc* property.

Theorem 3.1. *Given a specific δ_t, $v_t(\boldsymbol{\pi}_t, \delta_t)$ is pwlc with respect to $\boldsymbol{\pi}_t \in \Pi$; that is,*

$$v_t(\boldsymbol{\pi}_t, \delta_t) = \min_{\boldsymbol{\gamma} \in \Gamma_t(\delta_t)} \boldsymbol{\gamma}' \boldsymbol{\pi}_t,$$

where $\Gamma_t(\delta_t)$ is a finite set of $(Y+1)$-dimensional vectors.

We next explore the benefits of including uncertainty intervals in the observation space. For ease of expression, we refer to the POMDP model that includes (does not include) uncertainty intervals in the observation space as POMDP-UI (POMDP-B). We present the structure of POMDP-B in Appendix B.1.

Theorem 3.2. *The minimum total expected cost of POMDP-UI is less than or equal to that of POMDP-B.*

Theorem 3.2 demonstrates the superiority of including the uncertainty interval. Relative to POMDP-B, POMDP-UI combines predictions and uncertainty intervals that capture the uncertainty in patient condition, which provides another dimension of information for real-time patient risk monitoring and gives rise to an equal or lower total expected costs for patients.

3.4.2 Algorithm

In this part, we propose an algorithm to construct $\Gamma_t(\delta_t)$ in the *pwlc* function. When $u_t = 0$ and $\pi_t(A) = 0$, $\delta_{t+1} = \delta_t + 1$, similar to Krishnamurthy (2016).The set $\Gamma_t(\delta, u = 0)$ can be constructed as follows:

$$\Gamma_t(\delta_t, u_t, \hat{y}_{t+1}, \theta_{t+1}) = \{\frac{c_t(u_t)}{(Y+1)|\Theta|} + \rho P(u_t)\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})\boldsymbol{\gamma}_{t+1}|\boldsymbol{\gamma} \in \Gamma_{t+1}(\delta_{t+1})\}$$

$$\Gamma_t(\delta_t, u_t) = \bigoplus\nolimits_{\hat{y}_{t+1},\theta_{t+1}}\Gamma_t(\delta_t, u_t, \hat{y}_{t+1}, \theta_{t+1}),$$

where $\bigoplus$ denotes the cross-sum operator, and $A\bigoplus B$ consists of all pairwise additions of vectors from these two sets.

When $u_t = 1$, $\delta_{t+1} = 1$, the set $\Gamma_t(\delta_t, u_t = 1)$ can be constructed as follows:

$$\Gamma_t(\delta_t, u_t, \tilde{y}_t, \hat{y}_{t+1}, \theta_{t+1}) = \{\frac{c_t(u_t)}{(Y+1)^2|\Theta|} + \rho\tilde{L}_{\tilde{y}_t}P(u_t)\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})\boldsymbol{\gamma}_{t+1}|\boldsymbol{\gamma}_{t+1} \in \Gamma_{t+1}(\delta_{t+1})\}$$

$$\Gamma_t(\delta_t, u_t, \tilde{y}_t) = \bigoplus\nolimits_{\hat{y}_{t+1},\theta_{t+1}}\Gamma_t(\delta_t, u_t, \tilde{y}_t, \hat{y}_{t+1}, \theta_{t+1})$$

$$\Gamma_t(\delta_t, u_t) = \bigoplus\nolimits_{\tilde{y}_t}\Gamma_t(\delta_t, u_t, \tilde{y}_t).$$

Then $\Gamma_t(\delta_t)$ can be constructed as $\Gamma_t(\delta_t) = \cup_{u_t \in \mathcal{U}_t}\Gamma_t(\delta_t, u_t)$.

We further present a modified incremental pruning algorithm (MIPA) to compute

Algorithm 1 MIPA

```
**Input:** $\Gamma_{t+1}(\delta_{t+1}), \delta_{t+1} \in \Delta_{t+1}$
**Output:** $\Gamma_t(\delta_t), \delta_t \in \Delta_t$

for each $\delta_t \in \Delta_t$ do
    $u_t = 0$
    for each $\hat{y}_{t+1} \in \mathcal{O}, \theta_{t+1} \in \Theta$ do
        $\Gamma_t(\delta_t, u_t = 0, \hat{y}_{t+1}, \theta_{t+1}) \longleftarrow \text{prune}(\{\frac{c_t(u_t)}{(Y+1)|\Theta|} + \rho P(u_t)\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})\gamma_{\mathbf{t+1}}|\gamma_{\mathbf{t+1}} \in \mathbf{\Gamma_{t+1}}(\delta_{\mathbf{t+1}})\})$
        $\Gamma_t(\delta_t, u_t = 0) \longleftarrow \text{prune}(\Gamma_t(\delta_t, u_t = 0)\bigoplus\Gamma_t(\delta_t, u_t = 0, \hat{y}_{t+1}, \theta_{t+1}))$
    end for
    $\Gamma_t(\delta_t) = \text{prune}(\Gamma_t(\delta_t) \cup \Gamma_t(\delta_t, u_t = 0))$
    $u_t = 1$
    for each $\tilde{y}_t \in \mathcal{O}$ do
        for each $\hat{y}_{t+1} \in \mathcal{O}, \theta_{t+1} \in \Theta$ do
            $\Gamma_t(\delta_t, u_t = 1, \tilde{y}_t, \hat{y}_{t+1}, \theta_{t+1}) \longleftarrow \text{prune}(\{\frac{c_t(u_t)}{(Y+1)^2|\Theta|} + \rho\tilde{L}_{\tilde{y}_t}P(u_t)\tilde{B}_{\hat{y}_{t+1},\theta_{t+1}}(\delta_{t+1})\gamma_{\mathbf{t+1}}|\gamma_{\mathbf{t+1}} \in \mathbf{\Gamma_{t+1}}(\delta_{\mathbf{t+1}})\})$
            $\Gamma_t(\delta_t, u_t = 1, \tilde{y}_t) \longleftarrow \text{prune}(\Gamma_t(\delta_t, u_t = 1, \tilde{y}_t)\bigoplus\Gamma_t(\delta_t, u_t = 1, \tilde{y}_t, \hat{y}_{t+1}, \theta_{t+1}))$
        end for
        $\Gamma_t(\delta_t, u_t = 1) \longleftarrow \text{prune}(\Gamma_t(\delta_t, u_t = 1)\bigoplus\Gamma_t(\delta_t, u_t = 1, \tilde{y}_t))$
    end for
    $\Gamma_t(\delta_t) = \text{prune}(\Gamma_t(\delta_t) \cup \Gamma_t(\delta_t, u_t = 1))$
end for
```

$\Gamma_t(\delta_t)$ in Algorithm 1, which is based on the incremental pruning algorithm introduced in Cassandra et al. (1997). Note that δ_t is the time elapsed from the last test period, so the upper bound of δ_t (denoted as Δ_t) in epoch t should be $\min\{t, |\Delta|\}$. Concretely, $\Delta_t = \{1, 2, ..., t\}$ if $t < |\Delta|$, otherwise, $\Delta_t = \Delta$.

As for the "prune" function, since our optimal value function preserves the *pwlc* property, suppose there is a $\boldsymbol{\gamma} \in \Gamma_k$ such that $\forall \boldsymbol{\pi} \in \Pi$, $\gamma' \pi \geq \tilde{\gamma}' \pi$ for all vectors $\tilde{\gamma} \in \mathbf{\Gamma_t} - \{\gamma\}$. Then γ dominate all the other vectors in the belief space, and the prune function would eliminate such vectors. Specifically, we introduce the following linear programming, that if α yields a solution $\alpha < 0$, then γ dominate other vectors and can be eliminated.

$$\begin{aligned} \text{maximize} \quad & \alpha \\ \text{subject to} \quad & (\tilde{\gamma} - \gamma)^{'} \pi \geq \alpha && \forall \tilde{\gamma} \in \Gamma_t - \{\gamma\} \\ & \pi(i) \geq 0 && i \in \mathcal{Y} \\ & \mathbf{1}^{'}_{\mathbf{Y+1}} \pi = \mathbf{1} \end{aligned}$$

3.5 Numerical Experiments

In this section, we conduct a set of numerical experiments using synthetic examples to evaluate the performance of the proposed model (POMDP-UI). We describe the setting and parameters in Section 3.5.1, and illustrate the main results in Section 3.5.2.

3.5.1 Experimental Settings and Parameters

Consider the status of a patient across a planning horizon of 48 hours, then $T = 48$ if we use 1 hour as the decision epoch. Let r_t denote the value of the key indicator in epoch t. We adopt a random walk to mimic patients' progression. Specifically, we assume $r_t = r_{t-1} + w_t$, where w_t is a Gaussian white noise process with mean 0 and variance 1 (case with a high disease incidence rate; denoted by H-case) or variance 0.5 (case with a low incidence prevalence rate; denoted by L-case). We simulate 3,000 patients, for each of whom the starting r_0 is assumed to be uniformly distributed on the interval $[0, 10)$. We classify patients into three clinical classes. Patients with $r < 10$ are classified into normal class, i.e., $y = 0$; the abnormal class $y = 1$ contains patients with $r \geq 10$; the absorbing class includes patients under treatment. Specifically, we assume that a patient will receive treatment if and only a diagnostic test is performed, and the result suggests $r \geq 10$. $|\Delta|$ is set to 8, i.e., individuals have to receive at least one diagnostic test in 8 epochs. Then the system state can be defined accordingly based on Section 3.3. The transition matrix between three clinical classes can be estimated from the simulated data (cf. Appendix

B.3.1 for the details), and the transition probabilities among different states can be specified according to Section 3.3. We randomly generate the information matrices for both POMDP-UI and POMDP-B (cf. Appendix B.3.1 for the details). We assume delay cost $d(y_t) = 0$ for patients with $y_t = 0$ and $d(y_t) = 10$ for patients with $y_t = 1$; intervention cost $h(y_t) = 8$ for patients with $y_t = 1$; and terminal cost $c_T(\boldsymbol{\pi}_T) = [0, 30, 0]$. Test cost k is from the set $\{0, 0.2, 0.5, 1, 1.5, 2, 3, 5\}$. Discount factor ρ is assumed to be 1.

For ease of expression, we refer to the optimal policies derived from POMDP-UI and POMDP-B as OP-POMDP-UI and OP-POMDP-B, respectively. We evaluate the performance of the proposed model by comparing OP-POMDP-UI to four benchmark testing policies with different fixed testing intervals—i.e., conducting a diagnostic test every 2 hours, 3 hours, 4 hours, and 6 hours as policy R2, policy R3, policy R4, and policy R6, respectively. We further compare OP-POMDP-UI with OP-POMDP-B to examine whether there are any additional benefits to incorporating the uncertainty interval into the modeling framework.

3.5.2 Numerical Results

We compare OP-POMDP-UI to the benchmark policies and OP-POMDP-B for H-case in terms of multiple metrics: *Number of Tests*, *Detection Time*, and *Missed Periods*. *Number of Tests* is the total number of diagnostic tests performed, *Detection Time* denotes the epoch in which the disease is detected, and *Missed Periods* represents the total number of epochs during which patients are in the abnormal state until the disease is detected. Specifically, to evaluate each policy, we count the number of tests performed for all patients and use the average value as a proxy for *Number of Tests*. We next collect the detection time of the patients detected by the policy and adopt the average value as a measure for *Detection Time*. Then, we sum up the missed periods for all individuals, including those not detected by the policy, and use the average value as a measure for *Missed Periods*.

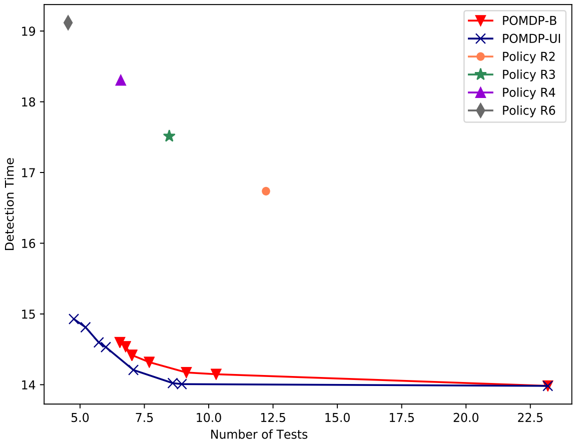


Figure 3.2: Trade-off between *Number of Tests* and *Detection Time* in numerical experiments (H-case)

Figure 3.2 illustrates the trade-off between *Number of Tests* and *Detection Time*. Note that the *Number of Tests* for policy R2, R3, R4, and R6 are 12.2, 8.5, 6.6, and 4.5 times, respectively; and the *Detection Time* for policy R2, R3, R4, and R6 are 16.7, 17.5, 18.3 and 19.1, respectively. OP-POMDP-UI is able to achieve earlier detection with fewer tests compared with Policy R2, R3, and R4. For example, OP-POMDP-UI can detect patients' deterioration at time 14.9 with 4.8 tests, which reduces *Number of Tests* by 60.7%, 43.5%, and 27.3%, and advances the *Detection Time* by 10.8%, 14.9%, and 18.6% compared with Policy R2, R3, and R4. In comparison to policy R6, OP-POMDP-UI can advance the detection time by 22% with little sacrifice in *Number of Tests*. In addition, we find that the performance of OP-POMDP-UI is superior to OP-POMDP-B with respect to *Detection Time* and *Number of Tests*, the observation further addresses the benefit of considering the uncertainty of the estimation.

The trade-off between *Number of Tests* and *Missed Periods* is shown in Figure 3.3. We find that *Missed Periods* for policy R6, R4, R3, R2, OP-POMDP-U and the OP-POMDP-UI is in a descending order, which is consistent with the early detection capability discussed above. Notably, *Missed Periods* is 0.5 for policy R2 that conducts a diagnostic test every 2 hours; OP-POMDP-UI can reduce *Missed Periods* to 0.32 with 5.2 tests. The improvement is even greater compared with policy R3, R4, and R6, with *Missed Periods* of 1.0, 1.4, and 2.2, respectively. In terms of percentages, OP-POMDP-UI can reduce *Missed Periods* by 36.0%, 68.0% and 77.1%, and reduce *Number of Tests* by 57%, 38.9%, and 21.2% compared to R2, R3 and R4. Furthermore, it can reduce *Missed Periods* by 85%, with a 16% increase in *Number of Tests* compared to R6. Besides, we observe that the performance of OP-POMDP-UI is superior to OP-POMDP-B with respect to *Missed Periods* and *Number of Tests*.

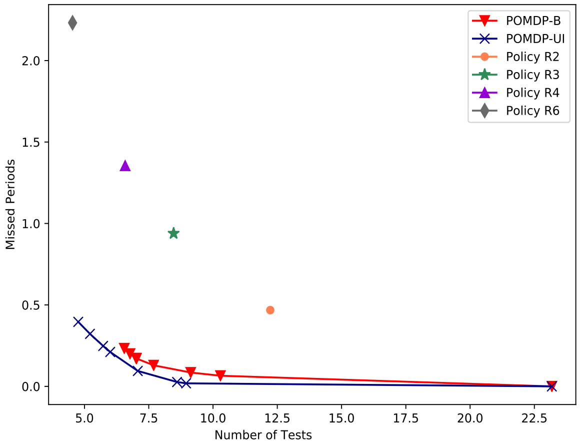


Figure 3.3: Trade-off between *Number of Tests* and *Missed Periods* in numerical experiments (H-case)

In summary, OP-POMDP-UI greatly outperforms four benchmark policies in terms of *Detection Time* and *Missed Periods*. Furthermore, it enables earlier detection with fewer tests compared to R2, R3, R4 and OP-POMDP-B. The results for L-case are similar to the results we discussed above. We relegate all the relative results in Appendix B.3.2.

3.6 Case Study

To further evaluate the performance of the proposed model in practical environments, we conduct a case study in this section. We apply our model to a cardiothoracic ICU of a national hospital to optimize the prescription of diagnostic tests in blood glucose monitoring. We introduce the background of blood glucose monitoring in Section 3.6.1, followed by the data selection process and summary statistics in Section 3.6.2. We calibrate the model settings and parameters based on the real data in Section 3.6.3, and present the numerical results in Section 3.6.4.

3.6.1 Background

Broad observational studies (e.g., Barsheshet et al., 2006, Capes et al., 2000) have shown that hyperglycemia is associated with higher morbidity and mortality in hospitalized patients, and it's essential to provide prompt treatment for these patients. Langley and Adams (2007) conduct a systematic review and conclude that maintaining normoglycaemia and treatment with insulin-based regimens is beneficial for limiting organ damage and significantly reduces both morbidity and mortality in critically ill patients who require intensive treatments. In recent decades, point of care testing (POCT) for blood glucose levels has developed steadily, and glucose meters are widely used in ICUs for POCT (Tonyushkina and Nichols, 2009). In contrast to other diagnostic tests, which require that the specimen be sent from the point of care, followed by waiting for results, POCT yields prompt test results for patients' glucose levels. However, frequent testing with a glucose meter requires finger pricking and increases patients' anxiety and discomfort (Ong et al., 2014), while delays in testing may lead to delayed treatment and adverse health outcomes. Thus, in this part, we aim to investigate the optimal time to administer a blood glucose test in ICUs.

Our partner hospital recommends testing blood glucose levels every 6—8 hours for patients not receiving insulin treatments. However, in the data, the average testing interval between two consecutive glucose tests is 4.5 hours, which indicates more frequent orderings than the guideline. We apply the proposed model developed in Section 3.3 to address the diagnostic overutilization problem in blood glucose monitoring.

3.6.2 Data Selection and Summary Statistics

Of all 5,351 ICU patients in our dataset, we first eliminate patients who didn't have glucose testing records and who had diabetes or hypoglycemia prior to their admission ($N = 477$). Moreover, we do not consider patients who had abnormal blood glucose levels at the first blood glucose test or who were on insulin therapy before their first tests. ($N = 551$). Note that the average LOS for all patients is 3.9 days. We further remove patients with ICU LOS longer than 10 days ($N = 265$), as they may have different or complicated conditions. The data points recorded at mealtimes are also excluded. We provide summary statistics for the study sample in terms of demographics and clinical variables in Table 3.1.

3.6.3 Model Setting and Parameter

We next calibrate model settings and discuss the details of parameter estimation.

Time horizon: For each patient, the time horizon is from the first testing time to the discharge time, so the time horizon differs from patient to patient. The decision epoch is 1 hour for all patients. We use $\mathcal{T}_i = \{1, 2, 3, ..., T_i\}$ to denote the time horizon for patient i.

State space: According to the protocol in our partner hospital, patients with glucose levels no less than 10 $mmol/L$ should be intervened with insulin treatment. The cut-off point of 10 $mmol/L$ is also in line with the medical literature (Ichai and Preiser, 2010). Thus, we classify patients into three clinical classes: normal (glucose level is less than 10 $mmol/L$), abnormal (glucose level is no less than 10 $mmol/L$) and absorbing. We suppose $|\Delta| = 12$, i.e., patients must receive at least one diagnostic

Variable	Mean	SD	Variable	Count	Percentage
Demographics					
Age, year	59.9	13.4	Gender: male	3,225	75.5
Weight, kg	65.5	14.7	Race		
			Chinese	2,840	66.5
			Malay	618	14.6
			Indian	344	8.1
			Other	470	10.8
	Clinical Variables	Mean	SD		
	HR, beats/min	83.8	15.9		
	RR, times/min	19.1	6.0		
	Temperature, °C	36.5	0.9		
	SpO2, %	98.6	3.0		
	Mean arterial BP, mmHg	78.6	13.9		
	Systolic BP, mmHg	117.8	20		
	Diastolic BP, mmHg	59.8	13		
	CVP, mmHg	9.0	5.7		
	GCS	12.8	3.8		
	Arterial pCO2, mmHg	40.3	7.4		
	Arterial pH	7.4	0.1		
	Arterial SaO2, %	93.3	12.3		
	FiO2, %	39.9	9.0		
	Urine volume, mL	42.8	82.7		
	Glucose, mmol/L	7.9	2.2		

SD: standard deviation; HR: heart rate; RR: respiratory rate; GCS: Glasgow Coma Scale; BP: blood pressure; CVP: central venous pressure.

Table 3.1: Summary statistics in blood glucose monitoring case

test every healf day. The system state can be defined accordingly.

Transition probability: Since the decision period is 1 hour, we extract all glucose records for each patient and identify all consecutive tests that were performed within 1 hour. Based on the selected data, we use maximum likelihood estimation to derive the transition probabilities between patients' clinical classes $p(y_{t+1} = j|y_t = i, u)$, which is shown below.

$$P(u=0)=\begin{pmatrix} 0.92 & 0.08 & 0 \\ 0.20 & 0.80 & 0 \\ 0 & 0 & 1 \end{pmatrix}, P(u=1)=\begin{pmatrix} 0.92 & 0.08 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Hence, $p^{'}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, u)$ can be easily generated according to the definition in Section 3.3.

Observation space: We adopt an OLS model introduced in Section 3.3 to construct the observations. To select variables, we first perform 5-fold cross-validation and select the combination of variables that yields the lowest average mean squared error. Then, we incorporate all selected variables into the OLS model, and the final model is presented in Appendix B.4. Based on the final model, we can predict the individuals' indicators and calculate confidence intervals for the predicted outcomes $\hat{r}$. If $\hat{r} < 10$, the predicted outcome will be classified as normal; if $\hat{r} \geq 10$, it will be classified as abnormal. we further divide confidence intervals into two levels based on the criteria of whether the confidence interval covers both normal and abnormal classes, i.e., if the lower bound of the confidence interval is categorized as normal and the upper bound is categorized as abnormal, then the uncertainty level is "high"; otherwise, the uncertainty level is "low". In addition, patients' absorbing state can be observed directly.

Information matrix: Based on individuals' true clinical classes, time lags, the predicted clinical classes and uncertainty levels, we construct the information matrices using maximum likelihood estimation (cf. Appendix B.4 for the details).

Cost: Due to the lack of relevant data, we are unable to estimate the associated costs. We assume a delay cost of 0 for patients with blood glucose level $< 10mmol/L$ and 10 for patients with blood glucose level $\geq 10mmol/L$; intervention cost of 8 for the patients with abnormal classes and terminal cost of $[0, 30, 0]$. Test cost k is from the set $\{0, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 1.2, 1.5, 2, 3, 5\}$.

Remark 1: At some epochs, our optimal policy may suggest a test action, while in reality, blood glucose may not be tested in those epochs. Therefore, we cannot determine whether treatments should be initiated using incomplete real data. To address this problem, we sample 100 groups of blood glucose levels from a normal distribution, with the mean being the predicted blood glucose level and the standard error being the standard error of prediction.

3.6.4 Numerical Results

We compare OP-POMDP-UI with OP-POMDP-B and the practical policy adopted in our partner ICU in terms of *Number of Tests*, *Detection Time* and *Missed Periods*.

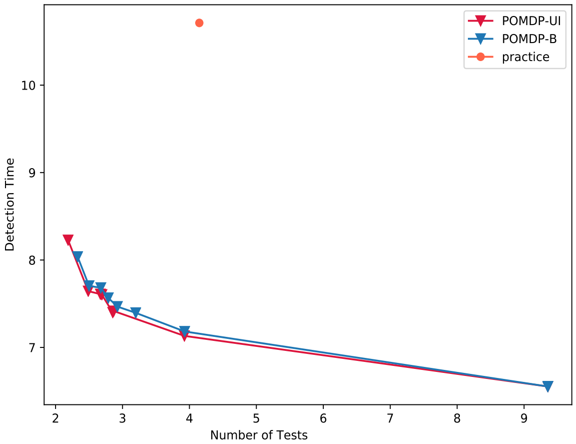


Figure 3.4: Trade-off between *Number of Tests* and *Detection Time* in glucose monitoring

Figure 3.4 illustrates the trade-off between *Number of Tests* and *Detection Time*. The *Number of Tests* and the *Detection Time* for the practical policy is 4.1 and 10.7, respectively. OP-POMDP-UI is able to detect patients' deterioration at time 7.4 with 2.9 tests, which greatly reduces *Number of Tests* by 29.7% and advances the *Detection Time* by 30.8% relative to the practical policy. It is worth noting that the mean onset of hyperglycemia is 6.6, and OP-PODMP-UI detects the symptoms with little delay. Compared to OP-POMDP-B, OP-POMDP-UI is able to achieve slightly earlier detection with fewer tests. The improvement is subtle, suggesting that the inclusion of predictive information is capable of bringing significant benefits in this case.

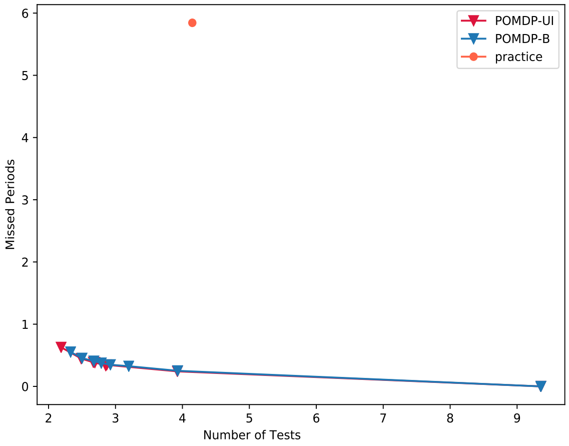


Figure 3.5: Trade-off between *Number of Tests* and *Missed Periods* in glucose monitoring

The trade-off between *Number of Tests* and *Missed Periods* is presented in Figure 3.5. The *Missed Periods* is 5.8 for the practical policy; OP-POMDP-UI can significantly reduce *Missed Periods* to 0.4 with 2.5 tests. Besides, we observe that OP-POMDP-UI outperforms OP-POMDP-B with respect to *Missed Periods* and *Number of Tests*.

In summary, the proposed policy can achieve earlier detection with fewer tests. Furthermore, the proposed policy can also significantly reduce the delay time in detection with fewer tests.

3.7 Conclusion

The overuse of diagnostic tests is an emerging and critical problem in the development of the healthcare ecosystem. However, the existing researches have not explicitly addressed the overutilization problem of repetitive tests in acute diseases contexts. On the one hand, physicians need frequently administer diagnostic tests to track patients' health conditions as delay in tests may lead to delay in detection and treatment, which violates people-centric philosophy in healthcare service delivery; on the other hand, overuse of diagnostic tests may result in financial burden and negative health effect on the patient, moreover, it also incurs huge medical resources waste. In this chapter, we propose a POMDP framework to optimize the prescription of diagnostic tests under real-time monitoring contexts where people undergo frequent routine, unharmful, and cheap medical tests. Based on the high-frequency updated clinical information, we develop a predictive model, whose accuracy depends on the time lag of diagnostic tests, to predict the patient's clinical results. We further embed the predictions in the POMDP framework for information update and decision support. We propose a modified incremental pruning algorithm to derive the optimal policy and evaluate the performance of the proposed model for glucose monitoring in ICUs. Our results show that the proposed framework can achieve early detection of deterioration with fewer diagnostic tests compared with the practical policy. The framework can also be applied to many other health monitoring systems, especially those powered by smart wearable health devices for acute and chronic diseases.

This chapter has several limitations. First, we have not theoretically explored the specific value of imperfect predictive information and uncertainty intervals. Second, we assume that the diagnostic tests for key indicators are perfect. However, in reality, these diagnostic tests are also subject to errors. We believe it will not influence our main insights, and we can incorporate it in our model if the relevant data is available. Besides, in our numerical experiments and case studies, as the real cost data is not available, we use hypothetical cost data to derive the optimal policy and evaluate the performance of the proposed model. We leave these problems for future research.

Chapter 4

Optimizing Initial Screening for Colorectal Cancer Detection with Adherence Behavior

4.1 Introduction

Cancer is the second leading cause of death globally. In 2021, more than 1.9 million new cancer diagnoses are expected in the United States (Siegel et al., 2021). Despite that cancer is a fatal disease, cancer-related mortality has been declining steadily. The five-year relative survival rate for all cancers has increased from 49% in 1975-1977 to 69% in 2009-2015 (Siegel et al., 2020). Much of this reduction in mortality can be attributed to the efforts in detecting cancers early and advances in cancer treatment (Berry et al., 2005, Office for National Statistics, 2019). Take colorectal cancer (CRC) as an example; the 5-year relative survival rates for CRC are 90%, 71%, and 14% at the local, regional, and distant stages, respectively (American Cancer Society 2020a). Starting the treatment before it progresses to advanced stages would significantly reduce the number of death. However, early detection of cancer remains a challenge mostly due to the asymptomatic nature of certain cancers at an early stage. Therefore, regular public cancer screening becomes essential.

According to the World Health Organization, the aim of a cancer screening program is to “identify individuals with abnormalities suggestive of specific cancer or pre-cancer who have not developed any symptoms and refer them promptly for diagnosis and treatment". The targeted population is average-risk individuals without medical history and family history of cancer. A well-designed screening guideline would be able to effectively detect individuals at risk with the consideration of the resulting economic costs of the screening protocol. In practice, many health systems introduce cancer screening protocols in the form of two-stage medical tests. Examples include screenings for CRC, breast cancer, and prostate cancer. Our study specifically focuses on the two-stage CRC screening program.

CRC is the third most diagnosed cancer and the second leading cause of death for cancer patients worldwide (Sung et al., 2021). It is a cancer developed from abnormal cells in the colon or rectum. There are three CRC stages, namely local, regional and distant stages (American Cancer Society, 2020a). Approximately 4.4% of men and 4.1% of women will be diagnosed with CRC in their lifetime (Siegel et al., 2020). Most of the countries and regions adopt two-stage CRC screening programs for early detection. Individuals above a certain age (e.g., 45 in the US; 50 in Singapore) are advised to do the fecal immunochemical test (FIT) annually, followed by a colonoscopy if any positive outcome is received from the FIT (American Cancer Society, 2020b, Singapore Health Hub, 2011). Typically, the first stage of the screening (i.e., an initial screening/test) uses a less invasive, less costly, but less accurate test, whereas the second-stage tests are more invasive, more expensive, but usually considered gold-standard with almost 100% accuracy. Having an initial screening before the gold-standard test aims to make the screening guideline more accessible to the public, inducing a higher acceptance level and take-up rate. Additionally, second-stage tests would be performed on a smaller group of individuals who have already received risk alarms from the initial test and, therefore, gives rise to a considerate economic saving.

Given that a screening program is a complex public health intervention, the design of the initial tests would significantly affect the performance of two-stage screening programs. Most initial tests compare the biomarker concentrations or estimated risk levels to a pre-specified cut-off point to report a positive or negative result. FIT screens for occult (hidden) blood in the stool sample. The fecal-hemoglobin (f-Hb) concentration of the sample is compared with a cut-off value (e.g., 20 $\mu g/g$ for the test kit used in Singapore) so that a concentration value higher (lower) than this cut-off point will trigger a positive (negative) outcome. The choice of cut-off points can result in different test sensitivity and specificity. The sensitivity of the test kit is defined as the probability that an individual with CRC receives a positive outcome; specificity is the probability that a healthy individual receives negative outcome. Given the hemoglobin concentration distributions for healthy individuals and individuals with CRC (Figure 4.1), a higher cut-off point will decrease test sensitivity and increase specificity.

There are two forms of FIT kits adopted in practice, *qualitative* FIT and *quantitative* FIT. Qualitative FIT uses fixed pre-specified threshold values chosen by the manufacturers. When it is used by the health system, only binary outcomes (positive or negative) are reported. Quantitative FIT, on the other hand, directly reports the values of f-Hb concentrations (Fraser, 2011). Heath systems can adjust the test cut-off points and control the test sensitivity and specificity. It is heatedly debated in the medical community on the different impacts of these two types of FIT kits (Fraser et al., 2012, Allison et al., 2014b) because the selection of the initial test cut-off point would directly impact both the effectiveness and efficiency of the two-stage screening program. Tests with higher sensitivity can detect more cancer incidences and thus contribute to the screening effectiveness; however, higher sensitivity is usually accompanied by lower specificity (a higher false-positive rate), leading to unnecessary second-stage tests and a higher economic burden. In fact, colonoscopy is an expensive procedure, most countries do not have sufficient resources to screen the entire target population, and in some regions, patients may suffer from long waiting times to get tested (Kolata, 2003, Yu et al., 2008, Hubers et al., 2020). On the other hand, colonoscopy resource is overused within some regions because of a high rate of inappropriate recommendations for patients with low risk (Zorzi et al., 2016, Kruse et al., 2015, Murphy et al., 2016). Hence, many research works advocate the benefits of quantitative tests so that health systems have higher flexibility in designing their own FIT according to different practical considerations (Fraser et al., 2012, Robertson et al., 2017). According to the 5^{th} meeting of the World Endoscopy Organization CRC screening committee, the primary concern of the selection of the cut-off point in the quantitative FIT test is to balance the trade-off between test effectiveness and test efficiency (Allison et al., 2014a). For example, Spain used 20 $\mu g/g$ as the cut-off point to encourage more follow-ups from CRC patients but was experiencing pressure on colonoscopy resource; the Netherlands has increased the cut-off point of the FIT they used for CRC screening from 15 $\mu g/g$ to 47 $\mu g/g$ due to insufficient colonoscopies (Allison et al., 2014a, Toes-Zoutendijk et al., 2020).

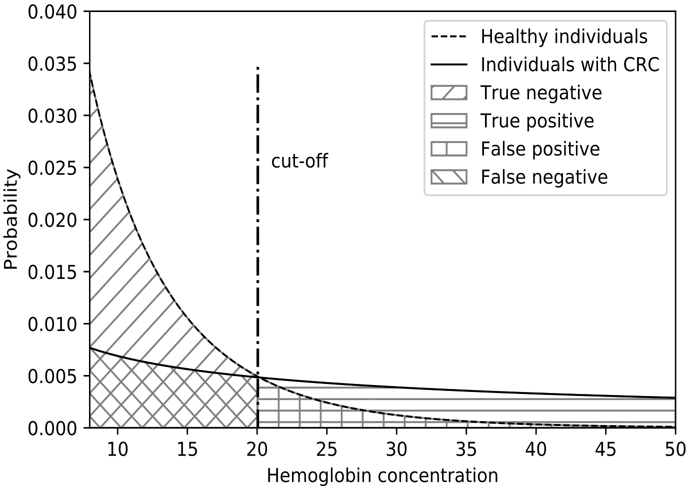


Figure 4.1: Probability density function

In addition, the actual demand for the second-stage test also depends on individuals' adherence to the protocol. The screening program advises all individuals receiving positive outcomes from the initial test follow up with doctors for the second-stage test. However, in practice, not all individuals would comply, and in some countries, the adherence rates are far below desired. For example, evidence from prior literature regarding CRC screening programs highlights that the adherence rate to the second-stage colonoscopy after receiving positive FIT outcomes varies from 54.8% to 92.5% in different countries (Navarro et al., 2017). Based on the national-wide survey ($n = 3,920$) we conducted in Singapore, among individuals who are tested positive ($n = 274$) in the first-stage FIT, more than 28% ($n = 77$) did not follow up with a colonoscopy. The challenge of low adherence to screening guidelines is well documented in the medical literature. Reasons behind the low adherence rate are related to various behavioral factors such as the lack of knowledge about the disease and screening, anxiety about getting bad news, etc (Bynum et al., 2012, Gimeno Garcia, 2012). More importantly, in real practice, the initial test accuracy information determined by the cut-off points is revealed to the public and the public can access it through the website or inquiring healthcare professionals. It is reported that the initial test with different cut-off points, which directly affect the accuracy of the initial tests, would lead to various adherence behavior, possibly due to individuals' trust on test (Plumb et al., 2017, Lee and Lee, 2018). Therefore, it is crucial to incorporate individuals' endogenous behavioral responses when designing the initial tests to balance the screening effectiveness and the economic implication with a limited second-stage test capacity.

4.1.1 Research Questions and Methodology

Despite all the discussions, no systematic approach has been proposed to address the initial test design problem, taking all the mentioned considerations into account. In this chapter, we aim to fill this gap by proposing an optimization framework that selects the cut-off value in the initial screening (i.e., FIT) to balance the trade-off between test effectiveness (i.e., detecting more cancer incidences) and the second-stage test (i.e., colonoscopy) demand, considering individuals' adherence behavior. The problem involves two sequential decisions —the health system designs a FIT by strategically choosing the cut-off point to maximize the number of cancer incidences detected while controlling the colonoscopy demand; given the designed FIT, individuals who received positive outcomes decide whether to follow up with a colonoscopy. We summarize our model and methodology in Figure 4.2.

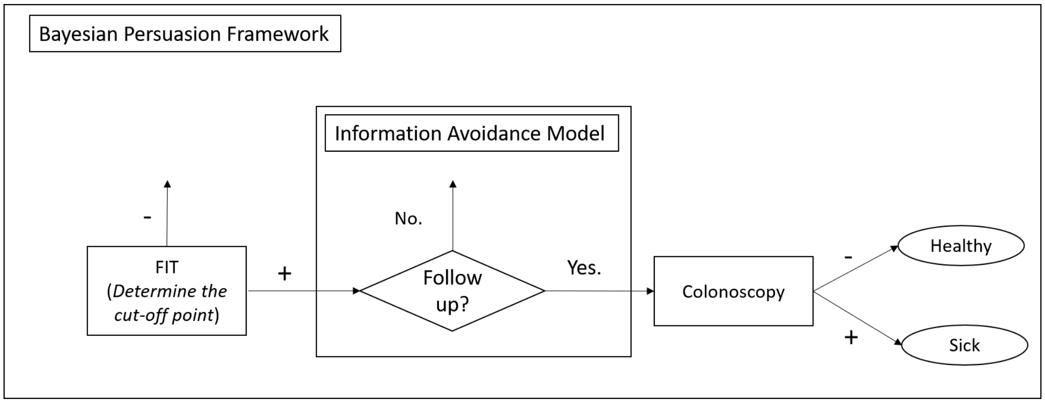


Figure 4.2: Model framework

We adopt a Bayesian persuasion framework to model interactions between the health system and the targeted population (eligible average-risk individuals). Bayesian persuasion paradigm proposed in Kamenica and Gentzkow (2011) is an information model where a sender chooses a signal to reveal to a receiver, who then takes an action that affects both players' payoff. The sender's problem is to maximize its payoff by designing information shared with the receiver in the form of signals to influence/persuade the receiver to take certain action; the receiver would choose actions that maximize his/her payoff after observing the signals. In our context, given each individual's prior belief of the health state (i.e., with or without cancer), the initial test serves as a natural signaling mechanism influencing the belief updating process. The selection of cut-off points explicitly dictates the likelihood of generating positive/negative outcomes, which, in turn, alters the individual's posterior belief of the health state. This would affect individual's decision to follow up with the second-stage test.

To model individuals' follow-up decisions and adherence behavior, we adopt a novel concept well-received in the economic literature, information avoidance. A growing theoretical and experimental literature suggests that information may directly enter the agent's utility function. This can create an incentive to avoid information, even when it is useful, free, and independent of strategic considerations (Golman et al., 2017). Given that the gold-standard second-stage test can confirm the health state accurately, individuals still choose to avoid the information. Golman et al. (2017) discuss several behavior factors that lead to information avoidance behavior. The most relevant ones include "optimism maintenance" and "risk, loss, and disappointment aversion". Optimism maintenance states that people are optimistic about their health states and choose to hold these optimistic beliefs; those whose beliefs conflict with the information tend to dismiss the information and question the source's quality or impartiality (Brunnermeier and Parker, 2005). Risk, loss, and disappointment aversion refer to the perception that the possible perceived disappointment or loss might overweight the benefit of knowing the actual health states (Golman et al., 2017). Several utility models are proposed in the literature to capture information avoidance behavior. In this chapter, we adopt one of the most applicable utility models, "optimal expectation" proposed by Brunnermeier and Parker (2005) to express an individual's follow-up utility. Specifically, individuals' expected utility is a weighted summation of an objective utility and a subjective utility. The objective utility is the expected utility with respect to the actual risk of developing cancer. The subjective utility highly relies on an individual's self-belief of their risk of developing cancer and the perceived cost of follow-up, which can be significantly different from the objective ones. When individuals are optimistic, or the perceived cost of receiving "bad news" is high, they would not follow up with a second-stage test if they trust their beliefs more than the objective information. In our survey data, we indeed observe that participants' subjective beliefs of having CRC after receiving positive FIT outcome were significantly smaller than the posterior objective risks (with a p-value less than 0.05), suggesting that individuals are overly optimistic about their health status, which may lead to non-follow-up decisions. The optimal expectation model assumes that the subjective belief is also rationally optimized to maximize the total utility. Given that we have the leverage of obtaining the subjective beliefs from real data, we relax this assumption and directly use the calibrated subjective beliefs.

By incorporating the information avoidance utility model into the Bayesian persuasion framework, we can explicitly capture the key elements of the problem of interest. Furthermore, we conduct a nationwide survey in Singapore to study the people's perceptions of CRC cancer screening guideline and their adherence behavior. The survey targets Singapore residents aged 50 and above. It covers various CRC and CRC screening-related questions, including the respondents' CRC knowledge and screening awareness, risk attitude and perceptions, factors influencing their screening adherence decisions, heath literacy, life satisfaction, etc. With other data on demographic and personal information, we have, in total, 7,899 variables from 3,920 respondents. This large-scale survey data facilitates the learning of individuals' subjective perceptions and calibration of the expected utility function. The optimal screening guideline obtained from our model gives rise to insightful and practical implications for the CRC screening design in Singapore.

Moving beyond the current practice of CRC screening, we leverage our framework to further investigate two extensions. Firstly, the current FIT is dichotomous (i.e., adopting a single cut-off point that generates positive or negative test results). We study alternative ways of reporting the first-stage test outcomes, in particular, using multiple cut-off points to generate different degrees of risk warnings. These types of tests are called *ordinal tests*, for example, a test with two cut-off points reporting three confidence ratings for the presence of disease - high risk, medium risk, low risk. When the biomarker concentration value is directed reported, the tests are called *continuous tests*, (e.g., the test of WBS count, and blood glucose level. We are interested to see whether a more refined risk stratification of the FIT would bring significant benefits. Secondly, we further explore the practice of using heterogeneous FIT kits for different sub-populations. It has been well acknowledged in the medical domain that the performance of the same FIT kit varies among different groups of individuals in terms of detection rates, adherence rate, etc. The benefits of population-cased FITs are discussed in the medical literature (Khalid-de Bakker et al., 2011, McDonald et al., 2012, Toes-Zoutendijk et al., 2020). In this chapter, we would further explore the optimal partitioning of the population and adopting heterogeneous cut-off points for each sub-population. We aim for an implementable population-cased practice that could be easily adopted by the health systems.

4.1.2 Key Results and Contributions

We summarize our main results and contributions below. Firstly, we obtain the optimal cut-off point for the FIT by balancing the trade-off between the number of CRC cases detected and demand for colonoscopy considering individuals' adherence behavior to the screening protocol. Intuitively, with a fixed adherence rate (i.e., a fixed proportion of positive cases would follow up with colonoscopy), a high- (low-) sensitivity FIT will detect more (fewer) CRC cases but result in high (low) demand for colonoscopy. Cut-off selection purely depends on the health system's cost-benefit trade-off, and every cut-off point will give rise to a FIT on the efficient frontier. However, via incorporating the practical observation that the individuals' adherence behavior endogenously depends on the design of FIT, we found that the trade-off is more involved. A well-chosen cut-off point can detect more cancer incidences with fewer colonoscopies. The conventional wisdom of pursuing high-sensitivity FITs that gives rise to a huge demand for colonoscopy could backfire and detect even fewer cancer incidences.

Secondly, our analysis contributes to the debate in the medical domain regarding the adoption of the quantitative test versus the qualitative test. The discussion is currently limited to qualitative arguments on the different implications of the two types of tests. We build an analytical tool to quantify the exact benefit of the quantitative test, which allows the health system to optimize the cut-off point. Our results further support its adoption and provide foundations for further investigation in clinical trials.

Thirdly, via modeling the initial test with possibly multiple cut-off points, we obtain the optimal structure of the test. Under some technical conditions, we show that when the health system has sufficient colonoscopy capacity and aims for all positive cases to follow up with colonoscopy, a FIT with a single cut-off point is optimal. This confirms the practice of using binary outcomes for the first-stage test. Under some other technical conditions, we also find that when the health system aims to maximize the detection rate, a continuous FIT that reports the hemoglobin values to individuals is optimal, which implies the potential benefits a continuous test might bring to cancer screening.

Finally, in our numerical study using the survey data of the Singapore population, we explore the benefits of using heterogeneous FIT kits to different sub-populations. We apply our framework to determine the optimal partitioning of the population and obtain the corresponding optimal cut-off point for each sub-population. When restricted to two sub-population groups, the optimal partitioning gives rise to one group with relatively high risk, which is assigned to a high-sensitivity FIT, and the other relatively low-risk group, which is assigned to a low-sensitivity FIT. We show that by using two different FIT kits customized to two sub-populations, significantly more CRC cases can be detected with fewer colonoscopies than a universal FIT test. To obtain a practical population-based screening policy, we further apply the interpretable clustering technique to search for implementable rules in partitioning the population. We find that by simply adopting an age-gender rule, we can recover the optimal partition.

It is worth mentioning that CRC generally starts with a polyp, a noncancerous growth developing in the colon or rectum's mucosal layer (American Cancer Society, 2020a). Polyps can also be detected by a two-stage CRC screening but with relatively low sensitivity. For instance, the sensitivity of seven different FIT brands fluctuates from 6% to 44% (Gies et al., 2018) for polyps. In this chapter, we only focus on applying the model framework to CRC detection, which, however, can also be generalized to study both polyps and CRC detection. We present this general case as an extension in Appendix C.6.

4.2 Literature Review and Related Studies

We divide the related studies on cancer screening into two main categories, studies in the operations management (OM) domain and those from medical literature. This is followed by a detailed review of the concepts and methodologies applied in our framework, information avoidance and Bayesian persuasion.

4.2.1 Operations Research and Management Science Studies on Cancer Screening

Designing cancer screening policies has received great attention in the OM field. Studies are conducted to optimize the screening protocol for various cancers, such as CRC (e.g., Erenay et al. (2014), Güneş et al. (2015)), breast cancer (e.g., Ayer et al. (2012) and Ayer et al. (2016), Cevik et al. (2018), Ayvaci et al. (2012)), and prostate cancer (e.g., Zhang et al. (2012)). A detailed review can be found from Alagoz (2011) and Alagoz et al. (2010).

For example, Ayer et al. (2012) built a POMDP model on the initial screening decisions of the breast cancer (i.e., either undergo a mammogram or wait). With the assumption that once the initial mammogram is positive, individuals will follow up with a perfect second-stage test (i.e., a biopsy), the authors found that individualizing mammography screening policy based on women's risk characteristics is crucial to maximizing the expected Quality-Adjusted Life-Years (QALYs). Erenay et al. (2014) investigated the optimal interval of performing a colonoscopy without the consideration of initial tests. With the objective of maximizing the expected QALYs, they developed a POMDP model and determined the optimal personalized CRC screening policy by incorporating age, gender, and risk of having CRC into the screening decisions. Different from the above-mentioned papers, which assume unlimited test capacity, Güneş et al. (2015) further incorporated the colonoscopy resource constraint into a dynamic compartmental model with the objectives of minimizing mortality or incidence rates when analyzing the optimal colonoscopy allocation policy for screening and diagnosis of colorectal cancer.

The studies mentioned above adopt multi-period decision models and aim to design screening policies for cancer prevention and surveillance from the perspective of the optimal screening frequency and starting/ending age of underdoing biopsies/colonoscopies. The key decisions involve whether to perform tests at each decision epoch, with test sensitivity and specificity fixed and given in advance. This chapter is a high-level test design problem, and we consider adjustable test sensitivity and specificity by optimizing the cut-off selection to balance the trade-off between test effectiveness and screening capacity.

Most OM literature assumes perfect adherence to the guideline (Ayer et al. 2012, Zhang et al. 2012). Erenay et al. (2014) and Ayer et al. (2016) acknowledged the issue of imperfect adherence in designing optimal screening policy. In Erenay et al. (2014), they numerically explored how different levels of adherence rates impact the optimal design of screening guidelines. Ayer et al. (2016) examined the effect of imperfect adherence in the initial mammography test for breast cancer by formulating a new POMDP model that sheds some light on the trade-offs inherent in different breast cancer screening policies. They concluded that screening strategies might be adjusted in clinical practice based on the adherence rate. However, the adherence rates remain fixed and exogenous, and the individual's adherence decision is disentangled from the policy design. Unlike the literature, in this chapter, we endogenize the individual's adherence decision to capture the screening policy's impact on adherence behavior.

4.2.2 Medical Studies on Screening Adherence and Cut-off Point Selection

In screening programs worldwide, the adherence rate to colonoscopy after receiving an abnormal FIT result is relatively low, with 65.7% in Taiwan, 58.6% in Chile, and 70.5% in France (Navarro et al. 2017, Pellat et al. 2018, Jen et al. 2018). According to numerous descriptive qualitative studies, many factors can influence individuals' follow-up decisions, such as economic status, education level, awareness of CRC, fear and anxiety about the colonoscopy procedures, the embarrassment of undergoing a colonoscopy, encouragement from family and friends, fear of being diagnosed with cancer, concerns about the accuracy of the initial test, health literacy and lack of assistance in making an appointment (Schneider et al. 2020, Wang et al. 2013, Plumb et al. 2017). Various interventions have been recommended to promote individual follow-up behaviors, such as helping individuals schedule appointments, providing concrete education about what the colonoscopy entails, proactively communicating costs and offering financial assistance (Schneider et al. 2020). This chapter does not focus on understanding the reasons behind imperfect adherence or propose interventions that can alter individuals' behavior. We aim to design the screening tests incorporating the adherence behavior into the model leveraging the survey data. We contribute to the medical literature by modeling the dependence between individuals' adherence decisions and the design of the screening guideline. We focus on using the initial test design as a lever to improve the overall screening effectiveness considering the strategic adherence behavior.

Regarding the selection of cut-off points, Itoh et al. (1996), Chen et al. (2007), Wong et al. (2004), Wilschut et al. (2011) and Hernandez et al. (2014) focused on improving test accuracy and effectiveness. These medical papers applied the existing decision rule in selecting the cut-off point. For example, Itoh et al. (1996) proposed three methods to find the optimal cut-off point: (1) choose the cut-off point with the highest positive predictive value; (2) choose the cut-off point that maximizes the sum of sensitivity and specificity (Youden's J index); (3) choose the cut-off point that minimizes the cost of effectiveness. The second method is widely used by researchers because of its simplicity but without considering the economic impact of the second-stage screening capacity. There are numerous studies considering the initial test designs under the constraint of second-stage screening resources. Some papers highlight that the policymaker can adjust the cut-off value based on the available colonoscopies resource as required for second-stage screening, and thus, avoid overextending the available endoscopic resources (Grazzini et al. 2009, Navarro et al. 2017, Toes-Zoutendijk et al. 2020). However, these papers do not consider the impact of individuals' adherence behavior on the demand for second-stage tests. We propose a holistic model to optimize the cut-off points considering the limited second-test capacity with imperfect adherence behavior.

There are also extensive studies that recommend customized cut-offs for subpopulations. By distinguishing subpopulations by age or gender, many medical papers advocate the use of personalized cut-offs (Khalid-de Bakker et al. 2011, McDonald et al. 2012, Toes-Zoutendijk et al. 2020). We also propose a population-based test scheme. Our optimization model is able to analytically partition the population and optimize the cut-off points for each sub-population.

4.2.3 Studies on Information Avoidance Behavior

A growing literature adopts the information avoidance concept to explain/analyze an individual's healthcare decisions. A detailed review can be found in Golman et al. (2017). Most recently, Li et al. (2020) discovered evidence in their field experiment for information avoidance in the context of testing for diabetes and cancer. Several belief-based utility models are applied to explain information avoidance. Following from Li et al. (2020), these utility models can be divided into three categories: the optimal expectations model (Brunnermeier and Parker 2005, Oster et al. 2013), the attention model (Karlsson et al. 2009, Golman and Loewenstein 2018, Golman et al. 2019, Ganguly and Tasoff 2017), and the curvature model (Caplin and Leahy 2001, Caplin and Eliaz 2003, Kőszegi 2003, Eliaz and Spiegler 2006). All three models assume that people derive anticipatory utility: beliefs about future events and outcomes affect current utility. In this chapter, we borrow the optimal expectations model to develop individuals' utility model. The optimal expectations model allows for self-manipulation of beliefs: Individuals can maintain biased beliefs to generate high anticipatory utility, thus avoiding the tests. We use the survey data to calibrate the utility function and characterize individual's follow-up decision. Most of the works mentioned above discuss information avoidance behavior in a conceptual way to provide high-level insights. Our work contributes to information avoidance literature by demonstrating its applicability in modeling the individual's healthcare decision in cancer screening. We use the real data to show that the information avoidance utility model can well represent the actual screening adherence behavior.

4.2.4 Studies on Bayesian Persuasion

Kamenica and Gentzkow (2011) has a seminal contribution by posing the "Bayesian persuasion" problem, in which a single informed principal (sender) chooses which information to collect and communicate to an uninformed agent (receiver) to motivate her to act in the desired way. They later extended the model to large state spaces and multiple agents (Gentzkow and Kamenica 2016, 2017). More recently, this basic framework has been developed for different domains with applications including price discrimination (Bergemann et al. 2015), monopoly pricing (Roesler and Szentes 2017), auctions (Bergemann et al. 2017), unobserved queuing system (Lingenbrink and Iyer 2019), and medical testing or treatment (Schweizer and Szech 2018, Xiang 2020). Specifically, (Schweizer and Szech (2018)) studied optimal information revelation scheme motivated by Huntington's Disease. They aimed to show revealing partial information of test outcome is beneficial compared with conventional tests, which always report precise outcomes. Xiang (2020) focused on physician-patient interaction in cervical spondylosis treatment decisions. She characterized this interaction in a Bayesian persuasion framework and test model implication using health insurance claims data. Our work also applies the Bayesian persuasion framework to the medical testing design; however, focusing on a concrete context of cancer screening program design to balance the trade-off between screening effectiveness and economic burden. Among all the known papers adopting this framework, we are the first one using real data to showcase the applicability power of Bayesian persuasion.

Outline of this chapter: The rest of the paper is organized as follows. Section 1.3 details the problem setup and the two-stage optimization model for the initial test design problem. Section 1.4 presents the theoretical results of the optimal test structure. In Section 1.5, we conduct a case study in the context of Singapore CRC screening and evaluate the performance of the optimal initial tests. We conclude our research by highlighting the major findings and show an outlook on future research in Section 1.6. All the proofs are relegated to Appendix C.2.

4.3 Model

In this work, following the Bayesian persuasion paradigm, we propose a two-stage optimization framework to study the initial test design problem in the context of CRC screening. In the first stage, the health system aims to maximize the overall expected probability of follow-up from individuals with CRC and simultaneously control the total expected demand for colonoscopy by strategically designing the FIT. In the second stage, based on the designed initial test, individuals' goal is to maximize their total expected utility by choosing whether to follow up with a colonoscopy after receiving FIT results. For easy reference, all the notations are presented in Appendix C.1.

4.3.1 Initial Test

Let $s \in S = \{0, 1\}$ denote the CRC state for individuals, where $s = 1$ ($s = 0$) represents a health state with (without) CRC. We use ζ to denote f-Hb concentration level tested from FIT. Conditioning on the individual's health state, ζ follows different distributions. Let $H_1(\cdot)$ ($H_0(\cdot)$) and $h_1(\cdot)$ ($h_0(\cdot)$) be the cumulative distribution function (CDF) and probability density function (PDF) of f-Hb concentration for individuals with (without) CRC (Figure 4.1). Suppose the range of f-Hb concentration is $[\underline{\zeta}, \bar{\zeta}]$. We assume that $h_0(\zeta), h_1(\zeta) > 0$ and are continuous in ζ for $\zeta \in [\underline{\zeta}, \bar{\zeta}]$.

To design a quantitative test, the health system chooses a set of cut-off points from the feasible range $[\underline{\zeta}, \bar{\zeta}]$. Denote the test design decision variable as a pair $(T, \mathcal{C}_T)$, where T refers to number of cut-off points selected, and $\mathcal{C}_T$ as the set of cut-off point values, i.e., $\mathcal{C}_T = \{c_1, c_2, .., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$. Given $(T, \mathcal{C}_T)$, the health system reports the test outcome after observing f-Hb value, ζ. $(T, \mathcal{C}_T)$ gives rise to $T + 1$ test outcomes in outcome set $\Gamma^{\mathcal{C}_T} = \{0, 1, 2, ..., T\}$: test outcome is 0 if $\zeta \leq c_1$, t if $c_t < \zeta \leq c_{t+1}$, $t \in \{1, 2, ..., T - 1\}$, or T if $\zeta > c_T$. Note that when $T = 1$, we have a dichotomous test which generates positive (1) or negative (0) outcomes. $T > 1$ corresponds to ordinal tests. We define $\sigma^{\mathcal{C}_T}(t|s)$ as the likelihood of receiving test outcome t if individual's state is s. Given the distribution functions of f-Hb concentrations for healthy and sick individuals, we have, for $s \in \{0, 1\}$, $\sigma^{\mathcal{C}_T}(0|s) = H_s(c_1)$, $\sigma^{\mathcal{C}_T}(t|s) = H_s(c_{t+1}) - H_s(c_t)$ if $t \in \{1, 2, ..., T-1\}$ and $\sigma^{\mathcal{C}_T}(T|s) = 1 - H_s(c_T)$. This information is announced to the public. Based on the likelihood of receiving each test outcome, individuals will update their belief of having cancer. Figure 4.3 gives an illustrative example of an initial test with two cut-off points. Once $(T, \mathcal{C}_T)$ is determined, we can fully characterize the initial test in terms of the test outcomes and likelihoods. We will omit the dependence on $\mathcal{C}_T$ in the notations if it is clear from the context.

In practice, a well-designed cancer screening initial test should possess a property that individuals with cancer are more likely to receive severe test outcomes than healthy individuals. This property can be represented by the "Monotone Likelihood Ratio (MLR)" stochastic order condition: A random variable X (with probability mass function q_1) is said to dominate Y (with probability mass function q_0) in the sense of MLR if $\frac{q_1(\cdot)}{q_0(\cdot)}$ is an increasing function (Eeckhoudt et al. 2011). Given that the test outcomes of individuals with and without CRC are two random vectors of dimension $T + 1$ with probability mass function $\sigma^{\mathcal{C}_T}(\cdot|0)$ and $\sigma^{\mathcal{C}_T}(\cdot|1)$, we define an initial test that possesses this nice property as a "MLR-feasible" initial test.

Property 1. *An initial test $(T, \mathcal{C}_T)$ is MLR-feasible if $\frac{\sigma^{\mathcal{C}_T}(t|1)}{\sigma^{\mathcal{C}_T}(t|0)}$ is increasing in t, $t \in \Gamma^{\mathcal{C}_T}$.*

Property 1 implies that if an initial test is MLR-feasible, then individuals with CRC are more likely to receive severe outcomes than healthy individuals. In the rest of this chapter, we restrict our study to MLR-feasible initial test.

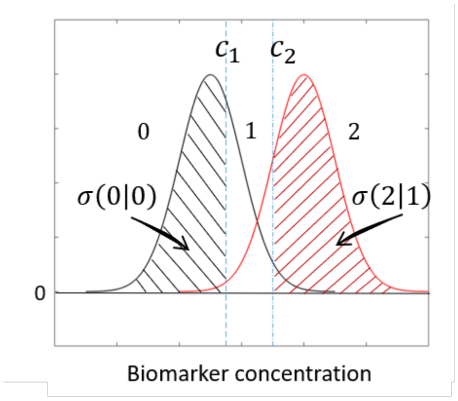


Figure 4.3: An initial test with two cut-off points
The left curve and the right curve present PDFs of biomarker concentrations for healthy and sick individuals. Suppose $\mathcal{C}_T = \{c_1, c_2\}$, where $c_1 < c_2$. The test outcome set is $\Gamma^{\mathcal{C}_T} = \{0, 1, 2\}$. Based on the definition of $\sigma^{\mathcal{C}_T}(t|s)$, the left shaded area denotes the likelihood of healthy individuals receiving test outcome 0, and the right one presents the likelihood of sick individuals receiving test outcome 2.

4.3.2 Individual's Follow-Up Problem

Individual's belief updating process.

We consider a population of size N. For an individual i, $i \in \{1, 2, ..., N\}$, the prior risk of developing CRC is denoted by p_i^0. After receiving a test outcome $t \in \Gamma$, the risk of developing CRC is updated to $p_i^s(t)$ (i.e., the posterior risk) following Bayesian updating process. Let $\lambda_i(t)$ denote the total probability of receiving test outcome t, i.e., $\lambda_i(t) = \sigma(t|0)(1 - p_i^0) + \sigma(t|1)p_i^0$, $\forall t \in \Gamma, i \in \{1, 2, ..., N\}$. We have

$$p_i^s(t) = \frac{\sigma(t|1)p_i^0}{\lambda_i(t)}, \quad i \in \{1, 2, ...N\},\ t \in \Gamma.$$

The following lemma establishes the monotonicity of $p_i^s(t)$ for an MLR-feasible initial test.

Lemma 4.1. *The posterior risk of developing CRC after taking a MLR-feasible initial test, $p_i^s(t)$, increases in t.*

Lemma 4.1 implies that individuals have higher posterior risks of having cancer if receiving worse test outcomes.

Individuals' utility model.

We denote individual i's follow-up action observing initial test outcome t as $a_i(t) \in \{0, 1\}$, where $a_i(t) = 1$ refers to a follow-up with the second-stage test and $a_i(t) = 0$ indicates the opposite. We use $u_i(s_i, a_i(t))$ to denote the utility for individual i if his/her health state is s_i and the follow-up action is $a_i(t)$. Following the convention in healthcare practice, we adopt QALYs as the performance measure of the individual's utility. QALYs is a generic measure of the value of health outcomes, including both the quality and the length of life lived. Mathematically, it is the product of the length of life in years and quality of life, where the quality of life is measured on a scale of 0 to 1, with 1 indicating perfect health and 0 indicating death.

Adapted from the utility model in information avoidance literature (Brunnermeier and Parker 2005, Oster et al. 2013), we construct the individual's utility function when making a follow-up decision as a weighted sum of an *subjective utility* and a *objective utility* by incorporating both the subjective beliefs and objective risks of their health status, subtracted by an *expected perceived disutility* of taking a colonoscopy if he/she follows up.

Subjective utility and objective utility. We denote the weighting parameters of subjective and objective utility by δ_i and $1-\delta_i$, respectively. Recall that individual's objective risk of developing CRC after receiving a test outcome t from the initial test is the posterior risk, $p_i^s(t)$. We define a subjective counterpart, denoted as, $\pi_i^s(t)$, to represent individual i's subjective belief of having CRC after receiving a test outcome t from the initial test. Therefore, we have first two components of the total expected utility given as follows:

$$\delta_i E[u_i(s_i, a_i(t))|\pi_i^s(t)] + (1 - \delta_i) E[u_i(s_i, a_i(t))|p_i^s(t)].$$

Note that to obtain above expected utility, we need to know the individuals' subjective posterior belief, $\pi_i^s(t)$, for any given $t \in \Gamma$. However, in the current practice, the dichotomous test with binary test outcomes, i.e., $t \in \{0, 1\}$, is being used. In order to obtain the subjective belief for any initial test design with possibly more than two test outcomes, we model individual i's subjective belief $\pi_i^s(t)$ as a function of the objective risk $p_i^s(t)$, i.e., $\pi_i^s(t) = \Phi(p_i^s(t))$, $t \in \Gamma$, where function Φ is carefully calibrated using the survey data. The correlation between objective risk and subjective belief can be well-interpreted. For instance, if an individual maintains a good (bad) lifestyle and has a lower (higher) posterior risk of having CRC, the self-belief of having CRC is also likely to be low (high). Moreover, in classic information avoidance literature (Oster et al. 2013), the researchers analyze the relationship between individuals' subjective beliefs and objective risk and find that the optimal subjective belief is a linear function of objective risk.

Expected perceived disutility of follow-up. There is a cost incurred if an individual chooses to follow up and takes a colonoscopy. The cost stems not only from the possible adverse effect from taking the colonoscopy, such as the risk of perforation or even death, but also other personal resistance and concerns individual have towards taking a colonoscopy and the follow-up treatment if confirmed with CRC, such as their age, medical history, family support, trust on doctors, perception on the discomfort and embarrassment of taking a colonoscopy, etc. We refer to the overall cost as the "perceived disutility" of the follow-up action, denoted as $d_i(s_i)$ for individual i if the health state is s_i. Because it is only incurred when $a_i(t) = 1$, we omit the dependence on $a_i(t)$ in the notation. If individual i follows up, the expected perceived disutility is $E[d_i(s_i)|\pi_i^s(t)]$. Note that the subjective belief of developing CRC is used as the probability distribution of s given that the disutility is an individual's perceived factor other than an objective evaluation.

To summarize, the total expected utility of participant i for action $a_i(t)$ is given as follows:

$$\begin{aligned} U_i(a_i(t) = 0) =& \delta_i E[u_i(s_i, a_i(t) = 0)|\pi_i^s(t)] + (1 - \delta_i) E[u_i(s_i, a_i(t) = 0)|p_i^s(t)] + \epsilon_i^0, \\ U_i(a_i(t) = 1) =& \delta_i E[u_i(s_i, a_i(t) = 1)|\pi_i^s(t)] + (1 - \delta_i) E[u_i(s_i, a_i(t) = 1)|p_i^s(t)] - E[d_i(s_i)|\pi_i^s(t)] + \epsilon_i^1. \end{aligned}$$

Where $\epsilon_i^{a_i}$ is a random error term that captures the impact of all unobservable factors which affect the utility of choosing action a_i by individual i.

Individual's follow-up decision.

In the context of cancer screening, an individual's follow-up decision does not solely rely on the utility of different choices. For some specific test outcomes, individuals would automatically not follow up with a second-stage test. Take a dichotomous FIT test as an example. When an individual receives a negative outcome (0) which indicates a by-default health state, he/she will not follow up. In the case of a positive outcome, the follow-up decision would be made by weighing the utility of different actions. To reflect this additional feature, we obtain the following condition that determines individuals' follow-up behavior.

$$a_i(t) = \begin{cases} 1 & \text{if } t \neq 0 \text{ and } U_i(a_i(t) = 1) > U_i(a_i(t) = 0), \\ 0 & \text{if } t = 0 \text{ or } U_i(a_i(t) = 0) \geq U_i(a_i(t) = 1). \end{cases}$$

Specifically, we add one additional condition that if an individual receives the best test outcome (i.e., $t = 0$), he/she would not follow up. Given that the total utility $U_i(a_i(t))$ is random, for an individual i with an initial test outcome t, the probability of following up with the second-stage test, $f_i(t)$, is expressed as follows:

$$f_i(t) = \begin{cases} 0 & \text{if } t = 0, \\ \text{Prob}(U_i(a_i(t) = 1) > U_i(a_i(t) = 0)) & \text{otherwise.} \end{cases} \tag{4.1}$$

Since the subjective belief $\pi_i^s(t)$ is a function of the objective risk $p_i^s(t)$, individuals' follow-up probability can be written as a function of $p_i^s(t)$ if $t \neq 0$. In the rest of the paper, we denote this by function $W(p_i^s(t))$.

$$f_i(t) = \begin{cases} 0 & \text{if } t = 0, \\ W(p_i^s(t)) & \text{otherwise.} \end{cases} \tag{4.2}$$

4.3.3 Health System Test Design Problem

The health system's objective is to maximize the expected follow-up probability from individuals with CRC and control the total expected demand for colonoscopy by selecting the set of cut-off points. Given $\mathcal{C}_T$ is the decision variable denoting a set of cut-off points with the corresponding test outcome set $\Gamma^{\mathcal{C}_T}$ and likelihood set $\{\sigma^{\mathcal{C}_T}(t|s)|t \in \Gamma^{\mathcal{C}_T}, s \in S\}$, we formulate the health system's problem as follows.

$$\begin{array}{ll} \max_{T,\mathcal{C}_T} & \sum\limits_{i=1}^{N} \sum\limits_{t \in \Gamma^{\mathcal{C}_T}} [f_i(t)\sigma^{\mathcal{C}_T}(t|1)p_i^0 - \tau f_i(t)\sigma^{\mathcal{C}_T}(t|0)(1-p_i^0)] \\ \text{s.t.} & f_i(t) \text{ satisfying (4.2)} \end{array} \tag{4.3}$$

Specifically, the first term in the objective function represents the follow-up rate from individuals with CRC; the second term corresponds to the follow-up rate from healthy individuals. We introduce a multiplier τ in front of the second term to capture the "cost" of extra colonoscopy demand from healthy individuals. Adjusting values of τ would give rise to different levels of control over colonoscopy demand.

4.4 Optimal Initial Test Design

There are two challenges in analyzing Problem (4.3). First, the space of feasible designs of the initial test is large, where, essentially, any value between $[\underline{\zeta}, \bar{\zeta}]$ is a feasible cut-off point value. Given that the health system can adopt a test with multiple cut-off points, there are exponentially many feasible solutions even with discretized feasible space. Secondly, the test likelihood functions σ and the best response follow-up probability functions, f, are highly nonconvex.

To address those challenges, we first study two special cases and explore the optimal structure of the initial tests theoretically. Specifically, we investigate (1) how many cut-off points should be utilized and (2) the optimal values of the cut-off points. The first special case is that the health system aims to encourage all individuals receiving risk alarms to follow up with the second-stage tests. This scenario refers to achieving a full screening guideline compliance where all individuals receiving alarms are not differentiated and should confirm their health status via a second-stage test. With this goal, the capacity of colonoscopy is not a primal concern. Mathematically, it corresponds an objective function with $\tau = -1$. We refer to this scenario as the "compliance maximization case". The second special case relates to enhancing the screening guideline's effectiveness by maximizing the expected follow-ups of individuals with CRC. This corresponds to an objective function with $\tau = 0$. We call this scenario the "effectiveness maximization case".

For a general objective function with an arbitrary value of τ, answering how many cut-off points should be selected is challenging. We restrict our analysis to finding the optimal cut-off point value for a dichotomous initial test from a pre-specified discretized set of cut-off candidates. We then solve Problem (4.3) analytically via an integer programming reformulation.

4.4.1 Compliance Maximization Case

For most of the two-stage CRC screening guidelines in practice, individuals receiving positive FIT outcomes are all encouraged to follow up with a second-stage test. This guideline aims to achieve higher individuals' compliance without differentiating these individuals based on their potential risk of developing CRC. In this section, we particularly focus on this initiative and characterize the optimal initial test design (i.e., $\tau = -1$).

Specifically, we say an initial test outcome, t, is a risk alarm for an individual i if $t \neq 0$. The health system aims to maximize the total follow-up probability for all individuals receiving risk alarms (Problem (4.4)).

$$\begin{aligned} \max_{T,\mathcal{C}_T} \quad & \sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T}} \left[f_i(t)\sigma^{\mathcal{C}_T}(t|1)p_i^0 + f_i(t)\sigma^{\mathcal{C}_T}(t|0)(1-p_i^0) \right] \\ s.t. \quad & f_i(t) \text{ satisfying } (4.2) \end{aligned} \tag{4.4}$$

Notice that in Problem (4.4), the follow-up probability is summed over all the test outcomes $t \in \Gamma^{\mathcal{C}_T}$. If t is not a risk alarm, i.e., $t = 0$, the follow-up probability is 0. Hence, Problem (4.4) is equivalent to only maximizing the follow-up probability of individuals receiving risk alarms. The following theorem presents the optimal structure of the initial test.

Theorem 4.2. *For the compliance maximization case, if $W(p_i^s(t))$ is concave in $p_i^s(t)$, a dichotomous initial test with cut-off point value $\underline{\zeta}$ is optimal.*

Theorem 4.2 confirms the current practice of adopting a dichotomous initial test. In terms of selecting cut-off values, it is suggested in the literature that a lower cut-off value should be applied for FIT screening to achieve a higher detection rate without considering the high false-positive rate and the demand for unnecessary colonoscopies (Hol et al. 2009). Our result further supplements this claim by showing that high sensitivity FIT can be optimal under certain conditions, even considering the adverse impact of a high false-positive rate on adherence behavior.

We also consider the case when $W(\cdot)$ is not a concave function and analyze the performance of adopting the dichotomous FIT with the highest sensitivity. We found that the performance gap between such a FIT test with the optimal test design is bounded by a finite value measured by the "modulus of concavity" of $W(\cdot)$. The detailed analysis is relegated to Appendix C.3.

4.4.2 Effectiveness Maximization Case

In this section, we examine the optimal initial test design when the health system aims to enhance the effectiveness of CRC screening by maximizing the overall follow-up probability from individuals with CRC. This corresponds to an objective with $\tau = 0$, which gives the following health system's optimization problem.

$$\begin{aligned} \max_{T,\mathcal{C}_T} \quad & \sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T}} f_i(t) \sigma^{\mathcal{C}_T}(t|1) p_i^0 \\ s.t. \quad & f_i(t) \text{ satisfying (4.2)} \end{aligned} \tag{4.5}$$

The following theorem establishes the optimal initial test for the effectiveness maximization case.

Theorem 4.3. *For the effectiveness maximization case, if $W(p_i^s(t))$ is convex and nondecreasing in $p_i^s(t)$, it's optimal to adopt the continuous initial test which directly report individuals' f-Hb values.*

In the proof, we establish the optimally of a continuous test by firstly show that for any initial test with a finite number of cut-off points, the objective value of Problem (4.5) is nondecreasing if arbitrarily adding one more cut-off point from $[\underline{\zeta}, \bar{\zeta}]$. Then we prove that if we uniformly add infinitely many cut-off points, the objective value of Problem (4.5) converges to that of the continuous test[1]. We note that in practice, continuous tests are adopted in certain health screenings (e.g., blood glucose, WBC count). Our result sheds light on the potential benefits a continuous test might bring to cancer screening.

4.4.3 Analytical Model for the Initial Test Design

In this section, we investigate the initial test design under a general setting where the adjusting term τ can take any arbitrary value. We restrict our analysis to designing an optimal dichotomous initial test given a pre-specified candidate set of cut-off points, denoted by $\mathcal{V}_m \equiv \{v_1, v_2, ..., v_m | v_j \in [\underline{\zeta}, \bar{\zeta}], v_1 < v_2 < ... < v_m\}$. Apart from tractability, the reason that we introduce a candidate set is due to practical considerations. In practice, biomarker measurements usually come with precision issues, and the displayed biomarker values are generally not continuous. In addition, professional practitioners are likely to have a set of preferred cut-off values, and discretized cut-off point values are more interpretable than continuous values.

In the following content, we study the initial test designs under two different

[1] For the continuous CRC test, it should possess a similar property that individuals with cancer are more likely to receive severe outcomes than healthy individuals. The belief updating process is similar to the ordinal test. We relegate all the descriptions in Appendix C.2.3.

settings: the *universal test* and the *customized test.* For a universal initial test, the health system adopts a one-FIT-all test in which the same FIT kit is used for all individuals. For a customized initial test, different subgroups of the population are assigned to different FIT kits (i.e., FIT kits with different cut-off points).

Universal dichotomous test design

To design the universal test for all N individuals, the health system chooses one cut-off from the candidate set $\mathcal{V}_m$. We use a binary variable x_j to denote whether cut-off value v_j is selected, $j \in \{1, 2, ..., m\}$. If v_j is chosen, x_j equals 1; otherwise x_j equals 0. For each candidate cut-off point value, say v_j, follow-up probability, $f_i^j(t)$ and likelihood, $\sigma_j(t|s)$, can be evaluated. We have the health system's design problem formulated as an integer programming problem as follows.

$$\begin{aligned} \max_{\mathbf{x}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{m} f_i^j(1)[\sigma_j(1|1)p_i^0 - \tau\sigma_j(1|0)(1-p_i^0)]x_j \\ s.t. \quad & \sum_{j=1}^{m} x_j = 1 \\ & \mathbf{x} \in \{0,1\} \end{aligned} \tag{4.6}$$

Customized dichotomous test design

In this section, we consider customized tests for different individuals and optimize the values of cut-off points for each test to elicit the highest objective value. Suppose no more than L types of dichotomous tests are designed for all participants N, where $L << N$. Given the candidate set $\mathcal{V}_m$, we use binary variables x_j to denote whether a FIT kit with cut-off point v_j is selected, and binary variable q_{ij} to denote whether the FIT kit with cut-off point v_j is assigned to individual i. The initial test design problem again can be formulated as an integer programming problem.

$$\begin{aligned}
\max_{\mathbf{x},\mathbf{q}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{m} f_i^j(1)[\sigma_j(1|1)p_i^0 - \tau\sigma_j(1|0)(1-p_i^0)]q_{ij} & \\
s.t. \quad & \sum_{j=1}^{m} q_{ij} = 1, & \forall i \in [N] \\
& \sum_{j=1}^{m} x_j \leq L & \\
& q_{ij} \leq x_j, & \forall j \in [m], i \in [N] \\
& \mathbf{x} \in \{0,1\}, \mathbf{q} \in \{0,1\} &
\end{aligned} \tag{4.7}$$

Data-driven interpretable clustering test design

Notably, the output of the customized test design gives rise to the optimal cut-off points for L initial tests and a partition of the population into L clusters that maximizes the health system's objective. However, in most cases, the obtained clusters are not partitioned based on factors that can be easily interpreted and the partition cannot be easily implemented.

We further propose another method that designs customized initial tests to L sub-populations that are well-partitioned based on demographic features that are observable to the health system. We adopt the interpretable clustering framework (Bertsimas and Kallus 2020, Mundru 2019) which is a two-step method composed of prediction and optimization. A decision-tree model is firstly trained using individuals' demographic data given the cluster of each individual obtained from the customized test design (Problem (4.7)). The obtained decision tree generates a new partition of the population which only depends on the demographic information. The optimal initial tests are then obtained via solving the optimal universal test model (Problem (4.6)) for each cluster obtained from the new partition.

4.5 CRC Screening Test Design in Singapore

With increasing advocacy on CRC screening, the Health Promotion Board in Singapore launched the national CRC screening programme in July 2011 to encourage regular screenings for Singaporean or permanent residents (PRs) aged 50 and above (Singapore Health Hub 2011). FIT kits are distributed free of charge at the Singapore Cancer Society and can also be purchased through participating stores such as pharmacy outlets at an affordable price. Individuals are taught how to collect the stool samples and are required to return the kits within a week. Results will be ready in a month from the submission of the kits to the testing lab, and should there be positive results, the health system staff will contact participants and help them make an appointment with a hospital for further testing. Singapore has adopted two-sample FIT initial test scheme where individuals are required to collect two stool samples on two seperate days, and if at least one positive outcome is reported, a second-stage test is recommended. Two-sample FIT scheme has been shown to increase sensitivity by potentially detecting CRC that has been missed in one-sample FIT (Lim et al. 2020, Chew et al. 2009). In this section, we conduct a comprehensive study to explore the optimal initial test design in Singapore.

4.5.1 Survey and Data

The data used in our numerical study is mainly from nationwide survey data we collected in Singapore, public data sources and related literature. In this section, we mainly present the details of the survey data. Other relevant data will be introduced when they are used in the model calibration.

Survey The survey was conducted through the Singapore Life Panel (SLP), a high-frequency survey panel involving elderly participants aged from 50 years and above, which coincides with our research target audience. Each month, the panel is surveyed on their demographics, lifestyle, insurance information, health and probability literacy, life satisfaction, etc. Our survey on CRC screening is an added module of one of the monthly surveys, which consists of questions related to past CRC screening experiences, CRC and CRC screening knowledge, perception over test accuracy, and factors influencing adherence behavior, etc. The participation of our survey module is on a voluntary basis, and those who complete the whole module are rewarded with a five-dollar voucher. The study was approved by the Institutional Review Board at Singapore Management University.

SLP invited 7,539 participants and there is a total of 3,920 responses in our survey module. Given that the target population of CRC screening guideline is individuals with average risks of CRC (Ministry of Health Singapore 2020b), we exclude individuals with a family history of CRC and medical history of CRC or polyps ($n = 638$). Thus, we end up with a total number of 3,282 data points. Out of the 3,282 participants, 1,400 participants have undergone FIT screening, and 202 of them have ever received a positive FIT result. However, only 72% of the participants with positive FIT results ($n = 145$) followed up with the second-stage tests.

Survey data pre-processing There are two potential issues with the survey data, missing values and selection bias. We treat these missing values through multiple imputations using the multivariate imputation by chained equations given that missing values accounted for less than 5% for every question. Selection bias can stem from two sources, the initial selection of SLP, and the subsequent participation of our survey module, which is on a voluntary basis. Given that the survey data is used in optimizing the initial test design applied to the entire Singapore population, it is essential to address these two sources of potential bias. For the first source, according to Vaithianathan et al. (2018), the SLP is capable of representing Singapore's elderly population. We further compare the demographics of the SLP with the Singapore census of population data (Statistics Singapore 2018), and find a close match in terms of age, gender, marital status, ethnicity, education, labor force status, income and expenditure. For the second source of selection bias, we apply a response propensity weight adjustment introduced by Brick (2013) by taking the inverse of the estimated propensities of the respondents.

4.5.2 Parameter Estimation

The key inputs that need careful calibrations include f-Hb concentration distribution functions for individuals with and without CRC, h_1 and h_0; and individual's follow-up utility function, U_i. Specifically, the establishment of U_i requires learning of the subjective belief of having CRC given any test outcome t, QALYs estimation for utility term $u_i(s_i, a_i(t))$, constructing the perceived disutility when individual i follows up, $d_i(s_i)$, and determining the weighting parameter δ_i. The specific functional forms of h_1 and h_0 are extracted from Peng et al. (2019), as for the other variables, we present the estimation details, and a summary of estimation methods and data required is given in Table 4.1.

Parameter/ Function		Estimation Method	Data
$U_i(a_i(t))$	$\pi_i^s(t) = \Phi(p_i^s(t))$	Regression	Survey data
	$u_i(s_i, a_i(t))$	Simulation	Medical literature Public data
	δ_i	MLE*	Survey data Medical literature
	$d_i(s_i)$		

* MLE refers to maximum likelihood estimation.

Table 4.1: A summary of estimation methods and data sources

Subjective belief estimation

Recall that an initial test outcome is a risk alarm if it is not the best outcome. Given that individuals who do not receive risk alarms will not follow up, we only need to focus on the subjective belief estimation for those receiving risk alarms. Following the literature (Oster et al. 2013), the linear form between the subjective belief and objective risk is assumed. Specifically, the relationship between the subjective belief of having CRC and objective posterior risk, i.e., $\pi_i^s(t) = \Phi(p_i^s(t))$, is linear and same for any test outcome $t(\neq 0)$. Given the data pair of p^s and π^s for each participant i in our survey, we can, therefore, calibrate Φ using a linear regression. The obtained relationship is $\pi_i^s(t) = 0.89 \times p_i^s(t) - 0.0042$. The coefficients are significant at 1%

level.

Estimation of $u_i(s_i, a_i(t))$

We use individuals' expected remaining QALYs as a measure of utility, $u(s, a(t))$. To obtain the expected QALY, we build a simulation model of the natural progress of CRC for an individual with CRC screening intervention at individualized risk for CRC, adopted from the model in Ladabaum et al. (2001).

An individual without CRC will receive the full expected remaining QALYs[2]. The expected remaining QALYs vary by age and gender and are calculated via data extracted from Singapore life table (Department of Statistics 2019). We assume the individuals participating in the FIT are asymptomatic; otherwise, they would bypass the CRC screening and directly consult their physicians for further treatment, and therefore, are not our study audience. Hence, for an individual with CRC, he/she is possibly staying in the localized or regional CRC stage. Following Ladabaum et al. (2001), we assume that CRC cases progress from localized to regional (2 years in each state) to distant unless symptoms lead to diagnosis and treatment. If this occurs, patients will enter postcancer surveillance. If an individual with CRC follows up after receiving a positive FIT outcome, they will receive cancer treatment; if not, they will only receive treatment until the presentation of symptoms, and they may experience natural death during the CRC progression process. Data used in calculating QALYs for CRC individuals are from the public database and medical literature. Please refer Appendix C.4.1 for the detailed simulation model and QALYs derivation.

Utility functional form estimation

In this part, we present the estimation of the weighting parameter of the subjective utility (i.e., δ_i), the perceived disutility if individuals follow up with the second-stage

[2] Note the follow-up decision does not affect the health individual's QALYs calculation in this part because the QALYs loss from a colonoscopy if he/she follows up is incorporated in the disutility function d_i.

test (i.e., $d_i(s_i)$).

δ_i **and δ-features** People with different characteristics demonstrate various attitudes towards screening guidelines and may exhibit drastic differences in their objectivity towards test outcomes, captured by the weighting parameter δ_i. We utilize the demographics and personal characteristics data from the survey to calibrate the heterogeneity in δ. Specifically, from a total 7,899 variables in the survey data, we perform an initial variable selection (cf. Appendix C.4.2 for the details). We refer to all the selected variables as *δ-features.* Specifically, δ-features contain individuals' age, knowledge about colonoscopy, knowledge about CRC incidence rate, whether they have private insurance, and frequency of taking FIT. δ_i is modeled as a linear function of the δ-features.

d_i **and d-features** The perceived disutility of follow-up consists of two components: the QALYs loss due to a colonoscopy and a perceived cost due to concerns about the colonoscopy and treatment if detected with CRC. Firstly, colonoscopy is accompanied by a certain risk of perforation, which may further lead to perforation-related death. This QALYs loss is estimated similarly to the QALYs estimation in Section 4.5.2 with details relegated in Appendix C.4.2. Secondly, individuals generally have various personal concerns over the follow-up decision. In our survey, we asked individuals about the factors they were concerned about when deciding whether to follow up. After an initial variable selections (cf. Appendix C.4.2 for the details), the following key factors are identified, medical history, age (i.e., too old for treatment), trust on doctors, whether they want to know health condition, price of a colonoscopy and family support. Together with QALYs loss, all the factors are termed as *d-features.* Perceived disutility is modeled as a linear function of the d-features.

Maximum likelihood estimation (MLE) of utility function. We assume the random noises ϵ_i^0 and ϵ_i^1 follow Gumbel distribution with location parameter 0 and scale parameter 1. As a result, the follow-up probability given positive test outcomes follows a multinomial logit choice model. We then perform MLE to estimate the linear coefficients of δ-features and d-features (cf. Appendix C.4.2 for the details).

Performance evaluation of the utility model

To assess the performance of our proposed utility model, we first test the capability of our model in predicting individuals' follow-up behavior using the average AUC from 3-fold cross-validation as a performance measurement. Due to the imbalanced dataset (the number of follow-ups is 145 and the number of non-follow-ups is 57), we adopt the resampling method to generate a balanced dataset (145 follow-ups and 145 non-follow-ups). The average AUC is 0.82.

In addition, we compare the performance with a logistic regression model, which directly predicts individuals' follow-up behaviors using variables in the survey data. We perform stepwise logistic regression (cf. Appendix C.4.3) for variable selection and its average AUC of 3-fold cross-validation is 0.82, which is same as the performance of our structural model.

Although we can employ regression models to characterize individuals' follow-up decisions, our utility model is superior to the logistic model in two regards. First, using the structural model, we can capture the belief updating process and endogenize the behavior response to the test design, which allows us to further investigate the optimal test design. This is not achievable with a regression model. Second, we are able to capture individuals' information avoidance behavior and to understand how individuals evaluate subjective and objective utility, which enables us explicitly identify the impact of individualized factors on their adherence behavior.

4.5.3 Optimal FIT Design

In this section, we apply our framework to design FIT test kits in the context of Singapore CRC screening given the parameters and estimates obtained. Our study sample is the propensity-adjusted whole participant cohort. Notably, one-sample and two-sample FIT are widely adopted in different national screening programs. As previously introduced, the two-sample FIT is performed in Singapore as it can detect more CRC patients relative to the one-sample test. In some other regions and countries (e.g., Taiwan, Spain), the one-sample FIT is recommended. The U.S. Preventive Services Task Force also suggests a one-sample annual FIT screening (Robertson et al., 2017). There is a growing debate on how many samples should be utilized in the FIT test screening programs (van Roon et al., 2011, Goede et al., 2013). In the following, we explore the optimal FIT design for both one-sample and two-sample cases. The design outcomes would contribute to the performance comparison of the two cases in terms of screening effectiveness and colonoscopy demand. For the sake of brevity, the details of how we construct the candidate set of cut-off points are presented in Appendix C.4.4.

One-sample FIT design

In this section, we focus on the design of one-sample FIT. We specifically study dichotomous test design and explore the options of universal test, customized test and interpretable clustering test.

Universal dichotomous test

We consider a population base of 10,000 individuals, of which 14.23 individuals have CRC on expectation based on the CRC incidence rate. We obtain the optimal universal dichotomous test for various values of adjusting term, τ, from -1 to 1. We present the optimal cut-off point and other related performance metrics in Table 4.2.

Firstly, we observe that when the adjusting term τ increases from -1 to 1, the optimal cut-off value increases, and the corresponding sensitivity (specificity) decreases (increases). This is an intuitive result that if the follow-up penalty for healthy individuals is high, the health system should reduce the false-positive rate, which can be accomplished by increasing the cut-off value.

Secondly, given that sensitivity (specificity) decreases (increases) as the cut-off

Adjusting term (τ)	-1	-0.05	-0.02	-0.005	0	0.005	0.01	0.05	1
The optimal cut-off point, $\mu g/g$	10	13	27	31	33	35	38	50	75
Sensitivity, %	87.48	85.55	78.64	77.06	76.30	75.58	74.53	69.38	63.38
Specificity, %	84.15	89.88	98.50	99.10	99.30	99.45	99.62	99.95	100.00
Expected number of positives *	1594.98	1023.09	160.78	100.76	80.72	65.26	48.33	14.95	9.34
Expected number of positives from CRC patients*	12.45	12.18	11.19	10.97	10.86	10.76	10.61	9.87	9.02
Expected number of positives from healthy individuals*	1582.53	1010.91	149.59	89.79	69.86	54.50	37.72	5.08	0.32
Expected colonoscopy demand*	344.47	340.63	136.78	95.47	78.59	64.43	48.10	14.95	9.34
Expected number of follow-ups from CRC patients*	3.18	4.48	10.14	10.68	10.73	10.71	10.60	9.87	9.02
Expected number of follow-ups from healthy individuals*	341.29	336.15	126.64	84.79	67.86	53.72	37.50	5.08	0.32

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.

Table 4.2: The optimal universal dichotomous test (one-sample FIT)

point value becomes larger, the expected number of positive outcomes from both individuals with and without CRC decline. This trend implies that under the assumption of a constant adherence (i.e., a fixed proportion of the individuals receiving positive outcomes will follow up and demand a colonoscopy), higher test effectiveness, which, in this case, corresponds to more positive outcomes from CRC individuals, is always accompanied by a higher demand for colonoscopies. Each cut-off point would correspond to a FIT kit on the efficient frontier. On the contrary, by considering the imperfect test-dependent adherence behavior, the expected number of the actual follow-ups from CRC individuals is not monotone. This observation is an outcome of two opposing effects when cut-off point values change. On the one hand, FIT's sensitivity decreases when the cut-off point value increases, and therefore, fewer CRC patients are detected. On the other hand, as the cut-off value increases, individuals are more likely to follow up after receiving a positive outcome due to higher posterior CRC risk and belief. In particular, we find that when the optimal cut-off point value is less than 33 $\mu g/g$, the impact on follow-up probability dominates. The FIT kits in this domain (i.e., high-sensitivity FIT kits) are no longer on the efficient frontier. A FIT with a larger cut-off point value can detect more CRC cases with fewer colonoscopies. When the optimal cut-off point value is higher or equal to 33 $\mu g/g$, the impact on detection rate dominates, and the trade-off between test effectiveness and colonoscopy demand appears. Our results shed light on the criticality of considering the strategic imperfect adherence behavior. The wrong assumption on constant adherence rate would recommend a FIT kit with high sensitivity (e.g., cut-off point is 10 $\mu g/g$) when colonoscopy capacity is relatively sufficient. This FIT kit will essentially fail in practice, not only by creating excessive unnecessary colonoscopy demand but also by causing much fewer CRC cases being detected.

Customized dichotomous test

Customized FIT would provide additional flexibility to the health system in improving the effectiveness of CRC screening. While adopting a large number of different types of test kits is not practical, we explore the benefit of promoting two types of the dichotomous FITs to two subpopulations (i.e., $L = 2$ in Problem (4.7)). The optimal design of three-type case (i.e., $L = 3$) is also obtained with details given in Appendix C.5.1. Table 4.3 presents the optimal two-type customized dichotomous test with the adjusting term τ equals 0.006[3].

	Cluster 1	Cluster 2	Total	Cluster 1$^{\&}$	Cluster 2$^{\&}$	Total$^{\&}$	Current practice
Number of individuals*	4022.48	5977.52	10000	4022.48	5977.52	10000	10000
Expected number of CRC patients*	9.18	5.05	14.23	9.18	5.05	14.23	14.23
The optimal cut-off point, $\mu g/g$	31	39	-	-	-	33	20
Sensitivity, %	77.06	74.20	-	-	-	76.30	81.77
Specificity, %	99.10	99.67	-	-	-	99.30	96.22
Expected number of positives*	43.17	23.72	66.89	35.09	45.63	80.72	389.37
Expected number of positives from CRC patients*	7.08	3.74	10.82	7.01	3.85	10.86	11.64
Expected number of positives from healthy individuals*	36.09	19.98	56.07	28.08	41.78	69.86	377.73
Expected colonoscopy demand*	43.12	23.56	66.68	35.09	43.50	78.59	227.01
Expected number of follow-ups from CRC patients*	7.07	3.74	10.81	7.01	3.72	10.73	7.21
Expected number of follow-ups from healthy individuals*	36.05	19.82	55.87	28.08	39.78	67.86	219.80

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.
$^{\&}$ The results are from the universal test when the adjusting term τ equals 0.

Table 4.3: The optimal customized dichotomous test when $\tau = 0.006$ and $L = 2$ (one-sample FIT)

It is worthwhile to examine the population features of the obtained two clusters. We present the key feature variables that differentiate the two clusters in Appendix (Table C.6). Overall, the sub-population in Cluster 1 has more CRC individuals and more elderly males who are of higher risks to develop CRC (National Registry of

[3] We present the result of $\tau = 0.006$ to show a case of a high follow-up rate from the CRC patients with a low total colonoscopy demand. The performance of the test designs under different adjusting terms is presented as efficient frontiers in the latter part of the numerical discussion (Figure 4.5).

Diseases Office, 2015) compared with Cluster 2. This would help explain the optimal test design for the two clusters.

We compare the optimal customized tests to the optimal universal test to show the benefit of customization[4]. To facilitate the comparison, for the optimal universal test with $\tau = 0$, we further get the expected number of positive outcomes and colonoscopy demands for the same two clusters (columns with superscript $^{\&}$ in Table 4.3).

We first observe that the optimal customized test design suggests a higher sensitivity FIT for Cluster 1 (adjusting the cut-off from 33 $\mu g/g$ to 31 $\mu g/g$). An increase in the FIT sensitivity will enhance the test effectiveness for this cluster that contains more CRC individuals. In addition, due to the higher risk of developing CRC, the sub-population in this cluster have a relatively higher follow-up probability after receiving positive test outcomes. This inherent high follow-up probability leaves room for strategically choosing a cut-off point such that the increase in the sensitivity only slightly reduces the adherence level (7.07 out of 7.08 v.s. 7.01 out of 7.01 in the universal test). The overall effect from a higher cut-off point benefits the test effectiveness for this cluster.

Secondly, for Cluster 2 that contains more relatively lower risk individuals, a lower sensitivity FIT is suggested compared with the optimal universal FIT (altering the cut-off point from 33 $\mu g/g$ to 39 $\mu g/g$). Interestingly, we not only reduce the demand for colonoscopies but also enhances the screening effectiveness in terms of detecting more CRC individuals. Despite the fact that we raise fewer positive cases for CRC individuals in Cluster 2 under the customized test due to low sensitivity (3.74 v.s. 3.85 in the universal test), the resulting lower false positive rate increases the adherence rate (from 3.72 out of 3.85 in the universal test to 3.74 out of 3.74) for CRC individuals. Overall, we have more follow-ups from CRC individuals (3.74 v.s. 3.72 in the universal test).

[4] we compare with the universal test with $\tau = 0$ as this case generates the highest follow-up rate from CRC patients

Interpretable clustering dichotomous test

We further apply the interpretable clustering method to design two customized tests based on the clusters generated from the customized test design result. Specifically, the demographic data used to train the decision tree model to predict the individuals' clusters include age, gender and marital status that is easily accessible to the health system. The decision tree obtained is shown in Figure 4.4 and the average AUC of 5-fold cross-validation is 0.99, where age and gender are two important splitting variables.

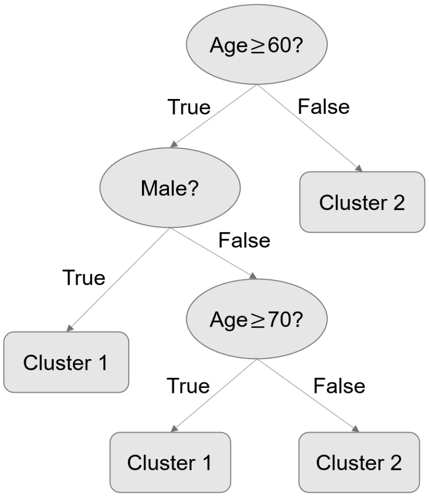


Figure 4.4: Decision tree (one sample)

We note that individuals who are male (female) and older than 60 (70) years old are classified in Cluster 1, and the remaining individuals are included in Cluster 2. As the elderly and males have higher risks of developing CRC, Cluster 1 contains high risk individuals. We further observe that 94.04% (3782.91 of 4022.48) individuals who originally belonged to Cluster 1 in the customized test design result remain in Cluster 1, and 100% (5977.52 of 5977.52) individuals who previously were part of Cluster 2 remain in Cluster 2. The decision tree based clustering which is purely based on age and gender can recover the optimal partition in the customized test design.

For each newly generated cluster, we solve the optimal cut-off value and report the expected number of positive results and follow-ups in the two clusters in Table C.8. The optimal cut-off values in the two clusters are 31 $\mu g/g$ and 39 $\mu g/g$, which are exactly the same as the ones obtained from the customized test design.

The interpretable clustering initial test design is able to propose an implementable customized screening policy that only depends on age and gender. The partition rule of the population is not only interpretable but also aligns with the CRC risk profile that is well understood by healthcare practitioners. By simply partition the population into two groups, a substantial improvement of the screening effectiveness and efficiency can be materialized.

Two-sample FIT design

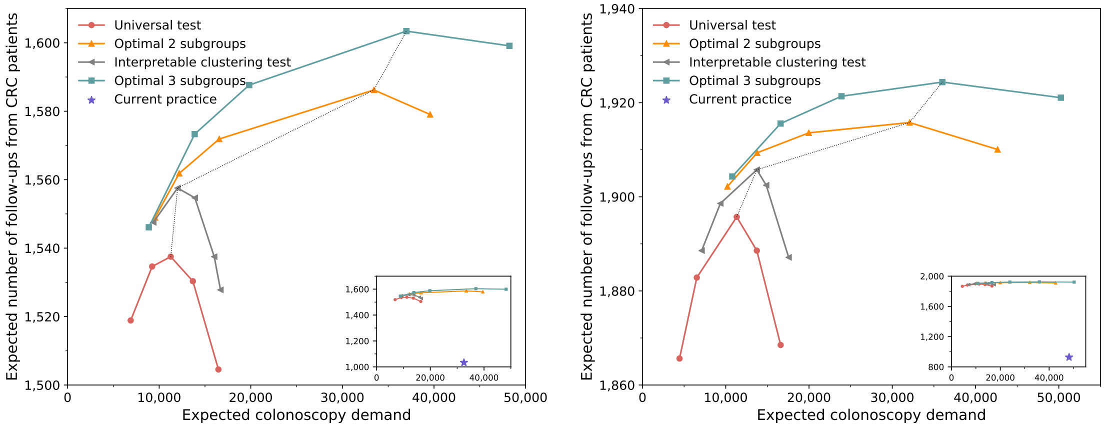
We further explore the optimal initial test design if the health system utilizes the two-sample FIT in the screening program. Singapore has adopted two-sample FIT in the national CRC screening programme, that a positive result will be reported once the hemoglobin concentration in any of the two samples exceeds the pre-determined cut-off point. We assume independence between the two FIT kits as they are collected on two different days. Under the independence assumption, the aggregate level of two-sample FIT's sensitivity and specificity can be calculated. Note that we can also incorporate any correlation between two FIT kits if relevant data is available. We also verify that two-sample dichotomous FIT initial tests are MLR-feasible if each test kit is MLR-feasible (cf. Appendix C.4.4).

Similar to one-sample FIT design, optimal cut-offs are obtained for the universal dichotomous test, customized dichotomous test, and interpretable clustering test. The optimal cut-offs are much higher than those of one-sample FIT because two-sample test induces higher overall test sensitivity and lower specificity. The main findings and insights are similar to those in one-sample FIT. For the universal test, a FIT with high sensitivity is again not desirable. By considering two heterogeneous FIT kits, one high risk group and one low risk group are identified. Furthermore, the interpretable clustering method shows that by simply partition the population by an age threshold of 60, we can recover the two clusters generated from optimal customized test design. We relegate the detailed results in Appendix C.5.2.

Practical implication

Currently, Singapore adopts a FIT test kit with the cut-off point value of 20 $\mu g/g$. For ease of expression, we refer to the choice of 20 $\mu g/g$ as the current practice.

To compare the our optimal designs with the current practice for both one-sample and two-sample cases, we plot the efficient frontiers of each optimal design in Figure 4.5 (one point on each line represents a design for a particular adjustment term, τ). As already discussed in previous sections, the customized tests and the interpretable clustering tests are superior to the universal test. The customized test with three heterogeneous test kits is superior to the customized test with two test kits, but the benefit is modest. Should a universal test be adopted, a test with a well-chosen cut-off point significantly dominates the current practice and achieves higher screening effectiveness with fewer demand for the second-stage colonoscopy.



(a) One-sample FIT

(b) Two-sample FIT

Figure 4.5: Trade-off between test effectiveness and test efficiency
The grey dotted line connects the cut-off points that generate the highest detection rates under different screening designs, and all strategies that lie on the right-hand side of this line are not on the efficient frontiers.

To further demonstrate the benefit of optimal designs, we present the test performance of one point on the efficient frontier for each design in Table 4.4, and compare them with the current practice. The values presented in the table correspond to a population base equals to the actual Singapore population.

	One-sample FIT				Two sample-FIT			
	Current practice	Universal test ($\tau = 0$)	Customized test ($\tau = 0.006$, $L = 2$)	Interpretable test ($\tau = 0.007$)	Current practice	Universal test ($\tau = 0$)	Customized test ($\tau = 0.0035$, $L = 2$)	Interpretable test ($\tau = 0.0035$)
Cut-off point	20	33	31,39	31,39	20	39	37, 45	37, 45
Sensitivity , %	81.77	76.30	77.06, 74.20	77.06, 74.20	96.68	93.34	93.69, 92.31	93.69, 92.31
Specifiity , %	96.22	99.30	99.10, 99.67	99.10, 99.67	92.58	99.33	99.15, 99.67	99.15, 99.67
Expected number of positives*	55,793	11,566	9,584	9,388	108,174	11,459	9,586	9,406
Expected number of positives from CRC patients*	1,668	1,556	1,550	1,549	1,972	1,903	1,901	1,900
Expected number of positives from healthy individuals*	54,125	10,010	8,034	7,839	106,202	9,556	7,685	7,506
Expected colonoscopy demand*	32,528	11,261	9,555	9,359	48,170	11,319	9,569	9,387
Expected number of follow-ups from CRC patients*	1,033	1,537	1,549	1,547	927	1,896	1,900	1,899
Expected number of follow-ups from healthy individuals*	31,495	9,724	8,006	7,812	47,243	9,423	7,669	7,488
Total cost of colonoscopies for citizens/PRs*, million	35.98	12.45	10.57	10.35	53.28	12.52	10.58	10.38
Total cost of colonoscopies for the government*, million	59.95	20.75	17.61	17.25	88.78	20.86	17.64	17.30

* The population base of Singapore is 1,432,897, of which 2,039 individuals have CRC and 1,430,858 do not.

Table 4.4: Performance comparision of screening designs

Regarding the one-sample FIT, our optimal initial test design result suggests an optimal universal test that alters the cut-off point from 20 $\mu g/g$ (current practice) to 33 $\mu g/g$. This optimal test is able to identify 504 more CRC incidences and, at the same time, reduce 21,267 colonoscopy demand from the whole population[5]. In particular, one colonoscopy costs the government an average of S$1,843 in terms of subsidizing the procedure, and an individual an average of S$1,106 (Ministry of Health Singapore (2020a)). The reduction in the colonoscopy demand could help the Singapore government reduce healthcare expenditure by S$39.20 million and also save Singapore citizens/permanent residents' spending by S$23.53 million. In addition, heterogeneous test kits tailored to different subgroups of populations could help the health system detect 516 more CRC incidences and simultaneously reduce 22,973

[5] According to Statistics Singapore (2018), there are 1,432,897 individuals aged 50 and above in Singapore. We obtained the two numbers by rescaling the results in Table 4.2 which assumes a 10,000 population base.

colonoscopies from the Singapore population, which leads to a reduction in healthcare expenditure by S$42.34 million and a saving in Singaporean/PR expenses by S$25.41 million. The interpretable clustering test could help to detect 514 more CRC incidences and reduce 23,168 colonoscopy demand, which amounts to a reduction of government healthcare expenditure by S$42.70 million and a saving of healthcare spending by S$25.63 million for citizens/PRs.

For the two-sample FIT, the improvement over the current practice with two-sample FIT is also significant. We omit the details here. In addition, the two sample FIT optimal design outperforms one-sample FIT in terms of test efficiency and test effectiveness under specific adjusting terms.

4.6 Conclusion

In this chapter, we focus on a two-stage CRC screening program and develop an optimization framework to design the initial test that balance the screening effectiveness and efficiency considering individuals' adherence behavior. The optimal design results suggest that a well designed initial test would be able to detect more CRC cases with fewer colonoscopies. Besides, we also show that adopting customized dichotomous tests or interpretable clustering dichotomous tests for different subpopulations could provide additional benefits.

Several interesting future research topics are worth to mention. Our work does not consider the first-stage adherence problem that not all the individuals will take up the intital test as suggested by the screening guideline. In practice, both the initial FIT take rate and the repeat screening rate among FIT-negative patients are far below desired. For instance, in one observational study, Nielson et al. (2019) show that the proportion of FIT tests completed was 46% in the patients' first year and 41% in the patients' second year. One interesting avenue to further extend our research is to incorporate individuals' first-stage decisions into the optimal test design. Moreover, it would be interesting to develop a multi-period model to investigate the optimal cut-off point considering the repeating screening behavior among FIT-negative individuals. Moreover, we have discussed with healthcare professionals regarding the application of our results in real practice. With possible integration of our suvery data and clinical data, we can future incorporate the post-colonoscopy CRC treatment into the study.

Chapter 5

Conclusion

Promoting effective and sustainable healthcare service delivery has become critical due to the increasing trend of the aging population, the rising incidence of various diseases, and the surge in health care expenditures. This dissertation integrates a combination of techniques from machine learning, optimization, game theory and survey design to improve the real-time patient monitoring systems and population screening programs.

Chapter 2 proposes a new framework to estimate real-time values of risk monitoring scores and uncertainties in risk assessment. We demonstrate that integrating predictive information into existing scoring systems can significantly improve the prognostic accuracy and discriminative ability on various predictive outcomes (e.g., 24-hour mortality, 30-day readmission). The proposed approach provides the basis for more detailed patient risk classification and decision recommendations. Moreover, the estimates of uncertainty in patient health status can be used to trigger evidence-based on-demand laboratory testing. It's worth noting that our results are based on the analysis of data from one medical center. Although the method is sufficiently generic to improve any risk scores, the improvement may differ in other units. In addition, external validation and clinical tries are warranted to confirm the benefits of enhanced risk monitoring systems. What's more, with the advancement of natural language processing, image recognition, video analysis and deep learning techniques, we can collect more data from different sources (e.g., electronic health records, spontaneous reporting databases, mobilized health records) in different formats (e.g., text, audio, video) to improve risk monitoring of patients. We leave these for future research.

Chapter 3 embeds the predictive model proposed in Chapter 2 to characterize the optimal prescription of diagnostic tests in the detection of acute diseases. We theoretically demonstrate that considering uncertainty in risk measurement can contribute to lower expected costs for patients. Utilizing the data from a medical center, we show that the proposed optimal strategy is able to advance the detection of diseases with fewer tests in an ICU context. Several interesting future research topics are worth mentioning. First, we can theoretically explore the specific value of predictive information and uncertainty measures. Second, we can consider the accuracy of diagnostic tests and incorporate it into our framework. Moreover, since reducing overuse of healthcare resources is a complicated and crucial issue, it is worthwhile to combine medical practices, machine learning techniques and optimization model to further explore such problems and validate the effectiveness of the proposed frameworks using more comprehensive healthcare data.

Chapter 4 focuses on a two-stage CRC screening program and further considers individual behavioral factors. We establish an optimization framework with information avoidance to optimize the initial test design, with the aim of balancing the trade-off between effectiveness and efficiency of the second-stage screening tests. We show that the proposed initial test design is able to detect more CRC cases with fewer second-stage screening tests. In addition, we show that the employment of customized dichotomous tests or interpretable clustered dichotomous tests for different subpopulations can provide additional benefits. This chapter does not consider individuals' adherence to the first-stage tests. Besides, it's also interesting to develop a multi-period model to investigate the repeating screening behavior among individuals receiving negative results from the initial tests. Furthermore, it is also critical to consider human behavioral factors in other health screening programs and medications (e.g., insulin therapy for diabetes, antiretroviral therapy for people with HIV). We can further investigate these potential research topics.

To summarize, by analyzing comprehensive datasets collected from multiple sources, this dissertation demonstrates that well-designed monitoring systems and screening programs can benefit individuals, health care providers, and health systems by improving the effectiveness and efficiency of healthcare delivery. Using more healthcare data from diverse sources and implementing more advanced technologies in information systems and operations management, we can further explore these directions and come up with more evidence-based frameworks to improve healthcare delivery in future studies.

Bibliography

Adhikari, N. K., Fowler, R. A., Bhagwanjee, S., and Rubenfeld, G. D. (2010). Critical care and the global burden of critical illness in adults. *The Lancet*, 376(9749):1339–1346.

Alagoz, O. (2011). Optimizing cancer screening using partially observable markov decision processes. In *Transforming Research into Action*, pages 75–89. INFORMS.

Alagoz, O., Ayer, T., and Erenay, F. S. (2010). Operations research models for cancer screening. *Wiley encyclopedia of operations research and management science*.

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P., McGinn, T., and Guyatt, G. (2017). Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *Jama*, 318(14):1377–1384.

Allison, J., Fraser, C., and Young, G. (2014a). Chicago 2014 5th meeting of the expert working group (ewg)–'fit for screening'. `https://www.worldendo.org/wp-content/uploads/2016/08/weo_expert_working_group_fit_meeting_report_chicago2014.pdf`.

Allison, J. E., Fraser, C. G., Halloran, S. P., and Young, G. P. (2014b). Population screening for colorectal cancer means getting fit: the past, present, and future of colorectal cancer screening using the fecal immunochemical test for hemoglobin (fit). *Gut and liver*, 8(2):117.

American Cancer Society (2020a). Colorectal cancer facts & figures. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2020-2022.pdf.

American Cancer Society (2020b). Colorectal cancer screening tests. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/screening-tests-used.

Are, C., Bartlett, D. L., Nissan, A., Dov, Z., Gupta, A., Savant, D., Bargallo-Rocha, J. E., Said, H. M., Oliveira, A. F., de Castro Ribeiro, H. S., et al. (2020). Global forum of cancer surgeons: Position statement to promote cancer surgery globally. *Annals of Surgical Oncology*, 27:2573–2576.

Aşuroğlu, T. and Oğul, H. (2021). A deep learning approach for sepsis monitoring via severity score estimation. *Computer Methods and Programs in Biomedicine*, 198:105816.

Aubin, J.-P. (2007). *Mathematical methods of game and economic theory*. Courier Corporation.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Ayer, T., Alagoz, O., Stout, N. K., and Burnside, E. S. (2016). Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Science*, 62(5):1339–1362.

Ayvaci, M. U., Alagoz, O., and Burnside, E. S. (2012). The effect of budgetary restrictions on breast cancer diagnostic decisions. *Manufacturing & Service Operations Management*, 14(4):600–617.

Aziz, K., Tarapiah, S., Ismail, S. H., and Atalla, S. (2016). Smart real-time healthcare monitoring and tracking system using gsm/gps technologies. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pages 1–7. IEEE.

Baig, M. M. and Gholamhosseini, H. (2013). Smart health monitoring systems: an overview of design and modeling. *Journal of medical systems*, 37(2):9898.

Barsheshet, A., Garty, M., Grossman, E., Sandach, A., Lewis, B. S., Gottlieb, S., Shotan, A., Behar, S., Caspi, A., Schwartz, R., et al. (2006). Admission blood glucose level and mortality among hospitalized nondiabetic patients with heart failure. *Archives of internal medicine*, 166(15):1613–1619.

Bergemann, D., Brooks, B., and Morris, S. (2015). The limits of price discrimination. *American Economic Review*, 105(3):921–57.

Bergemann, D., Brooks, B., and Morris, S. (2017). First-price auctions with general information structures: Implications for bidding and revenue. *Econometrica*, 85(1):107–143.

Beriault, D. R., Gilmour, J. A., and Hicks, L. K. (2021). Overutilization in laboratory medicine: tackling the problem with quality improvement science. *Critical Reviews in Clinical Laboratory Sciences*, pages 1–24.

Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D. F., and Feuer, E. J. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine*, 353(17):1784–1792.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Bertsimas, D., O'Hair, A., Relyea, S., and Silberholz, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science*, 62(5):1511–1531.

Bindraban, R. S., Ten Berg, M. J., Naaktgeboren, C. A., Kramer, M. H., Van Solinge, W. W., and Nanayakkara, P. W. (2018). Reducing test utilization in hospital settings: a narrative review. *Annals of laboratory medicine*, 38(5):402–412.

Boer, R., de Koning, H., Threlfall, A., Warmerdam, P., Street, A., Friedman, E., and Woodman, C. (1998). Cost effectiveness of shortening screening interval or extending age range of nhs breast screening programme: computer simulation study. *Bmj*, 317(7155):376–379.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353.

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.

Bynum, S. A., Davis, J. L., Green, B. L., and Katz, R. V. (2012). Unwillingness to participate in colorectal cancer screening: examining fears, attitudes, and medical mistrust in an ethnically diverse sample of adults 50 years and older. *American journal of health promotion*, 26(5):295–300.

Capes, S. E., Hunt, D., Malmberg, K., and Gerstein, H. C. (2000). Stress hyperglycaemia and increased risk of death after myocardial infarction in patients with and without diabetes: a systematic overview. *The Lancet*, 355(9206):773–778.

Caplin, A. and Eliaz, K. (2003). Aids policy and psychology: A mechanism-design approach. *RAND Journal of Economics*, pages 631–646.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, 116(1):55–79.

Caso, R., Fabrizio, A., and Sosin, M. (2020). Prolonged follow-up of colorectal cancer patients after 5 years: to follow or not to follow, that is the question (and how)! *Annals of translational medicine*, 8(5).

Cassandra, A., Littman, M. L., and Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 54–61. Morgan Kaufmann Publishers Inc.

Cevik, M., Ayer, T., Alagoz, O., and Sprague, B. L. (2018). Analysis of mammography screening policies under resource constraints. *Production and Operations Management*, 27(5):949–972.

Chalmers, J. D., Singanayagam, A., and Hill, A. T. (2008). Systolic blood pressure is superior to other haemodynamic predictors of outcome in community acquired pneumonia. *Thorax*, 63(8):698–702.

Chen, L.-S., Liao, C.-S., Chang, S.-H., Lai, H.-C., and Chen, T. H.-H. (2007). Cost-effectiveness analysis for determining optimal cut-off of immunochemical faecal occult blood test for population-based colorectal cancer screening (kcis 16). *Journal of medical screening*, 14(4):191–199.

Cheng, L.-F., Prasad, N., and Engelhardt, B. E. (2019). An optimal policy for patient laboratory tests in intensive care units. In *PSB*, pages 320–331. World Scientific.

Chew, M., Suzanah, N., Ho, K., Lim, J., Ooi, B., Tang, C., Eu, K., et al. (2009). Colorectal cancer mass screening event utilising quantitative faecal occult blood test. *Singapore Med J*, 50(4):348–353.

Chhatwal, J., Alagoz, O., and Burnside, E. S. (2010). Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations research*, 58(6):1577–1591.

Cismondi, F., Celi, L. A., Fialho, A. S., Vieira, S. M., Reti, S. R., Sousa, J. M., and Finkelstein, S. N. (2013). Reducing unnecessary lab testing in the icu with artificial intelligence. *International journal of medical informatics*, 82(5):345–358.

Clouzeau, B., Caujolle, M., San-Miguel, A., Pillot, J., Gazeau, N., Tacaille, C., Dousset, V., Bazin, F., Vargas, F., Hilbert, G., et al. (2019). The sustainable impact of an educational approach to improve the appropriateness of laboratory test orders in the icu. *PloS one*, 14(5):e0214802.

Deeks, J. J. and Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *Bmj*, 329(7458):168–169.

Department of Statistics (2019). Complete life tables for singapore resident population, 2017-2018. `https://www.singstat.gov.sg/publications/population/complete-life-table`.

Desautels, T., Das, R., Calvert, J., Trivedi, M., Summers, C., Wales, D. J., and Ercole, A. (2017). Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: a cross-sectional machine learning approach. *BMJ open*, 7(9).

Dhanani, J., Barnett, A., Lipman, J., and Reade, M. (2018). Strategies to reduce inappropriate laboratory blood test orders in intensive care are effective and safe: a before-and-after quality improvement study. *Anaesthesia and intensive care*, 46(3):313–320.

Eaton, K. P., Levy, K., Soong, C., Pahwa, A. K., Petrilli, C., Ziemba, J. B., Cho, H. J., Alban, R., Blanck, J. F., and Parsons, A. S. (2017). Evidence-based guidelines to eliminate repetitive laboratory testing. *JAMA internal medicine*, 177(12):1833–1839.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (2011). *Economic and financial decisions under risk*. Princeton University Press.

Eliaz, K. and Spiegler, R. (2006). Can anticipatory feelings explain anomalous choices of information sources? *Games and Economic Behavior*, 56(1):87–104.

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

Feldstein, C. and Weder, A. B. (2012). Orthostatic hypotension: a common, serious and underrecognized problem in hospitalized patients. *Journal of the American Society of Hypertension*, 6(1):27–39.

Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C., and Vincent, J.-L. (2001). Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758.

Finkelsztein, E. J., Jones, D. S., Ma, K. C., Pabón, M. A., Delgado, T., Nakahira, K., Arbo, J. E., Berlin, D. A., Schenck, E. J., Choi, A. M., et al. (2017). Comparison of qsofa and sirs for predicting adverse outcomes of patients with suspicion of sepsis outside the intensive care unit. *Critical care*, 21(1):1–10.

Forsman, R. W. (1996). Why is the diagnostic an afterthought for managed care organizations? *Clinical Chemistry*, 42(5):813–816.

Fraser, C. G. (2011). Screening for colorectal neoplasia with faecal tests. *The lancet oncology*, 6(12):516–517.

Fraser, C. G., Allison, J. E., Halloran, S. P., Young, G. P., and Expert Working Group on Fecal Immunochemical Tests for Hemoglobin, Colorectal Cancer Screening Committee, W. E. O. (2012). A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. *Journal of the National Cancer Institute*, 104(11):810–814.

Freund, Y., Lemachatti, N., Krastinova, E., Van Laer, M., Claessens, Y.-E., Avondo, A., Occelli, C., Feral-Pierssens, A.-L., Truchot, J., Ortega, M., et al. (2017). Prognostic accuracy of sepsis-3 criteria for in-hospital mortality among patients with suspected infection presenting to the emergency department. *Jama*, 317(3):301–308.

Ganguly, A. and Tasoff, J. (2017). Fantasy and dread: The demand for information and the consumption utility of the future. *Management Science*, 63(12):4037–4060.

Gentzkow, M. and Kamenica, E. (2016). A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review*, 106(5):597–601.

Gentzkow, M. and Kamenica, E. (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429.

Gies, A., Bhardwaj, M., Stock, C., Schrotz-King, P., and Brenner, H. (2018). Quantitative fecal immunochemical tests for colorectal cancer screening. *International journal of cancer*, 143(2):234–244.

Gimeno Garcia, A. Z. (2012). Factors influencing colorectal cancer screening participation. *Gastroenterology research and practice*, 2012.

Goede, S. L., van Roon, A. H., Reijerink, J. C., van Vuuren, A. J., Lansdorp-Vogelaar, I., Habbema, J. D. F., Kuipers, E. J., van Leerdam, M. E., and van Ballegooijen, M. (2013). Cost-effectiveness of one versus two sample faecal immunochemical testing for colorectal cancer screening. *Gut*, 62(5):727–734.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Golman, R. and Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, 5(3):143.

Golman, R., Loewenstein, G., Molnar, A., and Saccardo, S. (2019). The demand for, and avoidance of, information. *Information (May 31, 2019)*.

Grand View Research (2021). Clinical laboratory service market size, share and trends analysis report by test type (human and tumor genetics, clinical chemistry, medical microbiology and cytology), by service provider, by application, by region, and segment forecasts, 2021 - 2028. `https://www.grandviewresearch.com/industry-analysis/clinical-laboratory-services-market/methodology`. accessed April 23, 2021.

Grazzini, G., Visioli, C., Zorzi, M., Ciatto, S., Banovich, F., Bonanomi, A., Bortoli, A., Castiglione, G., Cazzola, L., Confortini, M., et al. (2009). Immunochemical faecal occult blood test: number of samples and positivity cutoff. what is the best strategy for colorectal cancer screening? *British journal of cancer*, 100(2):259–265.

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

Hall, M. J., Williams, S. N., DeFrances, C. J., and Golosinskiy, A. (2011). Inpatient care for septicemia or sepsis: a challenge for patients and hospitals.

Hernandez, V., Cubiella, J., Gonzalez-Mao, M. C., Iglesias, F., Rivera, C., Iglesias, M. B., Cid, L., Castro, I., de Castro, L., Vega, P., et al. (2014). Fecal immunochemical test accuracy in average-risk colorectal cancer screening. *World Journal of Gastroenterology: WJG*, 20(4):1038.

Hernández-Peña, P. (2019). Global spending on health: A world in transition [global report 2019]. *WHO/HIS/HGF/HFWorkingPaper/*, (19.4).

Hol, L., Wilschut, J., van Ballegooijen, M., Van Vuuren, A., van der Valk, H., Reijerink, J., van Der Togt, A., Kuipers, E., Habbema, J., and Van Leerdam, M. (2009). Screening for colorectal cancer: random comparison of guaiac and immunochemical faecal occult blood testing at different cut-off levels. *British journal of cancer*, 100(7):1103–1110.

Holder, A. L., Overton, E., Lyu, P., Kempker, J. A., Nemati, S., Razmi, F., Martin, G. S., Buchman, T. G., and Murphy, D. J. (2017). Serial daily organ failure assessment beyond icu day 5 does not independently add precision to icu risk-of-death prediction. *Critical care medicine*, 45(12):2014.

Howlader, N., Noone, A., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., and Cronin, K. e. (2018). Seer cancer statistics review, 1975-2016, national cancer institute. `https://seer.cancer.gov/csr/1975_2016/`.

Hubers, J., Sonnenberg, A., Gopal, D., Weiss, J., Holobyn, T., and Soni, A. (2020). Trends in wait time for colorectal cancer screening and diagnosis 2013-2016. *Clinical and Translational Gastroenterology*, 11(1).

Hwang, S. Y., Jo, I. J., Lee, S. U., Lee, T. R., Yoon, H., Cha, W. C., Sim, M. S., and Shin, T. G. (2018). Low accuracy of positive qsofa criteria for predicting 28-day mortality in critically ill septic patients during the early period after emergency department presentation. *Annals of emergency medicine*, 71(1):1–9.

Ichai, C. and Preiser, J.-C. (2010). International recommendations for glucose control in adult non diabetic critically ill patients. *Critical care*, 14(5):R166.

Itoh, M., Takahashi, K., Nishida, H., Sakagami, K., and Okubo, T. (1996). Estimation of the optimal cut off point in a new immunological faecal occult blood test in a corporate colorectal cancer screening programme. *Journal of medical screening*, 3(2):66–71.

Iturrate, E., Jubelt, L., Volpicelli, F., and Hochman, K. (2016). Optimize your electronic medical record to increase value: reducing laboratory overutilization. *The American journal of medicine*, 129(2):215–220.

Jen, H.-H., Hsu, C.-Y., Chen, S. L.-S., Yen, A. M.-F., Chiu, S. Y.-H., Fann, J. C.-Y., Lee, Y.-C., Wu, M.-S., Hsu, W.-F., Peng, S.-M., et al. (2018). Rolling-out screening volume affecting compliance rate and waiting time of fit-based colonoscopy. *Journal of Clinical Gastroenterology*, 52(9):821–827.

Jenkins, D. A., Sperrin, M., Martin, G. P., and Peek, N. (2018). Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and prognostic research*, 2(1):23.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and uncertainty*, 38(2):95–115.

Keehan, S. P., Cuckler, G. A., Sisko, A. M., Madison, A. J., Smith, S. D., Stone, D. A., Poisal, J. A., Wolfe, C. J., and Lizonitz, J. M. (2015). National health expenditure projections, 2014–24: spending growth faster than recent trends. *Health Affairs*, 34(8):1407–1417.

Khalid-de Bakker, C. A., Jonkers, D. M., Sanduleanu, S., de Bruïne, A. P., Meijer, G. A., Janssen, J. B., van Engeland, M., Stockbrügger, R. W., and Masclee,

A. A. (2011). Test performance of immunologic fecal occult blood testing and sigmoidoscopy compared with primary colonoscopy screening for colorectal advanced adenomas. *Cancer Prevention Research*, 4(10):1563–1571.

Kolata, G. (2003). 50 and ready for colonoscopy? doctors say wait is often long. *The New York Times*.

Koleva-Kolarova, R. G., Zhan, Z., Greuter, M. J., Feenstra, T. L., and De Bock, G. H. (2015). Simulation models in population breast cancer screening: a systematic review. *The Breast*, 24(4):354–363.

Kőszegi, B. (2003). Health anxiety and patient behavior. *Journal of health economics*, 22(6):1073–1084.

Krishnamurthy, V. (2016). *Partially Observed Markov Decision Processes*. Cambridge University Press.

Kruse, G. R., Khan, S. M., Zaslavsky, A. M., Ayanian, J. Z., and Sequist, T. D. (2015). Overuse of colonoscopy for colorectal cancer screening and surveillance. *Journal of general internal medicine*, 30(3):277–283.

Kumwilaisak, K., Noto, A., Schmidt, U. H., Beck, C. I., Crimi, C., Lewandrowski, K., and Bigatello, L. M. (2008). Effect of laboratory testing guidelines on the utilization of tests and order entries in a surgical intensive care unit. *Critical care medicine*, 36(11):2993–2999.

Ladabaum, U., Chopra, C. L., Huang, G., Scheiman, J. M., Chernew, M. E., and Fendrick, A. M. (2001). Aspirin as an adjunct to screening for prevention of sporadic colorectal cancer: a cost-effectiveness analysis. *Annals of internal medicine*, 135(9):769–781.

Ladabaum, U. and Mannalithara, A. (2016). Comparative effectiveness and cost effectiveness of a multitarget stool dna test to screen for colorectal neoplasia. *Gastroenterology*, 151(3):427–439.

Langley, J. and Adams, G. (2007). Insulin-based regimens decrease mortality rates in critically ill patients: a systematic review. *Diabetes/metabolism research and reviews*, 23(3):184–192.

Le Maguet, P., Asehnoune, K., Autet, L.-M., Gaillard, T., Lasocki, S., Mimoz, O., Demeure Dit Latte, D., Gergaud, S., Morcet, J., Seguin, P., et al. (2015). Transitioning from routine to on-demand test ordering in intensive care units: a prospective, multicentre, interventional study. *BJA: British Journal of Anaesthesia*, 115(6):941–942.

Lee, H. L. and Pierskalla, W. P. (1988). Mass screening models for contagious diseases with no latent period. *Operations research*, 36(6):917–928.

Lee, J. K., Liles, E. G., Bent, S., Levin, T. R., and Corley, D. A. (2014). Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Annals of internal medicine*, 160(3):171–181.

Lee, S.-Y. and Lee, E. E. (2018). Cancer screening in koreans: a focus group approach. *BMC public health*, 18(1):254.

Leening, M. J., Vedder, M. M., Witteman, J. C., Pencina, M. J., and Steyerberg, E. W. (2014). Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Annals of internal medicine*, 160(2):122–131.

Li, Y., Meng, J., Song, C., and Zheng, K. (2020). Information avoidance and medical screening: A field experiment in china. *Management Science*.

Lim, T.-Z., Lau, J., Wong, G. J., and Tan, K.-K. (2020). Colorectal cancer in patients with single versus double positive faecal immunochemical test results: A retrospective cohort study. *medRxiv*.

Lingenbrink, D. and Iyer, K. (2019). Optimal signaling mechanisms in unobservable queues. *Operations research*, 67(5):1397–1416.

Liu, V., Escobar, G. J., Greene, J. D., Soule, J., Whippy, A., Angus, D. C., and Iwashyna, T. J. (2014). Hospital deaths in patients with sepsis from 2 independent cohorts. *Jama*, 312(1):90–92.

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

Marik, P. E. and Taeb, A. M. (2017). Sirs, qsofa and new sepsis definition. *Journal of thoracic disease*, 9(4):943.

May, T. A., Clancy, M., Critchfield, J., Ebeling, F., Enriquez, A., Gallagher, C., Genevro, J., Kloo, J., Lewis, P., Smith, R., et al. (2006). Reducing unnecessary inpatient laboratory testing in a teaching hospital. *American journal of clinical pathology*, 126(2):200–206.

McDonald, P. J., Strachan, J. A., Digby, J., Steele, R. J., and Fraser, C. G. (2012). Faecal haemoglobin concentrations by gender and age: implications for population-based screening for colorectal cancer. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 50(5):935–940.

McLay, L. A., Foufoulides, C., and Merrick, J. R. (2010). Using simulation-optimization to construct screening strategies for cervical cancer. *Health Care Management Science*, 13(4):294–318.

Michaelson, J. S., Halpern, E., and Kopans, D. B. (1999). Breast cancer: computer simulation method for estimating optimal intervals for screening. *Radiology*, 212(2):551–560.

Ministry of Health Singapore (2020a). Cost of a colonoscopy. `https://www.moh.gov.sg/cost-financing/fee-benchmarks-and-bill-amount-information/`.

Ministry of Health Singapore (2020b). Definition of average risk. `https://www.healthhub.sg/live-healthy/106/screening_colorectal_cancer_nuhs`.

Morgan, C., McBeth, J., Cordingley, L., Watson, K., Hyrich, K. L., Symmons, D. P., and Bruce, I. N. (2015). The influence of behavioural and psychological factors on medication adherence over time in rheumatoid arthritis patients: a study in the biologics era. *Rheumatology*, 54(10):1780–1791.

Mundru, N. (2019). *Predictive and prescriptive methods in operations research and machine learning: an optimization approach*. PhD thesis, Massachusetts Institute of Technology.

Murphy, C. C., Sandler, R. S., Grubber, J. M., Johnson, M. R., and Fisher, D. A. (2016). Underuse and overuse of colonoscopy for repeat screening and surveillance in the veterans health administration. *Clinical Gastroenterology and Hepatology*, 14(3):436–444.

National Registry of Diseases Office (2015). Singapore cancer registry annual registry report 2015. `https://www.nrdo.gov.sg/docs/librariesprovider3/Publications-Cancer/cancer-registry-annual-report-2015_web.pdf?sfvrsn=10`.

Navarro, M., Nicolas, A., Ferrandez, A., and Lanas, A. (2017). Colorectal cancer population screening programs worldwide in 2016: An update. *World journal of gastroenterology*, 23(20):3632.

Nielson, C. M., Vollmer, W. M., Petrik, A. F., Keast, E. M., Green, B. B., and Coronado, G. D. (2019). Factors affecting adherence in a pragmatic trial of annual fecal immunochemical testing for colorectal cancer. *Journal of general internal medicine*, 34(6):978–985.

Office for National Statistics (2019). *Cancer survival in England: adult, stage at diagnosis and childhood - patients followed up to 2018.* DANDY BOOKSELLERS Limited.

Ong, M.-S., Magrabi, F., and Coiera, E. (2018). Delay in reviewing test results prolongs hospital length of stay: a retrospective cohort study. *BMC health services research*, 18(1):369.

Ong, W. M., Chua, S. S., and Ng, C. J. (2014). Barriers and facilitators to self-monitoring of blood glucose in people with type 2 diabetes using insulin: a qualitative study. *Patient preference and adherence*, 8:237.

Oster, E., Shoulson, I., and Dorsey, E. (2013). Optimal expectations and limited medical testing: evidence from huntington disease. *American Economic Review*, 103(2):804–30.

Osterberg, L. and Blaschke, T. (2005). Adherence to medication. *New England journal of medicine*, 353(5):487–497.

Özekici, S. and Pliska, S. R. (1991). Optimal scheduling of inspections: A delayed markov model with false positives and negatives. *Operations Research*, 39(2):261–273.

Paoli, C. J., Reynolds, M. A., Sinha, M., Gitlin, M., and Crouser, E. (2018). Epidemiology and costs of sepsis in the united states—an analysis based on timing of diagnosis and severity level. *Critical care medicine*, 46(12):1889.

Parmigiani, G. (1993). On optimal screening ages. *Journal of the American Statistical Association*, 88(422):622–628.

Pellat, A., Deyra, J., Coriat, R., and Chaussade, S. (2018). Results of the national organised colorectal cancer screening program with fit in paris. *Scientific reports*, 8(1):1–4.

Pencina, M. J., D'Agostino Sr, R. B., and Demler, O. V. (2012). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in medicine*, 31(2):101–113.

Pencina, M. J., D'Agostino Sr, R. B., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*, 30(1):11–21.

Peng, S.-M., Chiu, H.-M., Jen, H.-H., Hsu, C.-Y., Chen, S. L.-S., Chiu, S. Y.-H., Yen, A. M.-F., Fann, J. C.-Y., Lee, Y.-C., and Chen, H.-H. (2019). Quantile-based fecal hemoglobin concentration for assessing colorectal neoplasms with 1,263,717 taiwanese screenees. *BMC medical informatics and decision making*, 19(1):94.

Plumb, A. A., Ghanouni, A., Rainbow, S., Djedovic, N., Marshall, S., Stein, J., Taylor, S. A., Halligan, S., Lyratzopoulos, G., and von Wagner, C. (2017). Patient factors associated with non-attendance at colonoscopy after a positive screening faecal occult blood test. *Journal of Medical Screening*, 24(1):12–19.

Raith, E. P., Udy, A. A., Bailey, M., McGloughlin, S., MacIsaac, C., Bellomo, R., and Pilcher, D. V. (2017). Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *Jama*, 317(3):290–300.

Rapsang, A. G. and Shyam, D. C. (2014). Scoring systems in the intensive care unit: a compendium. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, 18(4):220.

Reuters (2019). Fitbit targets 1 million new users with singapore government tie-up. `https://www.asiaone.com/digital/fitbit-wins-deal-1-million-new-users-singapore-health-plan`. accessed August 22, 2019.

Robertson, D. J., Lee, J. K., Boland, C. R., Dominitz, J. A., Giardiello, F. M., Johnson, D. A., Kaltenbach, T., Lieberman, D., Levin, T. R., and Rex, D. K. (2017). Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the us multi-society task force on colorectal cancer. *Gastroenterology*, 152(5):1217–1237.

Robiner, W. N. (2005). Enhancing adherence in clinical research. *Contemporary Clinical Trials*, 26(1):59–77.

Roesler, A.-K. and Szentes, B. (2017). Buyer-optimal learning and monopoly pricing. *American Economic Review*, 107(7):2072–80.

Rudd, K. E., Kissoon, N., Limmathurotsakul, D., Bory, S., Mutahunga, B., Seymour, C. W., Angus, D. C., and West, T. E. (2018). The global burden of sepsis: barriers and potential solutions. *Critical Care*, 22(1):1–11.

Salisbury, A. C., Reid, K. J., Alexander, K. P., Masoudi, F. A., Lai, S.-M., Chan, P. S., Bach, R. G., Wang, T. Y., Spertus, J. A., and Kosiborod, M. (2011). Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. *Archives of internal medicine*, 171(18):1646–1653.

Sano, W., Hirata, D., Teramoto, A., Iwatate, M., Hattori, S., Fujita, M., and Sano, Y.

(2020). Serrated polyps of the colon and rectum: Remove or not? *World Journal of Gastroenterology*, 26(19):2276.

Schneider, J. L., Rivelli, J. S., Gruss, I., Petrik, A. F., Nielson, C. M., Green, B. B., and Coronado, G. D. (2020). Barriers and facilitators to timely colonoscopy completion for safety net clinic patients. *American journal of health behavior*, 44(4):460–472.

Schweizer, N. and Szech, N. (2018). Optimal revelation of life-changing information. *Management Science*, 64(11):5250–5262.

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. (2019). Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1):1–12.

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA: a cancer journal for clinicians*, 70(1):7–30.

Singapore Health Hub (2011). Singapore health hub. `https://www.healthhub.sg/programmes/61/Screen_for_Life`. accessed November 14, 2019.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Smallwood, R. D. and Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088.

Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Gomes, A. C. P., Payberah, A. H., Zottoli, M., Nazarzadeh, M., et al. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics*, 101:103337.

Spencer, J., Sudan, M., and Xu, K. (2014). Queueing with future information. *ACM SIGMETRICS Performance Evaluation Review*, 41(3):40–42.

Statistics Singapore (2018). Singapore population census 2018. `https://www.singstat.gov.sg/-/media/files/publications/population/population2018.pdf`.

Steyerberg, E. W., Pencina, M. J., Lingsma, H. F., Kattan, M. W., Vickers, A. J., and Van Calster, B. (2012). Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European journal of clinical investigation*, 42(2):216–228.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

Tian, S., Yang, W., Le Grange, J. M., Wang, P., Huang, W., and Ye, Z. (2019). Smart healthcare: making medical care more intelligent. *Global Health Journal*, 3(3):62–65.

Toes-Zoutendijk, E., Kooyker, A. I., Dekker, E., Spaander, M. C., Opstal-van Winden, A. W., Ramakers, C., Buskermolen, M., van Vuuren, A. J., Kuipers, E. J., van Kemenade, F. J., et al. (2020). Incidence of interval colorectal cancer after negative results from first-round fecal immunochemical screening tests, by cutoff value and participant sex and age. *Clinical Gastroenterology and Hepatology*, 18(7):1493–1500.

Tonyushkina, K. and Nichols, J. H. (2009). Glucose meters: a review of technical challenges to obtaining accurate results. *Journal of diabetes science and technology*, 3(4):971–980.

Tsujita, K., Nikolsky, E., Lansky, A. J., Dangas, G., Fahy, M., Brodie, B. R., Dudek, D., Möckel, M., Ochala, A., Mehran, R., et al. (2010). Impact of anemia on clinical outcomes of patients with st-segment elevation myocardial infarction in relation to gender and adjunctive antithrombotic therapy (from the horizons-ami trial). *The American journal of cardiology*, 105(10):1385–1394.

Udell, M. and Boyd, S. (2013). Maximizing a sum of sigmoids. *Optimization and Engineering*, pages 1–25.

Udell, M. and Boyd, S. (2016). Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378.

Vaithianathan, R., Hool, B., Hurd, M. D., and Rohwedder, S. (2018). High-frequency internet survey of a probability sample of older singaporeans: the singapore life panel®. *The Singapore Economic Review*, page 1842004.

van Roon, A. H., Wilschut, J. A., Hol, L., van Ballegooijen, M., Reijerink, J. C., Kranenburg, L. J., Biermann, K., van Vuuren, A. J., Francke, J., van der Togt, A. C., et al. (2011). Diagnostic yield improves with collection of 2 samples in fecal immunochemical test screening without affecting attendance. *Clinical Gastroenterology and Hepatology*, 9(4):333–339.

Vincent, J.-L., De Mendonça, A., Cantraine, F., Moreno, R., Takala, J., Suter, P. M., Sprung, C. L., Colardyn, F., and Blecher, S. (1998). Use of the sofa score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Critical care medicine*, 26(11):1793–1800.

Vincent, J.-L., Marshall, J. C., Ñamendys-Silva, S. A., François, B., Martin-Loeches, I., Lipman, J., Reinhart, K., Antonelli, M., Pickkers, P., Njimi, H., et al. (2014). Assessment of the worldwide burden of critical illness: the intensive care over nations (icon) audit. *The lancet Respiratory medicine*, 2(5):380–386.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure.

Vleugels, J. L., Hazewinkel, Y., Fockens, P., and Dekker, E. (2017). Natural history of diminutive and small colorectal polyps: a systematic literature review. *Gastrointestinal endoscopy*, 85(6):1169–1176.

Vogt, H., Green, S., Ekstrøm, C. T., and Brodersen, J. (2019). How precision medicine and screening with big data could increase overdiagnosis. *BMJ*, 366:l5270.

Vrijburg, K. and Hernández-Peña, P. (2020). Global spending on health: Weathering the storm 2020. *World Health Organization Working paper*, (19.4).

Wang, J., Moehring, J., Stuhr, S., and Krug, M. (2013). Barriers to colorectal cancer screening in hispanics in the united states: an integrative review. *Applied nursing research*, 26(4):218–224.

Wilschut, J. A., Hol, L., Dekker, E., Jansen, J. B., Van Leerdam, M. E., Lansdorp-Vogelaar, I., Kuipers, E. J., Habbema, J. D. F., and Van Ballegooijen, M. (2011). Cost-effectiveness analysis of a quantitative immunochemical test for colorectal cancer screening. *Gastroenterology*, 141(5):1648–1655.

Wong, S.-S., Leong, A. P. K., and Leong, T.-Y. (2004). Cost-effectiveness analysis of colorectal cancer screening strategies in singapore: a dynamic decision analytic approach. In *Medinfo*, pages 104–108.

Xiang, J. (2020). Physicians as persuaders: Evidence from hospitals in china. *Unpublished Manuscript*.

Xu, K. (2015). Necessity of future information in admission control. *Operations Research*, 63(5):1213–1226.

Xu, K. and Chan, C. W. (2016). Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18(3):314–331.

Yu, D., Hopman, W. M., and Paterson, W. G. (2008). Wait time for endoscopic evaluation at a canadian tertiary care centre: Comparison with canadian association of gastroenterology targets. *Canadian Journal of Gastroenterology*, 22.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547.

Zhi, M., Ding, E. L., Theisen-Toupal, J., Whelan, J., and Arnaout, R. (2013a). The landscape of inappropriate diagnostic testing: a 15-year meta-analysis. *PloS one*, 8(11):e78962.

Zhi, M., Ding, E. L., Theisen-Toupal, J., Whelan, J., and Arnaout, R. (2013b). The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PloS one*, 8(11):e78962.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., Bixby, H., Cowan, M. J., Ali, M. K., Taddei, C., et al. (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *The Lancet*, 387(10027):1513–1530.

Zorzi, M., Senore, C., Turrin, A., Mantellini, P., Visioli, C. B., Naldoni, C., de'Bianchi, P. S., Fedato, C., Anghinoni, E., Zappa, M., et al. (2016). Appropriateness of endoscopic surveillance recommendations in organised colorectal cancer screening programmes based on the faecal immunochemical test. *Gut*, 65(11):1822–1828.

Appendix A

Appendix of Chapter 1

A.1 Data and List of Bedside Variables

Table A.1 summarizes the basic statistics of the laboratory test variables in our dataset. Note that the last three test variables, urea, white blood cell (WBC) and prothrombin time (PT), are used in the multiple organ dysfunction score (MODS) and the logistic organ dysfunction score (LODS).

	Median (IQR)	Average time interval between two consecutive updates, hours
Creatinine, $\mu mol/L$	93 (84)	22.1
Platelets, $10^9/L$	175 (112)	21.3
Bilirubin, $\mu mol/L$	13 (13)	40.8
Urea, $mmol/L$	7.3 (5)	22.2
WBC, $10^9/L$	11.9 (6.5)	21.3
PT, seconds	15 (2.5)	18.9

WBC: white blood cell; PT: prothrombin time; IQR: interquartile range.

Table A.1: Summary statistics of laboratory test values of the study population

The bedside variables we considered include vital signs, results from bedside arterial blood gas (ABG) tests, a set of indicators on cardiac rhythm from bedside electrocardiograms, medication, and other readily available variables in real-time. We categorize and list all the beside variables used below.

Vital signs	ABG	Cardiac Rhythm	Medication	Others
Temperature (C)	Arterial pO2	Asystole	Adrenaline	Cumulative urine
Diastolic BP	Arterial pCO2	Sinus rhythm	Nor-adrenaline	GCS
Systolic BP	Arterial pH	Sinus bradycardia	Dobutrex	CVP
Mean Arterial BP	Arterial SaO2	Sinus tachycardia		MV
SpO2	Chloride	Atrial fibrillation		FiO2
Respiration rate	Potassium	Atrial flutter		Total Braden
Heart rate	Sodium	Heart block		
		Junctional rhythm		
		Ventricular fibrillation		
		Ventricular tachycardia		
		Paced rhythm		

ABG: arterial blood gas; BP: blood pressure; SpO2: oxygen saturation (measured by pulse oximeter); pO2: partial pressure of oxygen; pCO2: partial pressure of carbon dioxide; SaO2: oxygen saturation (measured by blood gas analysis); GCS: Glasgow Coma Scale; CVP: central venous pressure; MV: under mechanical ventilation; FiO2: the fraction of inspired oxygen.

Table A.2: Bedside variables

A.2 Prediction Models for Each Test Variable

Five-fold cross-validation is conducted to select the combination of bedside variables that produces the lowest root mean squared error in predicting each test variable. Next, the coefficients are obtained by retraining the selected model using the whole dataset. Note that the last three test variables, urea, white blood cell and prothrombin time, are used in the multiple organ dysfunction score (MODS) and the logistic organ dysfunction score (LODS).

For creatinine, there are 27,872 updated data points. The cross-validated RMSE and R-square of the model are 49.22 and 0.072, respectively.

For platelets, there are 28,936 updated data points. The cross-validated RMSE and R-square of the model are 45.12 and 0.177, respectively.

For bilirubin, there are 6,049 updated data points. The cross-validated RMSE and R-square of the model are 22.41 and 0.011, respectively.

For urea, there are 22,536 updated data points. The cross-validated RMSE and R-square of the model are 3.02 and 0.123, respectively.

For WBC, there are 23,660 updated data points. The cross-validated RMSE and

	Coefficients	Std. Error	p-value
ΔDiastolic BP	-0.1143	0.0025	$< 0.001^{***}$
ΔTemperature	1.2331	0.2377	$< 0.001^{***}$
ΔHeart rate	-1.5570	0.0022	$< 0.001^{***}$
ΔArterial pCO2	-1.2813	0.0071	$< 0.001^{***}$
ΔArterial pH	-164.4317	7.3071	$< 0.001^{***}$
ΔChloride	-0.6080	0.0939	$< 0.001^{***}$
ΔPotassium	16.5072	0.5499	$< 0.001^{***}$
ΔSodium	-0.6707	0.1005	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; BP: blood pressure; pCO2: partial pressure of carbon dioxide.
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.3: Model for predicting creatinine

Factors	Coefficients	Std. Error	p-value
ΔDiastolic BP	0.1904	0.0232	$< 0.001^{***}$
ΔRespiration rate	0.3282	0.0468	$< 0.001^{***}$
ΔHeart rate	0.2017	0.0203	$< 0.001^{***}$
ΔArterial pO2	-0.0558	0.0037	$< 0.001^{***}$
ΔArterial pH	-38.6486	4.7171	$< 0.001^{***}$
ΔPotassium	6.6602	0.4879	$< 0.001^{***}$
ΔSodium	-1.2316	0.1127	$< 0.001^{***}$
ΔPaced rhythm	-16.0756	1.3672	$< 0.001^{***}$
Δt	0.3428	0.0175	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; BP: blood pressure; pO2: partial pressure of oxygen; Δt is the time duration for two consecutive updates of platelets.
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.4: Model for predicting platelets

R-square of the model are 4.08 and 0.107, respectively.

For PT, there are 3,320 updated data points. The cross-validated RMSE and R-square of the model are 2.68 and 0.118, respectively.

Factors	Coefficients	Std. Error	p-value
ΔMean arterial BP	0.0948	0.0289	$< 0.01^{**}$
ΔTotal braden	-0.6067	0.2210	$< 0.001^{***}$
ΔPaced rhythm	5.4178	1.5086	$< 0.001^{***}$
ΔAtrial fibrillation	2.9466	1.6518	$< 0.1^{\cdot}$
ΔGCS	-0.4293	0.1794	$< 0.05^{*}$
ΔArterial pO2	-0.0809	0.0463	$< 0.1^{\cdot}$
ΔGCS/Δt	5.0459	3.0309	$< 0.1^{\cdot}$
Δt	-0.0157	0.0057	$< 0.01^{**}$

Δ indicates changes in respective variable since the previous test; BP: Blood pressure; GCS: Glasgow Coma Scale; pO2: partial pressure of oxygen; Δt is the time duration for two consecutive updates of bilirubin.
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.5: Model for predicting bilirubin

Factors	Coefficients	Std. Error	p-value
ΔHeart rate	-0.0179	0.0014	$< 0.001^{***}$
ΔArterial pO2	-0.0011	0.0002	$< 0.001^{***}$
ΔChloride	-0.0204	0.0061	$< 0.001^{***}$
ΔPotassium	0.6359	0.0347	$< 0.001^{***}$
ΔSodium	-0.0656	0.0063	$< 0.001^{***}$
ΔCumulative urine	0.0001	0.0000	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; pO2: partial pressure of oxygen.
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.6: Model for predicting urea

	Coefficients	Std. Error	p-value
ΔHeart rate	0.0216	0.0018	$< 0.001^{***}$
ΔArterial pO2	0.0064	0.0003	$< 0.001^{***}$
ΔArterial pCO2	- 0.0874	0.0058	$< 0.001^{***}$
ΔArterial pH	- 19.5523	0.5817	$< 0.001^{***}$
ΔPaced rhythm	1.6318	0.1091	$< 0.001^{***}$
ΔGCS	0.0238	0.0069	$< 0.001^{***}$
Δt	- 0.0194	0.0002	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; pO2: partial pressure of oxygen; pCO2: partial pressure of carbon dioxide; GCS: Glasgow Coma Scale; Δt is the time duration for two consecutive updates of WBC. (Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.7: Model for predicting white blood cell

	Coefficients	Std. Error	p-value
ΔDiastolic BP	-0.0103	0.0014	$< 0.001^{***}$
ΔArterial pO2	0.0033	0.0002	$< 0.001^{***}$
ΔArterial pCO2	-0.0884	0.0004	$< 0.001^{***}$
ΔArterial pH	-11.1302	0.3955	$< 0.001^{***}$
ΔArterial SaO2	-0.0098	0.0018	$< 0.001^{***}$
ΔSodium	0.0544	0.0048	$< 0.001^{***}$
ΔPaced rhythm	1.2264	0.0734	$< 0.001^{***}$
ΔFiO2	0.0190	0.0018	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; BP: Blood pressure; pO2: partial pressure of oxygen; pCO2: partial pressure of carbon dioxide; SaO2: oxygen saturation; FiO2: the fraction of inspired oxygen. (Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table A.8: Model for predicting prothrombin time

Appendix B

Appendix of Chapter 2

B.1 POMDP-B

In POMDP-B, the observation probability differs from that of POMDP-UI. let $q^{UI}_{t+1}(\hat{y}_{t+1}|y_{t+1}, \delta_{t+1})$ denote the probability of observing $\hat{y}_{t+1}$ given state y_{t+1} and time lag δ_{t+1}. Given a specific $\delta = \delta^{'}$, we use $B^{UI}_{i,j}(\delta^{'}) = \{q^{UI}(\hat{y} = j|y = i, \delta = \delta^{'})\}$ to denote the information matrix. We next introduce the belief updating process and the optimality equations for POMDP-B.

Belief update (POMDP-B):

Let $f^{UI}_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t = 0)$ stand for the conditional probabilty of observing $\hat{y}_{t+1}$ given $\boldsymbol{\pi}_t, \delta_{t+1}$ and $u_t = 0$. Then

$$f^{UI}_t(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t = 0) = \sum_{y_{t+1}\in\mathcal{Y}} q^{UI}(\hat{y}_{t+1}|y_{t+1}, \delta_{t+1}) \sum_{y_t\in\mathcal{Y}} p(y_{t+1}|y_t, u_t = 0)\pi_t(y_t).$$

If $u_t = 0$, in epoch $t + 1$, given the new observation $\hat{y}_{t+1}$ and time lag δ_{t+1}, the Bayesian update of the belief state $\boldsymbol{\pi}_t$ is computed as follows:

$$\boldsymbol{\pi}_{t+1} = T^{UI}(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}) = \frac{\tilde{B}^{UI}_{\hat{y}_{t+1}}(\delta_{t+1})P^{'}(u_t)\boldsymbol{\pi_t}}{f^{UI}_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t = 0)},$$

where $\tilde{B}^{UI}_{\hat{y}_{t+1}}(\delta_{t+1}) = diag\big(B^{UI}_{1,\hat{y}_{t+1}}(\delta_{t+1}), ..., B^{UI}_{Y,\hat{y}_{t+1}}(\delta_{t+1}), B^{UI}_{A,\hat{y}_{t+1}}(\delta_{t+1})\big)$ and

$$\begin{aligned} f^{UI}_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t = 0) &= \mathbf{1}^{'}_{Y+1}\tilde{B}^{UI}_{\hat{y}_{t+1}}(\delta_{t+1})P^{'}(u_t)\boldsymbol{\pi}_t \\ &= \sum_{y_{t+1}\in\mathcal{Y}} q^{UI}(\hat{y}_{t+1}|y_{t+1}, \delta_{t+1}) \sum_{y_t\in\mathcal{Y}} p(y_{t+1}|y_t, u_t = 0)\pi_t(y_t). \end{aligned}$$

Note that the *i-th* element of $\boldsymbol{\pi}_{t+1}$ equals to

$$\frac{q_{t+1}^{UI}(\hat{y}_{t+1}|y_{t+1}=i,\delta_{t+1})\sum_{y_t\in\mathcal{Y}}p(y_{t+1}=i|y_t,u_t=0)\pi_t(y_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t=0)}.$$

If $u_t=1$, which indicates a diagnostic test is ordered in epoch t, we can observe the patient's true clinical class $\tilde{y}_t$ and update the belief state to $\tilde{\boldsymbol{\pi}}_t=\frac{\tilde{L}_{\tilde{y}_t}\boldsymbol{\pi}_t}{\tilde{f}_t^{UI}(\tilde{y}_t|\boldsymbol{\pi}_t)}$, Then, in epoch $t+1$, the Bayesian update of the belief state is computed by

$$\boldsymbol{\pi}_{t+1}=T^{UI}(\tilde{\boldsymbol{\pi}}_t,\delta_{t+1},u_t,\hat{y}_{t+1}).$$

Optimality Equations (POMDP-B): If a patient is under treatment (i.e., $\pi(A)=1$), no extra costs will be incurred; otherwise, let $Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t)$ denote the total expected cost and $v_t^{UI}(\boldsymbol{\pi}_t,\delta_t)$ represent the minimum total expected cost of a patient. Then

$$\begin{aligned}Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t=0)&=\boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t+\rho\sum_{\hat{y}_{t+1}\in\mathcal{O}}v_{t+1}^{UI}(\boldsymbol{\pi}_{t+1},\delta_{t+1})\cdot f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)\\&=\boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t+\rho\sum_{\hat{y}_{t+1}\in\mathcal{O}}v_{t+1}^{UI}(T^{UI}(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_t+1)\cdot f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t),\end{aligned}$$

$$\begin{aligned}Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t=1)=&\boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t+\rho\sum_{\tilde{y}_t\in\mathcal{O}}\tilde{f}_t^{UI}(\tilde{y}_t|\boldsymbol{\pi}_t)\sum_{\hat{y}_{t+1}\in\mathcal{O}}v_{t+1}^{UI}(T^{UI}(\tilde{\boldsymbol{\pi}}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_{t+1}=1)\\&\cdot f_t^{UI}(\hat{y}_{t+1}|\tilde{\boldsymbol{\pi}}_t,\delta_{t+1},u_t)\end{aligned}$$

and

$$v_t^{UI}(\boldsymbol{\pi}_t,\delta_t)=\min_{u_t\in\mathcal{U}_t}Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t).$$

To summarize, for $t<T$,

$$v_t^{UI}(\boldsymbol{\pi}_t,\delta_t)=\begin{cases}0 & \text{if } \pi_t(A)=1\\ \min_{u_t\in\mathcal{U}_t}Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t) & \text{if } \pi_t(A)\neq 1,\delta_t<|\Delta|\\ Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t=0) & \text{otherwise.}\end{cases}$$

The terminal cost $v^{UI}(\boldsymbol{\pi}_T)$ is represented by

$$v^{UI}(\pi_T)=\boldsymbol{c}_T^{'}\boldsymbol{\pi}_T.$$

B.2 Technique Proofs

B.2.1 Proof of Theorem 1.

We prove Theorem 1 using backward induction.

(i): First, we prove $v_T(\boldsymbol{\pi}_T)$ is *pwlc* with respective to $\boldsymbol{\pi}_T$.

Note that $v_T(\boldsymbol{\pi}_T) = \boldsymbol{c}_T^{'}\boldsymbol{\pi}_T$. Let $\Gamma_T(\delta_T) = \{\boldsymbol{c}_T\}$ for $\forall \delta_T \in \Delta$. Then $v(\pi_T) = \min_{\boldsymbol{\gamma}_T \in \Gamma_T(\delta_T)} \boldsymbol{\gamma}_T^{'}\boldsymbol{\pi}_T$, which is *pwlc* with respect to the belief state π_T.

(ii): Second, we prove that for $t \in \{1, 2, ..., T-1\}$, if $v_{t+1}(\boldsymbol{\pi}_{t+1}, \delta_{t+1})$ is *pwlc* with respect to $\boldsymbol{\pi}_{t+1}$, then $v_t(\boldsymbol{\pi}_t, \delta_t)$ is *pwlc* with respect to $\boldsymbol{\pi}_t$.

Suppose $v_{t+1}(\boldsymbol{\pi}_{t+1}, \delta_{t+1})$ is *pwlc* with respect to $\boldsymbol{\pi}_{t+1}$, i.e.: $v_{t+1}(\boldsymbol{\pi}_{t+1}, \delta_{t+1}) = \min_{\boldsymbol{\gamma}_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} \boldsymbol{\gamma}_{t+1}^{'}\boldsymbol{\pi}_{t+1}$.

(a): If $\pi_t(A) = 1$, then $\Gamma_{t+1}(\delta_{t+1}) = 0$, $\Gamma_t(\delta_t) = 0$, so the conclusion holds.

(b): If $u_t = 0$ and $\pi_t(A) = 0$, then $\delta_{t+1} = \delta_t + 1$, we have

$$
\begin{aligned}
Q_t(\boldsymbol{\pi}_t, \delta_t, u_t) =& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} v_{t+1}(T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}), \delta_{t+1}) \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \boldsymbol{\pi}_t, \delta_{t+1}, u_t) \\
=& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} v_{t+1}\Big(\frac{\tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) P^{'}(u_t)\boldsymbol{\pi}_t}{f_t(\hat{y}_{t+1}, \theta_{t+1} | \boldsymbol{\pi}_t, \delta_{t+1}, u_t)}, \delta_{t+1}\Big) \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \boldsymbol{\pi}_t, \delta_{t+1}, u_t) \\
=& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} \min_{\boldsymbol{\gamma}_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} \rho \boldsymbol{\gamma}_{t+1}^{'} \tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) P^{'}(u_t)\boldsymbol{\pi}_t \\
=& \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} \min_{\boldsymbol{\gamma}_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} \Big(\frac{c_t(u_t)}{(Y+1)|\Theta|} + \rho P(u_t) \tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) \boldsymbol{\gamma}_{t+1}\Big)^{'} \boldsymbol{\pi}_t,
\end{aligned}
$$

which is a summation of finite *pwlc* functions, so $Q_t(\boldsymbol{\pi}_t, \delta_t, u_t)$ is *pwlc*.

(c): If $u_t = 1$, then $\delta_{t+1} = 1$, we have

$$\begin{aligned}
Q_t(\boldsymbol{\pi}_t, \delta_t, u_t) =& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\tilde{y}_t \in \mathcal{O}} \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} v_{t+1}(T(\tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}), \delta_{t+1}) \\
& \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t) \tilde{f}_t(\tilde{y}_t | \boldsymbol{\pi}_t) \\
=& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\tilde{y}_t \in \mathcal{O}} \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} v_{t+1}(\frac{\tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) P^{'}(u_t) \tilde{\boldsymbol{\pi}}_t}{f_t(\hat{y}_{t+1}, \theta_{t+1} | \tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t) \pi_t(\tilde{y}_t)}, \delta_{t+1}) \\
& \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t) \cdot \tilde{f}_t(\tilde{y}_t | \boldsymbol{\pi}_t) \\
=& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\tilde{y}_t \in \mathcal{O}} \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} \min_{\gamma_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} \boldsymbol{\gamma}_{t+1}^{'} \frac{\tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) P^{'}(u_t) \tilde{\boldsymbol{\pi}}_t}{f_t(\hat{y}_{t+1}, \theta_{t+1} | \tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t) \pi_t(\tilde{y}_t)} \\
& \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \tilde{\boldsymbol{\pi}}_t, \delta_{t+1}, u_t) \cdot \tilde{f}_t(\tilde{y}_t | \boldsymbol{\pi}_t) \\
=& \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\tilde{y}_t \in \mathcal{O}} \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} \min_{\gamma_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} \boldsymbol{\gamma}_{t+1}^{'} \tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) P^{'}(u_t) \\
& \cdot \frac{\tilde{L}_{\tilde{y}_t} \boldsymbol{\pi}_t}{1_{Y+1}^{'} \tilde{L}_{\tilde{y}_t} \boldsymbol{\pi}_t} \cdot 1_{Y+1}^{'} \tilde{L}_{\tilde{y}_t} \boldsymbol{\pi}_t \\
=& \sum_{\tilde{y}_t \in \mathcal{O}} \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} \min_{\gamma_{t+1} \in \Gamma_{t+1}(\delta_{t+1})} (\frac{c_t(u_t)}{(Y+1)^2 |\Theta|} + \rho \tilde{L}_{\tilde{y}_t} P(u_t) \tilde{B}_{\hat{y}_{t+1}, \theta_{t+1}}(\delta_{t+1}) \boldsymbol{\gamma}_{t+1})^{'} \boldsymbol{\pi}_t,
\end{aligned}$$

which is the summation of finite *pwlc* functions, so $Q_t(\boldsymbol{\pi}_t, \delta_t, u_t)$ is *pwlc*. In addition, since $v_t(\boldsymbol{\pi}_t, \delta_t) = \min_{u_t \in \mathcal{U}_t} Q_t(\boldsymbol{\pi}_t, \delta_t, u_t)$ and the minimization preserves the *pwlc* property, so $v_t(\boldsymbol{\pi}_t, \delta_t)$ is *pwlc* with respect to $\boldsymbol{\pi}_t$. □

B.2.2 Proof of Theorem 2.

We aim to prove that given specific $\boldsymbol{\pi}_t$ and δ_t, $v_t(\boldsymbol{\pi}_t, \delta_t) \leq v_t^{UI}(\boldsymbol{\pi}_t, \delta_t)$. Specifically, we prove it by backward induction.

(i): When $t = T$, $v(\pi_T) = v^{UI}(\pi_T) = \boldsymbol{c}_T^{'} \boldsymbol{\pi}_T$.

(ii): For $t \in \{1, 2, ..., T-1\}$, suppose $v_{t+1}(\boldsymbol{\pi}_{t+1}, \delta_{t+1}) \leq v_{t+1}^{UI}(\boldsymbol{\pi}_{t+1}, \delta_{t+1})$. We next prove $v_t(\boldsymbol{\pi}_t, \delta_t) \leq v_t^{UI}(\boldsymbol{\pi}_t, \delta_t)$.

If u_t=0,

$$Q_t(\boldsymbol{\pi}_t, \delta_t, u_t) = \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1} \in \mathcal{O}} \sum_{\theta_{t+1} \in \Theta} v_{t+1}(T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1}), \delta_t + 1) \cdot f_t(\hat{y}_{t+1}, \theta_{t+1} | \boldsymbol{\pi}_t, \delta_{t+1}, u_t) \tag{B.1}$$

and

$$Q_t^{UI}(\boldsymbol{\pi}_t, \delta_t, u_t) = \boldsymbol{c}_t^{'}(u_t)\boldsymbol{\pi}_t + \rho \sum_{\hat{y}_{t+1} \in \mathcal{O}} v_{t+1}^{UI}(T^{UI}(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}), \delta_t + 1) \cdot f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t). \tag{B.2}$$

Since

$$f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t) = \sum_{y_{t+1} \in \mathcal{Y}} q(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1}, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1}|y_t, u_t = 0)\pi_t(y_t) \tag{B.3}$$

and

$$f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t) \sum_{y_{t+1} \in \mathcal{Y}} q^{UI}(\hat{y}_{t+1}|y_{t+1}, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1}|y_t, u_t = 0)\pi_t(y_t), \tag{B.4}$$

under these two models, $q^{UI}(\hat{y}_{t+1}|y_{t+1}, \delta_{t+1}) = \sum_{\theta_{t+1} \in \Theta} q_(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1}, \delta_{t+1})$; thus, given the exactly same $\boldsymbol{\pi}_t, \delta_t$, and u_t, patients should have equal probability of observing $\hat{y}_{t+1}$ in epoch $t+1$, that is

$$f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t) = \sum_{\theta_{t+1} \in \Theta} f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t). \tag{B.5}$$

Recall that the *i-th* element of $T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1})$ equals

$$\frac{q(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1} = i, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1} = i|y_t, u_t)\pi_t(y_t)}{f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)} \tag{B.6}$$

and the *i-th* element of $T^{UI}(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1})$ equals

$$\frac{q^{UI}(\hat{y}_{t+1}|y_{t+1} = i, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1} = i|y_t, u_t)\pi_t(y_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}. \tag{B.7}$$

Let $\alpha(i) = \frac{f_t(\hat{y}_{t+1}, \theta_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}$. According to Equality (1), $\sum_{i \in \Theta} \alpha(i) = 1$. Since the *i-th* element of $\sum_{i \in \Theta} \alpha(i) T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1} = i)$ equals

$$\frac{\sum_{i \in \Theta} q(\hat{y}_{t+1}, \theta_{t+1}|y_{t+1} = i, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1} = i|y_t, u_t)\pi_t(y_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}$$
$$= \frac{q^{UI}(\hat{y}_{t+1}|y_{t+1} = i, \delta_{t+1}) \sum_{y_t \in \mathcal{Y}} p(y_{t+1} = i|y_t, u_t)\pi_t(y_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t, \delta_{t+1}, u_t)}.$$

Which is also equal to the *i-th* element of $T^{UI}(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1})$. Thus,

$$\sum_{i \in \Theta} \alpha(i) T(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}, \theta_{t+1} = i) = T^{UI}(\boldsymbol{\pi}_t, \delta_{t+1}, u_t, \hat{y}_{t+1}).$$

According to the concavity of Q_t, we have

$$\begin{aligned}
&\frac{\sum\limits_{\theta_{t+1}\in\Theta} v_{t+1}(T(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1},\theta_{t+1}),\delta_t+1)\cdot f_t(\hat{y}_{t+1},\theta_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)}\\
\le& v_{t+1}(\frac{T(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1},\theta_{t+1})\cdot f_t(\hat{y}_{t+1},\theta_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)}{f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)},\delta_t+1)\\
=& v_{t+1}(T^{UI}(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_t+1)\\
\Longrightarrow& \sum_{\theta_{t+1}\in\Theta} v_{t+1}(T(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1},\theta_{t+1}),\delta_t+1)\cdot f_t(\hat{y}_{t+1},\theta_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t)\\
&\le v_{t+1}(T^{UI}(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_t+1)\cdot f_t^{UI}(\hat{y}_{t+1}|\boldsymbol{\pi}_t,\delta_{t+1},u_t).
\end{aligned}$$

Since $v_{t+1}(\boldsymbol{\pi}_{t+1},\delta_{t+1}) \le v_{t+1}^{UI}(\boldsymbol{\pi}_{t+1},\delta_{t+1})$, so $v_{t+1}(T^{UI}(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_t+1) \le v_{t+1}^{UI}(T^{UI}(\boldsymbol{\pi}_t,\delta_{t+1},u_t,\hat{y}_{t+1}),\delta_t+1)$. According to Equality (B.1) and Equality (B.2), $Q_t(\boldsymbol{\pi}_t,\delta_t,u_t=0) \le Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t=0)$ holds.

Similarly, we can prove that $Q_t(\boldsymbol{\pi}_t,\delta_t,u_t=1) \le Q_t^{UI}(\boldsymbol{\pi}_t,\delta_t,u_t=1)$ (the proof is similar and therefore ommited). Then based on the definition of $v_t(\boldsymbol{\pi}_t,\delta_t)$ and $v_t^{UI}(\boldsymbol{\pi}_t,\delta_t)$, we finally reach the conclusion that $v_t(\boldsymbol{\pi}_t,\delta_t) \le v_t^{UI}(\boldsymbol{\pi}_t,\delta_t)$. □

B.3 Numerical Experiments

B.3.1 H-case

The transition matrix between the clinical classes are presented as follows:

$$P(u=0)=\begin{pmatrix}0.94 & 0.06 & 0\\ 0.09 & 0.91 & 0\\ 0 & 0 & 1\end{pmatrix}, P(u=1)=\begin{pmatrix}0.94 & 0.06 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\end{pmatrix}.$$

The information matrices for POMDP-B are shown as below:

$$B(\delta=1)=\begin{pmatrix}0.91 & 0.09 & 0\\ 0.10 & 0.90 & 0\\ 0 & 0 & 1\end{pmatrix}, B(\delta=2)=\begin{pmatrix}0.83 & 0.17 & 0\\ 0.19 & 0.81 & 0\\ 0 & 0 & 1\end{pmatrix},$$

,

$$B(\delta=3)=\begin{pmatrix}0.75 & 0.25 & 0\\ 0.27 & 0.73 & 0\\ 0 & 0 & 1\end{pmatrix},B(\delta=4)=\begin{pmatrix}0.63 & 0.37 & 0\\ 0.36 & 0.64 & 0\\ 0 & 0 & 1\end{pmatrix},$$

$$B(\delta=5)=\begin{pmatrix}0.56 & 0.44 & 0\\ 0.45 & 0.55 & 0\\ 0 & 0 & 1\end{pmatrix},B(\delta=6)=\begin{pmatrix}0.50 & 0.50 & 0\\ 0.51 & 0.49 & 0\\ 0 & 0 & 1\end{pmatrix},$$

,

$$B(\delta=7)=\begin{pmatrix}0.43 & 0.57 & 0\\ 0.56 & 0.44 & 0\\ 0 & 0 & 1\end{pmatrix},B(\delta=8)=\begin{pmatrix}0.38 & 0.62 & 0\\ 0.62 & 0.38 & 0\\ 0 & 0 & 1\end{pmatrix}.$$

The information matrices for POMDP-UI are shown as below:

$$B(\delta=1)=\begin{pmatrix}0.81 & 0.10 & 0.01 & 0.08 & 0\\ 0.01 & 0.09 & 0.83 & 0.07 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},B(\delta=2)=\begin{pmatrix}0.72 & 0.11 & 0.02 & 0.15 & 0\\ 0.03 & 0.16 & 0.68 & 0.13 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},$$

,

$$B(\delta=3)=\begin{pmatrix}0.58 & 0.17 & 0.05 & 0.20 & 0\\ 0.06 & 0.21 & 0.55 & 0.18 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},B(\delta=4)=\begin{pmatrix}0.44 & 0.19 & 0.11 & 0.26 & 0\\ 0.11 & 0.25 & 0.45 & 0.19 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},$$

,

$$B(\delta=5)=\begin{pmatrix}0.34 & 0.22 & 0.18 & 0.26 & 0\\ 0.17 & 0.28 & 0.34 & 0.21 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},B(\delta=6)=\begin{pmatrix}0.28 & 0.22 & 0.22 & 0.28 & 0\\ 0.23 & 0.28 & 0.27 & 0.22 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},$$

,

$$B(\delta=7)=\begin{pmatrix}0.23 & 0.20 & 0.27 & 0.30 & 0\\ 0.30 & 0.26 & 0.21 & 0.23 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix},B(\delta=8)=\begin{pmatrix}0.16 & 0.22 & 0.35 & 0.27 & 0\\ 0.37 & 0.25 & 0.15 & 0.23 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}.$$

B.3.2 L-case

The transition matrix between the clinical classes are presented as follows:

$$P(u=0) = \begin{pmatrix} 0.98 & 0.02 & 0 \\ 0.10 & 0.90 & 0 \\ 0 & 0 & 1 \end{pmatrix}, P(u=1) = \begin{pmatrix} 0.98 & 0.02 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The information matrices of POMDP-B and POMDP-UI are the same as those in the H case.

The trade-off between *Number of Tests* and *Detection time* and the trade-off between *Number of Tests* and *Missed Periods* are presented as follows (Table B.1). The insights are similar to those in the H case, and we omit the discussion.

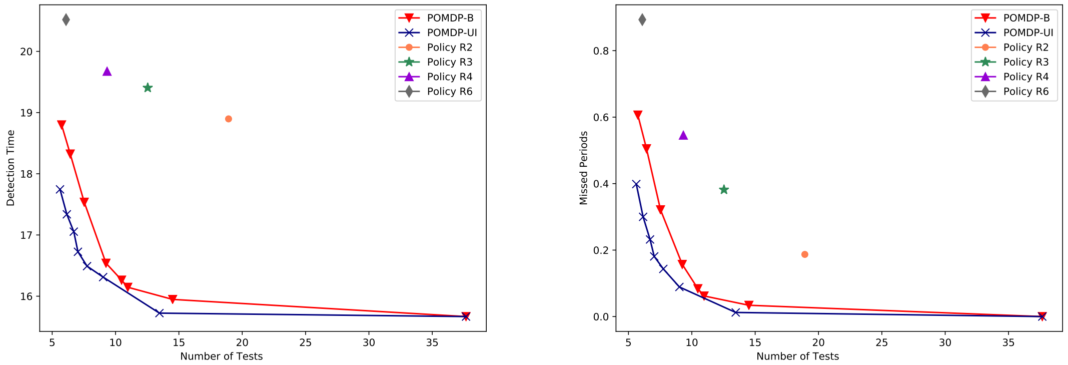


Figure B.1: Trade-off between *Number of Tests* and *Detection time* and trade-off between *Number of Tests* and *Missed Periods* in numerical experiments (L-case)

B.4 Case Study

B.4.1 Prediction Models for Blood Glucose Level

The regression model for blood glucose level is presented in Table B.1.

	Coefficients	Std. Error	p-value
ΔDiastolic BP	-0.01	0.00	$< 0.001^{***}$
ΔTotal Braden	-0.06	0.01	$< 0.001^{***}$
ΔGCS	0.02	0.01	$< 0.001^{***}$
ΔHeart Rate	0.01	0.00	$< 0.001^{***}$
ΔChloride	-0.03	0.00	$< 0.001^{***}$
ΔArterial pCO2	-0.03	0.00	$< 0.001^{***}$
ΔArterial pH	-7.24	0.31	$< 0.001^{***}$
ΔArterial SaO2	0.01	0.00	$< 0.001^{***}$
ΔFiO2	-0.02	0.00	$< 0.001^{***}$
Δt	-0.01	0.00	$< 0.001^{***}$

Δ indicates changes in respective variable since the previous test; BP: Blood pressure; pCO2: partial pressure of carbon dioxide; SaO2: oxygen saturation; FiO2: the fraction of inspired oxygen; Δt is the time duration for two consecutive updates of blood glucose level.
(Significance Level: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.')

Table B.1: Model for predicting blood glucose level

B.4.2 Information Matrices

The information matrices for POMDP-B are shown as below:

$$B(\delta = 1) = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.35 & 0.65 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B(\delta = 2) = \begin{pmatrix} 0.89 & 0.11 & 0 \\ 0.27 & 0.73 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 3) = \begin{pmatrix} 0.88 & 0.12 & 0 \\ 0.34 & 0.66 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B(\delta = 4) = \begin{pmatrix} 0.85 & 0.15 & 0 \\ 0.33 & 0.67 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 5) = \begin{pmatrix} 0.88 & 0.12 & 0 \\ 0.46 & 0.54 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B(\delta = 6) = \begin{pmatrix} 0.90 & 0.10 & 0 \\ 0.48 & 0.52 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 7) = \begin{pmatrix} 0.92 & 0.08 & 0 \\ 0.64 & 0.36 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B(\delta = 8) = \begin{pmatrix} 0.96 & 0.04 & 0 \\ 0.79 & 0.21 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 9) = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B(\delta >= 10) = \begin{pmatrix} 0.97 & 0.03 & 0 \\ 0.59 & 0.41 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The information matrices for POMDP-UI are shown as below:

$$B(\delta = 1) = \begin{pmatrix} 0.95 & 0.00 & 0.04 & 0.01 & 0 \\ 0.34 & 0.01 & 0.64 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, B(\delta = 2) = \begin{pmatrix} 0.88 & 0.01 & 0.11 & 0.00 & 0 \\ 0.26 & 0.01 & 0.72 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 3) = \begin{pmatrix} 0.87 & 0.01 & 0.12 & 0.00 & 0 \\ 0.33 & 0.01 & 0.64 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, B(\delta = 4) = \begin{pmatrix} 0.84 & 0.01 & 0.14 & 0.0.01 & 0 \\ 0.29 & 0.04 & 0.65 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 5) = \begin{pmatrix} 0.88 & 0.00 & 0.12 & 0.00 & 0 \\ 0.46 & 0 & 0.52 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, B(\delta = 6) = \begin{pmatrix} 0.89 & 0.01 & 0.10 & 0.00 & 0 \\ 0.47 & 0.01 & 0.52 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 7) = \begin{pmatrix} 0.92 & 0.00 & 0.07 & 0.01 & 0 \\ 0.63 & 0.01 & 0.34 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, B(\delta = 8) = \begin{pmatrix} 0.96 & 0 & 0.04 & 0.00 & 0 \\ 0.75 & 0.04 & 0.19 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

,

$$B(\delta = 9) = \begin{pmatrix} 0.94 & 0.01 & 0.04 & 0.01 & 0 \\ 0.45 & 0.05 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, B(\delta \geq 10) = \begin{pmatrix} 0.97 & 0.00 & 0.03 & 0.00 & 0 \\ 0.59 & 0 & 0.39 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

.

Appendix C

Appendix of Chapter 3

C.1 Notations

The notations we used in the model is summarized in Table C.1.

Notation	Description
Initial test	
$s \in S = \{0, 1\}$	Individuals' CRC state, where $s = 1$ ($s = 0$) represents a health state with (without) CRC
ζ	f-Hb concentration level tested from FIT
$[\underline{\zeta}, \bar{\zeta}]$	The range of f-Hb concentration
$H_1(\cdot)$ $(H_0(\cdot))$	The CDF of f-Hb concentration for individuals with (without) CRC
$h_1(\cdot)$ $(h_0(\cdot))$	The PDF of f-Hb concentration for individuals with (without) CRC
T	Number of cut-off points selected
$\mathcal{C}_T = \{c_1, c_2, .., c_T \| c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$	The set of cut-off point values
$(T, \mathcal{C}_T)$	Initial test design
$\Gamma^{\mathcal{C}_T} = \{0, 1, 2, ..., T\}$	The set of test outcomes if the initial test design is $(T, \mathcal{C}_T)$
$\sigma^{\mathcal{C}_T}(t\|s)$	The likelihood of receiving test outcome t if individual's state is s and the initial test design is $(T, \mathcal{C}_T)$

Individual's follow-up problem	
N	Population size
$i \in \{1, 2, ..., N\}$	Individual index
$\lambda_i(t)$	The total probability of receiving test outcome t for individual i
p_i^0	Individual i's prior risk of developing CRC
$p_i^s(t)$	Individual i's posterior risk of developing CRC given test outcome t
$\pi_i^s(t)$	Individual i's subjective belief of having CRC given test outcome t
$a_i(t) \in \{0, 1\}$	Individual i's follow-up action given test outcome t, where $a_i(t) = 1$ refers to a follow up with the second-stage test; $a_i(t) = 0$, otherwise.
$u_i(s_i, a_i(t))$	The utility for individual i given CRC state s_i and follow-up action $a_i(t)$
$d_i(s_i)$	Perceived disutility of the follow-up action for individual i with health state s_i
$\epsilon_i^{a_i}$	A random error term that captures the impact of all unobservable factors that affect the utility of choosing action a_i by individual i
τ	An adjusting term that captures the "cost" of extra colonoscopy demand from healthy individuals
$f_i(t)$	The probability of following up with the second-stage test for an individual i with an initial test outcome t
$\Phi(\cdot)$	Function capturing the relationship between $\pi_i^s(t)$ and $p_i^s(t)$ when $t \neq 0$, i.e., $\pi_i^s(t) = \Phi(p_i^s(t))$
$W(p_i^s(t))$	The probability of following up with the second-stage test for an individual i with an initial test outcome t if $t \neq 0$
Analytical model	
$\mathcal{V}_m \equiv \{v_1, v_2, ..., v_m \mid v_j \in [\underline{\zeta}, \bar{\zeta}], v_1 < v_2 < ... < v_m\}$	A pre-specified candidate set of cut-off points
$W(p_i^s(t))$	The probability of following up with the second-stage test for an individual i with an initial test outcome t if $t \neq 0$

x_j	A binary variable that denotes whether cut-off value v_j is selected, where $x_j = 1$ if v_j is chosen; $x_j = 0$, otherwise.
$f_i^j(t)$	If v_j is chosen, the probability of following up with the second-stage test for an individual i with an initial test outcome t
$\sigma_j(t\|s)$	The likelihood of receiving test outcome t if v_j is chosen given individual's state is s
L	The maximal number of test types
q_{ij}	A binary variable that denotes whether the FIT kit with cut-off point v_j is assigned to individual i

Table C.1: Model notation

C.2 Technique Proofs

C.2.1 Proof of Lemma 4.1.

Since

$$p_i^s(t) = \frac{\sigma(t|1)p_i^0}{\lambda_i(t)} = \frac{\sigma(t|1)p_i^0}{\sigma(t|0)(1-p_i^0) + \sigma(t|1)p_i^0} = \frac{p_i^0}{\frac{\sigma(t|0)}{\sigma(t|1)}(1-p_i^0) + p_i^0},$$

according to Property 1 that for MLR-feasible test, $\frac{\sigma(t|1)}{\sigma(t|0)}$ increases in t, we conclude that $p_i^s(t)$ increases in t. □

C.2.2 Proof of Theorem 4.2.

We prove this theorem via two steps: (1) we first prove that when $W(\cdot)$ is concave in $p_i^s(t)$, a dichotomous test is optimal; (2) we then prove that the optimal cut-off point value is $\underline{\zeta}$.

Let (T, C_T) be an initial test that has more than one cut-off point. That is $T > 1$ and $\mathcal{C}_T = \{c_1, c_2, .., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$. The corresponding test outcome set is $\Gamma^{\mathcal{C}_T} = \{0, 1, 2, ..., T\}$ and the likelihood of receiving test outcome t given state s_i is $\sigma^{\mathcal{C}_T}(t|s_i)$, $t \in \Gamma^{\mathcal{C}_T}, s_i \in S$.

Let $\lambda_i^{\mathcal{C}_T}(t)$ denote the total probability of receiving test outcome t, i.e., $\lambda_i^{\mathcal{C}_T}(t) = \sigma^{\mathcal{C}_T}(t|0)(1-p_i^0) + \sigma^{\mathcal{C}_T}(t|1)p_i^0$. Then for an individual i, the expected following up probability under $(T, \mathcal{C}_T)$ is

$$\sum_{t \in \Gamma^{\mathcal{C}_T}} \lambda_i^{\mathcal{C}_T}(t) f_i(t) = \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t) W(p_i^s(t)). \tag{C.1}$$

Consider another initial test that has one cut-off point, and the value is c_1. That is $T = 1$, $\mathcal{C}_1 = \{c_1\}$ and $\Gamma^{\mathcal{C}_1} = \{0, 1\}$. By construction, we have $\sigma^{\mathcal{C}_1}(0|s_i) = \sigma^{\mathcal{C}_T}(0|s_i), \sigma^{\mathcal{C}_1}(1|s_i) = \sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \sigma^{\mathcal{C}_T}(t|s_i)$. Let $\lambda_i^{\mathcal{C}_1}(j)$ denote the probability of receiving test outcome j ($j \in \Gamma^{\mathcal{C}_1}$), then $\lambda_i^{\mathcal{C}_1}(0) = \lambda_i^{\mathcal{C}_T}(0)$, $\lambda_i^{\mathcal{C}_1}(1) = \sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t)$. Let $\hat{f}_i(t)$ denote individual i's probability of following up under $(1, \mathcal{C}_1)$. Then for an individual i, the expected follow up probability under $(1, \mathcal{C}_1)$ is

$$\sum_{j \in \Gamma^{\mathcal{C}_1}} \lambda_i^{\mathcal{C}_1}(j) \hat{f}_i(j) = \lambda_i^{\mathcal{C}_1}(1) \hat{f}_i(1) = \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t) W\Big(\frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t)} \Big). \tag{C.2}$$

Followed by the concavity of $W(\cdot)$ is concave in $p_i^s(t)$, that is,

$$W\Big(\frac{\sum\limits_{t \in \mathcal{C}_T \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t \in \mathcal{C}_T \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t)} \Big) \geq \frac{\sum\limits_{t \in \mathcal{C}_T \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t) W(p_i^s(t))}{\sum\limits_{t \in \mathcal{C}_T \backslash \{0\}} \lambda_i^{\mathcal{C}_T}(t)}. \tag{C.3}$$

Therefore, the initial test design $(1, \mathcal{C}_1)$ induces an equal or higher probability of following up than $(T, \mathcal{C}_T)$ for any individual i ($i \in \{1, 2, ..., N\}$). A dichotomous test is optimal.

To prove (2), consider a dichotomous test with cut-off point $\underline{\zeta}$ and another with cut-off point $c^{'}$ where $\underline{\zeta} < c^{'} \leq \bar{\zeta}$.

Given the concavity of $W(\cdot)$, we have

$$W\Big(\frac{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 + 0(\zeta)\cdot(1-p_i^0)d\zeta}\Big) \geq \frac{\int_{c'}^{\bar{\zeta}} h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)d\zeta}{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)d\zeta} \tag{C.4}$$

$$\cdot W\left(\frac{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 + h_0(\zeta)\cdot(1-p_i^0)d\zeta}\right)$$

$$+\frac{\int_{\underline{\zeta}}^{c^{'}} h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)d\zeta}{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)d\zeta} \tag{C.5}$$

$$\cdot W\left(\frac{\int_{\underline{\zeta}}^{c^{'}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{\underline{\zeta}}^{c^{'}} h_1(\zeta)\cdot p_i^0 + h_0(\zeta)\cdot(1-p_i^0)d\zeta}\right) \tag{C.6}$$

Therefore, we can establish that the follow-up probability of any individual i under the test with cut-off point $\underline{\zeta}$ is higher than that under the test with cut-off point c' via the following.

$$\underbrace{\Big[\int_{\underline{\zeta}}^{\bar{\zeta}}\Big(h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)\Big)d\zeta\Big]}_{\substack{\text{Probability of receiving outcome 1}\\ \text{(cut-off point} = \underline{\zeta})}}\cdot\underbrace{W\Big(\frac{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{\underline{\zeta}}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 + 0(\zeta)\cdot(1-p_i^0)d\zeta}\Big)}_{\substack{\text{Follow-up probablity if receiving outcome 1}\\ \text{(cut-off point} = \underline{\zeta})}}$$

$$\geq \int_{c'}^{\bar{\zeta}}\Big(h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)\Big)d\zeta\cdot W\left(\frac{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 + h_0(\zeta)\cdot(1-p_i^0)d\zeta}\right)$$

$$+\int_{\underline{\zeta}}^{c^{'}}\Big(h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)\Big)d\zeta\cdot W\left(\frac{\int_{\underline{\zeta}}^{c^{'}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{\underline{\zeta}}^{c^{'}} h_1(\zeta)\cdot p_i^0 + h_0(\zeta)\cdot(1-p_i^0)d\zeta}\right)$$

$$>\underbrace{\Big[\int_{c'}^{\bar{\zeta}}\Big(h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)\Big)d\zeta\Big]}_{\substack{\text{Probability of receiving outcome 1}\\ \text{(cut-off point} = c')}}\cdot\underbrace{W\Big(\frac{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 d\zeta}{\int_{c'}^{\bar{\zeta}} h_1(\zeta)\cdot p_i^0 + h_0(\zeta)\cdot(1-p_i^0)d\zeta}\Big)}_{\substack{\text{Follow-up probablity if receiving outcome 1}\\ \text{(cut-off point} = c')}}$$

This completes the proof. □

C.2.3 Proof of Theorem 4.3.

We prove this theorem via two steps: (1) we first show that for any initial test with a finite number of cut-off points, the objective value will not decrease by arbitrarily adding one more cut-off point from $[\underline{\zeta}, \bar{\zeta}]$; (2) we prove that if we uniformly add infinitely many cut-off points, the objective value converges to that of the continuous test.

To prove (1), we first establish the following result.

Corollary C.1. *If $\frac{a_j}{b_j}$ and w_j are both nondecreasing in j, where a_j, b_j and w_j are positive variables, $j \in \{1, 2, ..., J\}$. Then $\frac{\sum_{j=1}^{J} a_j w_j}{\sum_{j=1}^{J} a_j} \geq \frac{\sum_{j=1}^{J} (a_j+b_j) w_j}{\sum_{j=1}^{J} (a_j+b_j)}$.*

Proof. Because a_j, b_j and x_j are positive variables, so it's equivalent to prove

$$\frac{\sum_{j=1}^{J} b_j}{\sum_{j=1}^{J} a_j} \geq \frac{\sum_{j=1}^{J} b_j w_j}{\sum_{j=1}^{J} a_j w_j}. \tag{C.7}$$

Let $k_1 = \frac{b_1}{a_1}$ and $k_j = \frac{b_j}{a_j} - \frac{b_{j-1}}{a_{j-1}}$ for $j \in \{2, 3, ..., J\}$. Then we have $\frac{b_j}{a_j} = \sum_{l=1}^{j} k_l$ and $b_j = \sum_{l=1}^{j} k_l a_j$. Thus,

$$\frac{\sum_{j=1}^{J} b_j}{\sum_{j=1}^{J} a_j} = \frac{\sum_{j=1}^{J}\sum_{l=1}^{j} k_l a_j}{\sum_{j=1}^{J} a_j} = k_1 + k_2 \frac{\sum_{j=2}^{J} a_j}{\sum_{j=1}^{J} a_j} + k_3 \frac{\sum_{j=3}^{J} a_j}{\sum_{j=1}^{J} a_j} + ... + k_J \frac{\sum_{j=J}^{J} a_j}{\sum_{j=1}^{J} a_j}.$$

We also have

$$\frac{\sum_{j=1}^{J} b_j w_j}{\sum_{j=1}^{J} a_j w_j} = \frac{\sum_{j=1}^{J}\sum_{l=1}^{j} k_l a_j w_j}{\sum_{j=1}^{J} a_j w_j} = k_1 + k_2 \frac{\sum_{j=2}^{J} a_j w_j}{\sum_{j=1}^{J} a_j w_j} + k_3 \frac{\sum_{j=3}^{J} a_j w_j}{\sum_{j=1}^{J} a_j w_j} + ... + k_J \frac{\sum_{j=J}^{J} a_j w_j}{\sum_{j=1}^{J} a_j w_j}.$$

Since $\frac{a_j}{b_j}$ is nondecreasing in j; therefore, $k_j = \frac{b_j}{a_j} - \frac{b_{j-1}}{a_{j-1}} \leq 0$ for $j \in \{2, 3, ..., J\}$. Thus, to prove (C.7), it's sufficient to prove that for $l \in \{1, 2, ..., J\}$, the following inequality holds:

$$\frac{\sum\limits_{j=l}^{J} a_j}{\sum\limits_{j=1}^{J} a_j} \leq \frac{\sum\limits_{j=l}^{J} a_j w_j}{\sum\limits_{j=1}^{J} a_j w_j}. \tag{C.8}$$

To prove (C.8), it's equivalent to prove $\frac{\sum_{j=l}^{J} a_j}{\sum_{j=l}^{J} a_j w_j} \leq \frac{\sum_{j=1}^{J} a_j}{\sum_{j=1}^{J} a_j w_j}$. Consequently, it's sufficient if we can prove that for $l \in \{2, 3, ..., J\}$,

$$\frac{\sum\limits_{j=l}^{J} a_j}{\sum\limits_{j=l}^{J} a_j w_j} \leq \frac{\sum\limits_{j=l-1}^{J} a_j}{\sum\limits_{j=l-1}^{J} a_j w_j}. \tag{C.9}$$

Note that

$$\begin{aligned}
(\text{C.9}) \Longleftrightarrow & \frac{\sum\limits_{j=l}^{J} a_j}{\sum\limits_{j=l-1}^{J} a_j} \leq \frac{\sum\limits_{j=l}^{J} a_j w_j}{\sum\limits_{j=l-1}^{J} a_j w_j} \\
\Longleftrightarrow & \, 1 - \frac{\sum\limits_{j=l}^{J} a_j}{\sum\limits_{j=l-1}^{J} a_j} \geq 1 - \frac{\sum\limits_{j=l}^{J} a_j w_j}{\sum\limits_{j=l-1}^{J} a_j w_j} \\
\Longleftrightarrow & \frac{a_{l-1}}{\sum\limits_{j=l-1}^{J} a_j} \geq \frac{a_{l-1} w_{l-1}}{\sum\limits_{j=l-1}^{J} a_j w_j} \\
\Longleftrightarrow & \frac{1}{\sum\limits_{j=l-1}^{J} a_j} \geq \frac{1}{\sum\limits_{j=l-1}^{J} a_j \frac{w_j}{w_{l-1}}}. \qquad (\text{C.10})
\end{aligned}$$

Because w_j is increasing in j for $j \in \{1, 2, ..., J\}$, so (C.10) holds. Thus, we conclude that $\frac{\sum_{j=1}^{J} a_j w_j}{\sum_{j=1}^{J} a_j} \geq \frac{\sum_{j=1}^{J} (a_j + b_j) w_j}{\sum_{j=1}^{J} (a_j + b_j)}$. ■

We then show that for any initial test with a finite number of cut-off points, the objective value will not decrease by arbitrarily adding one more cut-off point from $[\underline{\zeta}, \bar{\zeta}]$.

Let (T, C_T) be an initial test with T cut-off points and $\mathcal{C}_T = \{c_1, c_2, .., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$. The corresponding test outcome set is denoted as Γ and the likelihood of receiving test outcome t given state s_i is $\sigma(t|s_i)$, $t \in \Gamma, s_i \in S$. The total probability of receiving test outcome t for an individual i is denoted as $\lambda_i(t), \forall t \in \mathcal{C}_T$.

The overall expected follow-up probability from CRC patients equals $\sum\limits_{i=1}^{N}\sum\limits_{t\in\Gamma}\sigma(t|1)p_i^0 f_i(t)$, which is also equal to

$$\sum_{i=1}^{N}\sum_{t\in\Gamma\backslash\{0\}}\sigma(t|1)p_i^0 W\Big(p_i^s(t)\Big). \tag{C.11}$$

Suppose we arbitrarily add one cut-off point, $c' \in [\underline{\zeta}, \bar{\zeta}]$ that is not in $\mathcal{C}_T$. Adding one more cut-off point will split one test outcome, say k, to two, denoted by k_1 and k_2. We denote the likelihood of receiving test outcome k_1 and k_2 given state s as $\hat{\sigma}(j|s), j = k_1, k_2$. By construction, $\sigma(k|s) = \hat{\sigma}(k_1|s) + \hat{\sigma}(k_2|s)$. For the new initial test, the overall follow-up probability from CRC patients is

$$\sum_{i=1}^{N}\left(\sum_{t\in\Gamma\backslash\{k\}}\sigma(t|1)f_i(t) + \sum_{j\in\{k_1,k_2\}}\hat{\sigma}(j|1)f_i(j)\right)p_i^0. \tag{C.12}$$

We next compare the values of (C.11) and (C.12) in the following two cases.

Case 1: If $k = 0$,

$$(C.12) = \sum_{i=1}^{N}\sum_{t\in\Gamma\backslash\{0\}}\sigma(t|1)p_i^0 W\Big(p_i^s(t)\Big) + \sum_{i=1}^{N}\sum_{j\in\{k_1,k_2\}}\hat{\sigma}(j|1)p_i^0 f_i(j) \geq (C.11).$$

Case 2: If $k > 0$, by construction, $k_1 > 0$ and $k_2 > 0$.

$$\begin{aligned}(C.12) =& \sum_{i=1}^{N}\sum_{t\in\Gamma\backslash\{0,k\}}\sigma(t|1)p_i^0 W\Big(p_i^s(t)\Big) + \sum_{i=1}^{N}\sum_{j\in\{k_1,k_2\}}\hat{\sigma}(j|1)p_i^0 f_i(j)\\ =& \sum_{i=1}^{N}\sum_{t\in\Gamma\backslash\{0,k\}}\sigma(t|1)p_i^0 W\Big(p_i^s(t)\Big) + \sum_{i=1}^{N}\sum_{j\in\{k_1,k_2\}}\hat{\sigma}(j|1)p_i^0 W\Big(\hat{p}_i^s(j)\Big).\end{aligned} \tag{C.13}$$

Where $\hat{p}_i^s(j)$ is the posterior risk of CRC given the test outcome $j(\in \{k_1, k_2\})$ under the new initial test. Thus, to prove (C.13) $\geq$ (C.11), it's equivalent to prove

$$\sum_{i=1}^{N}\sum_{j\in\{k_1,k_2\}}\hat{\sigma}(j|1)p_i^0 W\Big(\hat{p}_i^s(j)\Big) \geq \sum_{i=1}^{N}\sigma(k|1)p_i^0 W\Big(p_i^s(k)\Big). \tag{C.14}$$

Note that $\sigma(k|s) = \sum\limits_{j\in\{k_1,k_2\}}\hat{\sigma}(j|s), \forall s \in \{0,1\}$. Let the total probability of receiving test outcome j for an individual i under the new test be $\hat{\lambda}_i(j)$. Then,

$\lambda_i(k) = \sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)$. Therefore $p_i^s(k) = \frac{\sigma(k|1)p_i^0}{\lambda_i(k)} = \frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)p_i^0}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)} = \frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^s(j)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)}$. To prove Inequality (C.14), it's sufficient to prove that for each individual i, the following inequality holds:

$$\sum_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)W\Big(\hat{p}_i^s(j)\Big) \geq \sigma(k|1)W\Big(p_i^s(k)\Big) = \sum_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)W\Bigg(\frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^s(j)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)}\Bigg)$$

$$\Longrightarrow \frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)W\Big(\hat{p}_i^s(t)\Big)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\sigma}(t|1)} \geq W\Bigg(\frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^s(j)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)}\Bigg). \tag{C.15}$$

Since $W(\cdot)$ is nondecreasing , then following from Corollary C.1 and Property 1, we have

$$\frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)W\Big(\hat{p}_i^s(j)\Big)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\sigma}(j|1)} \geq \frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)W\Big(\hat{p}_i^s(j)\Big)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)}. \tag{C.16}$$

Following from the convexity of $W(\cdot)$, we have

$$\frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)W\Big(\hat{p}_i^s(j)\Big)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)} \geq W\Bigg(\frac{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^s(j)}{\sum\limits_{j\in\{k_1,k_2\}} \hat{\lambda}_i(j)}\Bigg). \tag{C.17}$$

Combining Inequality (C.16) with Inequality (C.17), we conclude Inequality (C.15) holds. Thus, it's always optimal to add one more cut-off point to the original set of cut-off points.

Before proceeding to prove (2), we first introduce the Bayesian updating process and the property of the continuous test.

Continuous CRC test Under continuous test, an individual receives a test outcome $\zeta \in [\underline{\zeta}, \bar{\zeta}]$, which is the exact f-Hb concentration. The property that CRC patients are more likely to receive worse results than healthy individuals should persist in the continuous test. This suggests that $\frac{h_1(\zeta)}{h_0(\zeta)}$ increases in ζ for $\zeta \in [\underline{\zeta}, \bar{\zeta}]$.

Similar to the ordinal CRC test, for the continuous test which directly reports individuals' f-Hb concentrations, if individual i receives a test outcome ζ, the posterior probability density of having CRC, denoted by $\check{p}_i^s(\zeta)$, becomes:

$$\check{p}_i^s(\zeta) = \frac{h_1(\zeta)p_i^0}{h_1(\zeta)p_i^0 + h_0(\zeta)(1-p_i^0)}, \quad i \in \{1, 2, ...N\},\ \zeta \in [\underline{\zeta}, \bar{\zeta}].$$

Let $\check{f}_i(\zeta)$ denote the follow-up probability density for individual i when adopting the continuous test, and $\check{f}_i(\zeta) = W(\check{p}_i^s(\zeta))$. Given the range of fecal hemoglobin concentration is $[\underline{\zeta}, \bar{\zeta}]$, the overall follow-up probability for all individuals with CRC under continuous test can be written as

$$\sum_{i=1}^{N} \int_{\underline{\zeta}}^{\bar{\zeta}} \check{f}_i(\zeta) h_1(\zeta) p_0^i d\zeta.$$

To bridge the link between continuous and ordinal CRC tests, we establish the following corollary.

Corollary C.2. *For an ordinal CRC test, if policymakers uniformly choose infinitely many cut-off points from $[\underline{\zeta}, \bar{\zeta}]$, the overall follow-up probability from individuals with CRC is the same as that under continuous test.*

Proof. The overall follow-up probability from individuals with CRC under the continuous test can be written as

$$\begin{aligned} &\sum_{i=1}^{N} \int_{\underline{\zeta}}^{\bar{\zeta}} \check{f}_i(\zeta) h_1(\zeta) p_0^i d\zeta \qquad (C.18) \\ =&\sum_{i=1}^{N} \lim_{J\to\infty} \sum_{j=0}^{J} [\check{f}_i(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})j}{J}) h_1(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})j}{J}) p_0^i] \cdot \frac{\bar{\zeta}-\underline{\zeta}}{J}. \qquad (C.19) \end{aligned}$$

Define an ordinal test $(T, \mathcal{C}_T)$, where

$$\mathcal{C}_T \equiv \{c_1, c_2, .., c_T | c_1 = \underline{\zeta}, c_{t+1} = c_t + \frac{\bar{\zeta}-\underline{\zeta}}{T-1}, \forall t = 1, ..., T-1\}.$$

Define $\Delta c(T) = \frac{\bar{\zeta}-\underline{\zeta}}{T-1}$. Then, $c_{t+1} = c_t + \Delta c(T) = \underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1}, \forall t = 1, ..., T-1$. The overall follow-up probability for individuals with CRC under this ordinal test is

$$\begin{aligned} &\sum_{i=1}^{N} \sum_{t\in\Gamma^{\mathcal{C}_T}} f_i(t) \sigma^{\mathcal{C}_T}(t|1) p_i^0 \\ =&\sum_{i=1}^{N} \sum_{t\in\Gamma^{\mathcal{C}_T}\setminus\{0\}} W\Big(p_i^s(t)\Big) \sigma^{\mathcal{C}_T}(t|1) p_i^0 \\ =&\sum_{i=1}^{N} \sum_{t=1}^{T-1} W\left(\frac{(H_1(c_{t+1}) - H_1(c_t))p_i^0}{(H_1(c_{t+1}) - H_1(c_t))p_i^0 + (H_0(c_{t+1}) - H_0(c_t))(1-p_i^0)}\right)\Big(H_1(c_{t+1}) - H_1(c_t)\Big) p_i^0. \end{aligned}$$

Because $W(.)$ is continuous on $\mathbb{R}$, and $\lim_{T\to\infty} \frac{H_s(c_{t+1})-H_s(c_t)}{\Delta c(T)} = \lim_{T\to\infty} \frac{H_s(c_t+\Delta c(T))-H_s(c_t)}{\Delta c(T)} = h_s(c_t)$ for $s \in \{0,1\}, t = 1, ..., T-1$. When T goes to infinity, the follow-up probability from CRC individuals becomes

$$\sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=1}^{T-1} W\Bigg(\frac{(H_1(c_{t+1})-H_1(c_t))p_i^0}{(H_1(c_{t+1})-H_1(c_t))p_i^0+(H_0(c_{t+1})-H_0(c_t))(1-p_i^0)}\Big)\big(H_1(c_{t+1})-H_1(c_t)\Bigg)\, p_i^0$$

$$\sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=1}^{T-1} W\big(\frac{h_1(c_t)p_i^0}{h_1(c_t)p_i^0+h_0(c_t)(1-p_i^0)}\big)h_1(c_t)p_i^0 \cdot \Delta c(T)$$

$$= \sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=1}^{T-1} [\check{f}_i(c_t)h_1(c_t)p_i^0] \cdot \Delta c(T)$$

$$= \sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=1}^{T-1} [\check{f}_i(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1})h_1(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1})p_0^i] \cdot \frac{\bar{\zeta}-\underline{\zeta}}{T-1}$$

=(C.19).

Thus, if the policymaker uniformly choose infinitely many cut-off points from $[\underline{\zeta}, \bar{\zeta}]$, the overall follow-up probability is the same as that under a continuous test. ■

Combining the result from (1) and (2), we complete the proof. □

C.3 Performance Gap

In Theorem 4.2, we demonstrate that for the compliance maximization case, if $W(p_i^s(t))$ is concave in $p_i^s(t)$, then a dichotomous initial test should be selected, with the optimal cut-off being the one with the highest sensitivity. In real practice, $W(\cdot)$ may not be a concave function. We next show that when $W(p_i^s(t))$ is not concave in $p_i^s(t)$, if policymakers select $\underline{\zeta}$ as the only cut-off point, the performance gap between such a FIT and the optimal test is bounded by a finite value. We start by introducing two concepts: concave envelope and the "modulus of non-concavity".

Definition C.1. *(Udell and Boyd, 2013) The tightest concave approximation to $W(\cdot)$, denoted as $\hat{W}(\cdot)$, is a concave envelope of the function $W(\cdot)$, which is defined as the pointwise infimum over all concave functions that are greater than or equal to $W(\cdot)$.*

Definition C.2. *(Aubin, 2007, Udell and Boyd, 2016) Let $W : X \to \mathbb{R}$ be a function defined on a convex set X, and $\hat{W}(\cdot)$ be a concave envelope of the function $W(\cdot)$. We define the "modulus of non-concavity" of $W(\cdot)$, denoted as $\rho(W)$, as below.*

$$\rho(W) = \sup_{x \in X} \big(\hat{W}(x) - W(x)\big)$$

We then show the performance gap between a dichotomous test design with cut-off point value $\underline{\zeta}\}$ and the optimal test design is bounded by a finite value measured by the "modulus of non-concavity".

Theorem C.1. *For the compliance maximization case, the total expected follow-up probability for all the N individuals between test design $(1, \mathcal{C}_1 = \{\underline{\zeta}\})$ and the optimal test design is bounded by $N\rho(W)$, where $\rho(W) = \sup_{x\in[0,1]} \big(\hat{W}(x) - W(x)\big)$ and $\hat{W}(\cdot)$ is a concave envelope of $W(\cdot)$.*

Proof. We first prove that the optimal set of cut-off points includes $\underline{\zeta}$ by showing that we can always get an equal or higher probability of following up by adding $\underline{\zeta}$ into any set of cut-off points.

Let $\mathcal{C}_T = \{c_1, c_2, .., c_T | c_t \in (\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$ be an arbitrary set of $T(> 1)$ cut-off points. The expected follow-up probability for an individual i is $\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) f_i(t)$.

Consider adding $\underline{\zeta}$ as a new cut-off point, and $\mathcal{C}_{T+1} = \{\underline{\zeta}, c_1, c_2, .., c_T | c_t \in (\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$. Let $\check{p}_i^s(t)$ and $\check{f}_i(t)$ denote individual i's posterior risk and follow-up probability after receiving test outcome t under this new test, respectively. Note for $1 \le t \le T$, because the cut-off point in $\mathcal{C}_T$ remains in $\mathcal{C}_{T+1}$, and the test outcome labels are simply shifted by 1, we have $\sigma^{\mathcal{C}_{T+1}}(t+1|1) = \sigma^{\mathcal{C}_T}(t|1)$, thus $\lambda_i^{\mathcal{C}_{T+1}}(t+1) = \sigma^{\mathcal{C}_{T+1}}(t+1|0)(1-p_i^0) + \sigma^{\mathcal{C}_{T+1}}(t+1|1)p_i^0 = \sigma^{\mathcal{C}_T}(t|0)(1-p_i^0) + \sigma^{\mathcal{C}_T}(t|1)p_i^0 = \lambda_i^{\mathcal{C}_T}(t)$, $\check{p}_i^s(t+1) = \frac{p_i^0}{\frac{\sigma^{\mathcal{C}_{T+1}}(t+1|0)}{\sigma^{\mathcal{C}_{T+1}}(t+1|1)}(1-p_i^0)+p_i^0} = \frac{p_i^0}{\frac{\sigma^{\mathcal{C}_T}(t|0)}{\sigma^{\mathcal{C}_T}(t|1)}(1-p_i^0)+p_i^0} = p_i^s(t)$ and $\check{f}_i(t+1) = f_i(t)$.

Therefore, the expected follow-up probability for individual i under the new test design

$$\sum_{t=1}^{T+1} \lambda_i^{\mathcal{C}_{T+1}}(t)\check{f}_i(t) = \lambda_i^{\mathcal{C}_{T+1}}(1)\check{f}_i(1) + \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) f_i(t) \ge \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) f_i(t).$$

We conclude that the optimal set of cut-off points includes $\underline{\zeta}$.

Let $\mathcal{C}_T^* = \{c_1, c_2, .., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], \underline{\zeta} = c_1 < c_2 < ... < c_T\}$ denote the optimal cut-off set with T ($T > 1$) cut-off points. For an individual i, the expected probability of following up under the optimal test design $(T, \mathcal{C}_T)$ equals

$$\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t) W(p_i^s(t)).$$

Consider the dichotomous test with cut-off point $\mathcal{C}_1 = \{\underline{\zeta}\}$. Let $\hat{p}_i^s(t)$ denote individual i's posterior risk after receiving test outcome t under the dichotomous test. Thus, for individual i, the expected probability of following up under equals

$$\lambda_i^{\mathcal{C}_1}(1) W(\hat{p}_i^s(1)). \tag{C.20}$$

Since $\lambda_i^{\mathcal{C}_1}(1) = \sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t)$, $\hat{p}_i^s(1) = \frac{\sigma^{\mathcal{C}_1}(1|1) p_i^0}{\lambda_i^{\mathcal{C}_1}(1)} = \frac{\sum_{t=1}^{T} \sigma^{\mathcal{C}_T^*}(t|1) p_i^0}{\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t)} = \frac{\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t) p_i^s(t)}{\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t)}$. Consequently, (C.20) equals $\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t) W\Big(\frac{\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t) p_i^s(t)}{\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T^*}(t)}\Big)$.

Therefore, the performance gap between test design $(T, \mathcal{C}_T^*)$ and $(1, \mathcal{C}_1)$ equals

$$\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) W(p_i^s(t)) - \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) W\Big(\frac{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t)}\Big). \tag{C.21}$$

Let $\hat{W}(\cdot)$ be a concave envelope of $W(\cdot)$. Given the concavity of $\hat{W}(\cdot)$, we have

$$\sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) \hat{W}\Big(\frac{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t)}\Big) \geq \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) \hat{W}(p_i^s(t)) \geq \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) W(p_i^s(t)). \tag{C.22}$$

Combining inequality (C.21) with (C.22), we have

$$\text{(C.21)} \leq \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) \hat{W}\Big(\frac{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t)}\Big) - \sum_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) W\Big(\frac{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) p_i^s(t)}{\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t)}\Big). \tag{C.23}$$

Note that $\sum\limits_{t=1}^{T} \lambda_i^{\mathcal{C}_T}(t) \leq 1$. Thus, performance gap for all the N individuals is bounded by $\sum\limits_{i=1}^{N} \sup\limits_{x \in [0,1]} \big(\hat{W}(x) - W(x)\big) = N\rho(W)$. □

C.4 Parameter Estimation

C.4.1 Utility Estimation

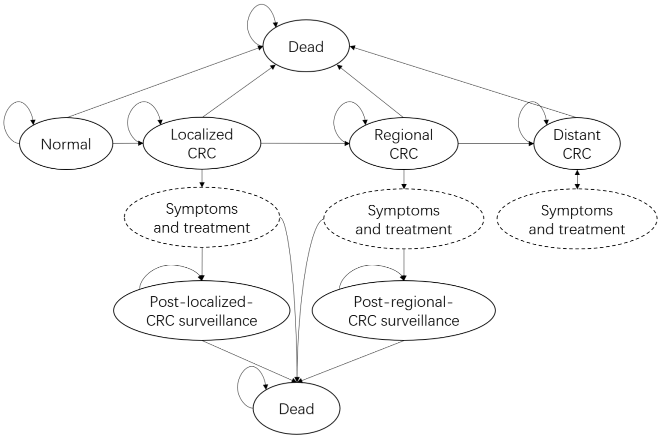


Figure C.1: CRC progression

We refer to the model constructed by Ladabaum et al. (2001) for the natural history of CRC in the average-risk population without CRC screening (Figure C.1). Individuals transition between principal health states of normal, localized CRC, regional CRC or distant CRC, and dead in 1-year cycles. The states for patients who survived CRC treatment (post-localized-CRC surveillance and post-regional-CRC surveillance) are also shown. As presented in Figure C.1, individuals could remain in the current state or transit to other states at the end of every cycle. The ovals without solid borders represent intermediate states associated with the symptomatic presentation of CRC. Specifically, individuals with distant CRC are symptomatic, whereas for individuals with localized or regional CRC, the associated symptoms may present with certain probabilities. We suppose that individuals with symptomatic presentation will consult doctors and receive the treatment. Typically, normal individuals may develop localized CRC. CRC cases progress from localized to regional (2 years in each stage) to distant unless symptoms lead to diagnosis and treatment. If individuals are diagnosed and treated, they will enter post-cancer surveillance.

Because the expected remaining life years vary widely among individuals with different ages, genders and CRC states, and the second-stage screening could detect CRC at an earlier stage when the treatment is usually more successful. Therefore, individuals' expected remaining QALYs differ by age, gender, CRC state and follow-up decision. Next, we present the estimation of the expected remaining QALYs for individuals with different aforementioned attributes. We use $u_{k,g}(s, a)$ to denote the expected remaining QALYs for individuals with age k, gender g, state s and follow-up action a. We next decribe the calculation of $u_{k,g}(s, a)$.

1. $u_{k,g}(0, 0)$: Following from Ladabaum et al. (2001), we suppose the quality of life for individuals without CRC is 1, and we use the expected remaining life years from the complete life tables for the Singapore resident population (Department of Statistics 2019) as a proxy of the expected remaining QALYs.

2. $u_{k,g}(0, 1)$: Notably, the QALYs loss from a colonoscopy is incorporated in the perceived cost of the follow-up action; therefore, we have $u_{k,g}(0, 1) = u_{k,g}(0, 0)$.

3. $u_{k,g}(1, 1)$: For asymptomatic CRC patients who follow up with the second stage tests, they will be diagnosed and treated. For those who survive CRC treatment, they will have high risks of post-treatment mortality in the initial 5 years; after the initial 5 years of post-cancer surveillance, the recurrence of CRC becomes less common (Caso et al., 2020), and we suppose the expected remaining QALYs of survivors are the same as those of normal individuals with the same age and gender. Then, we can derive $u_{k,g}(1, 1)$ according to the collected parameters.

4. $u_{k,g}(1, 0)$: For asymptomatic CRC patients who don't follow-up with the second-stage tests, they will be detected if the associated symptoms appear; otherwise, they will progress all the way to distant CRC or dead. In line with Ladabaum et al. (2001), we suppose there is a loss of quality for CRC patients who do not receive treatment. Besides, these patients may experience natural death (mortality not from CRC) during the process.

The values of all the aforementioned parameters are listed in Table C.2.

Parameters		
The expected remaining QALYs for healthy individuals	age-gender-specific	Department of Statistics (2019)
Distribution of CRC stages	stage-specific	Howlader et al. (2018)
Symptomatic presentation of localized CRC, %	22/year over 2 years	Ladabaum and Mannalithara (2016)
Symptomatic presentation of regional CRC, %	40/year over 2 years	Ladabaum and Mannalithara (2016)
Mortality rate from treated localized CRC, %	1.74/year in first 5 years	Ladabaum and Mannalithara (2016)
Mortality rate from treated regional CRC, %	8.6/year in first 5 years	Ladabaum and Mannalithara (2016)
Mean survival from distant CRC, *year*	1.9	Ladabaum and Mannalithara (2016)
Mortality rate from CRC treatment, %	2	Ladabaum and Mannalithara (2016)
Mortality rate not from CRC, %	age-gender-specific	Department of Statistics (2019)
Qualiy of life of localized CRC	0.90	Ladabaum and Mannalithara (2016)
Qualiy of life of regional CRC	0.80	Ladabaum and Mannalithara (2016)
Qualiy of life of distant CRC	0.76	Ladabaum and Mannalithara (2016)

Table C.2: Input parameters

C.4.2 Utility Functional Form Estimation

Variable selection

In this section, we introduce the variable selection process of δ-features and d-features.

δ-features and d-features To examine determinants and barriers towards CRC adherence behavior, we first divide the data points into two groups based on whether individuals followed up with the second-stage tests. We perform univariate statistical comparisons of the two groups by applying Welch's t-test and select the variables with p-value less than 0.1. The comparison result is presented in Table C.3. Then we adopt all significant variables as candidate variables for δ.

Besides, we inquired individuals about the factors they were concerned about when deciding whether to follow up in our survey. All these variables and the QALYs loss from a colonoscopy are initially selected as d-features. We then perform 3-fold cross-validation to choose the combinations of variables that produce the highest average AUC in predicting individuals' follow-up behaviors. The selected variable combinations are finally adopted as δ-features and d-features.

Variable	Mean of Group 0	Mean of Group 1	p-value
Age	62.68	64.21	< 0.1
Married	0.74	0.86	< 0.1
Own apartment	0.88	0.96	< 0.1
Own more than one apartment	0.14	0.24	< 0.1
Tobacco consumption	4.29	18.87	< 0.1
Have private insurance	0.84	0.71	< 0.05
Screening knowledge	2.18	2.57	< 0.05
Colonoscopy knowledge	0.89	1.12	< 0.01
Knowledge about CRC incidence rate	0.89	0.77	< 0.05
Real risk of having CRC	0.001	0.002	< 0.1
Frequency of receiving FIT	2.67	3.27	< 0.001

Table C.3: Univariate statistical comparisons

Estimation of the QALYs loss from a colonoscopy

We consider the possibility of colonoscopy perforation and perforation-related mortality rate. Let $\omega_{k,g}(s)$ denote the QALYs loss from a colonoscopy for individuals with age k, gender g and state s, then

$$\omega_{k,g}(s) = u_{k,g}(s,1) \times p(\text{perforation probability}) \times p(\text{perforation-related mortality}).$$

MLE estimation

We use $\boldsymbol{z}_i \equiv \{z_i^1, z_i^2, .., z_i^J\}$ and $\boldsymbol{y}_i \equiv \{y_i^1, y_i^2, ..., y_i^R\}$ to denote δ-features and d-features, respectively. Where y_i^R represents the QALYs loss from a colonoscopy for individual i.

In particular, we propose a multivariate linear model for δ_i, which takes the form:

$$\delta_i = \alpha^0 + \sum_{j=1}^{J} \alpha^j z_i^j.$$

Where α^0 denotes the constant and α^j $(j \in \{1, 2, ..., J\})$ denotes the coefficient for z_i^j.

In addition, we suppose $d_i(s_i = 0)$ and $d_i(s_i = 1)$ take the following functional forms:

$$d_i(0) = \beta^0 + \sum_{r=1}^{R} \beta^r y_i^r,$$

$$d_i(1) = \gamma^0 + \sum_{r=1}^{R} \gamma^r y_i^r.$$

Where β^0 and γ^0 are constants, β^r and γ^r denote the coefficients of y_i^r in the two respective functions, $r \in \{1, 2, ..., R\}$.

Note individuals will experience a loss of utility from the colonoscopy; therefore, we suppose $\beta^R, \gamma^R \geq 0$. For individuals who don't believe in having the disease, they will expend a utility loss from the follow-up action. Thus, we assume that $\beta^r \geq 0$ if $r \in \{1, 2, ..., R-1\}$. Conversely, for individuals who believe they have the disease, they can benefit from the subsequent diagnosis and treatment. Therefore, we assume that $\gamma^r \leq 0$ if $r \in \{1, 2, ..., R-1\}$.

Let e_i denote individual i's actual follow-up decision. For individual i, the likelihood of choosing action $e_i \in \{0, 1\}$ after receiving a positive test outcome ($t = 1$) equals

$$P_i(a_i(1) = e_i) = \frac{e^{U_i(a_i(1)=e_i)}}{e^{U_i(a_i(1)=0)} + e^{U_i(a_i(1)=1)}}.$$

Our purpose is to derive the parameters that maximize the log-likelihood function subject to the constraints of δ_i, β^r and γ^r. We formulate the parameter estimation problem as the following convex optimization problem:

$$\begin{aligned} \max_{\alpha^j, \beta^r, \gamma^r} \; & \sum_{i=1}^{N} \log(P_i(a_i(1) = e_i)) \\ \text{s.t. } & 0 \leq \delta_i \leq 1 \quad && i \in \{1, 2, ..., N\} \\ & \beta^r \geq 0 && r \in \{1, 2, ..., R\} \\ & \gamma^r \leq 0 && r \in \{1, 2, ..., R-1\} \\ & \gamma^R \geq 0 \end{aligned}$$

The descriptions of the estimated parameters are shown in Table C.4.

δ-features ($\boldsymbol{z}$)	$\boldsymbol{\alpha}$	
Age	0.02	
Have private insurance	0.07	
Screening knowledge	-0.22	
Colonoscopy knowledge	-0.06	
Frequency of receiving FIT	-0.08	
d-features ($\boldsymbol{y}$)	$\boldsymbol{\beta}$	γ
Concerns about the medical history	0.00	-100.29
Concerns about age	0.31	-194.10
Trust in doctors	0.02	-5180.09
Want to know health condition	0.00	-320.64
Concerns about the price of a colonoscopy	0.29	0.00
Support from family	46.87	-12857.20
Disutility from a colonoscopy	0.00	7.69

Table C.4: Descriptions of the estimated parameters

C.4.3 Logistic Regression Result

We perform stepwise logistic regression in the balanced dataset, and the regression result is shown in Table C.5.

C.4.4 Construction of the Discretized Candidate Set

The candidate set of cut-off points is selected to (1) cover the commonly adopted cut-off points in practice and also to (2) ensure that the initial test kits are MLR-feasible. Firstly, according to numerous systematic reviews (e.g., Lee et al. (2014), Robertson et al. (2017)), thresholds greater than 10 $\mu g/g$ and smaller than 100 ug/g are commonly adopted in real practice. Therefore, we choose to construct a discretized set from 10 $\mu g/g$ to 100 $\mu g/g$. Specifically, according to Figure 4.1, when the cut-off value increases from 10 $\mu g/g$ to 40 $\mu g/g$, the corresponding changes in test outcome probabilities are relatively large, which requires more granular increments. When the cut-off value is greater than 40, the changes are subtle. The candidate set, therefore, includes cut-off values from 10 $\mu g/g$ to 40 $\mu g/g$ in steps of 1 and from 40 $\mu g/g$ to 100 $\mu g/g$ in steps of 5.

	Coefficient	Standard error	p-value
δ-features			
Own apartment	0.62	0.22	< 0.01
Have private insurance	-1.29	0.30	< 0.01
Screening knowledge	1.17	0.25	< 0.01
Colonoscopy knowledge	0.95	0.33	< 0.01
Frequency of receiving FIT	-0.64	0.30	< 0.05
d-features			
Concerns about the medical history	2.55	0.72	< 0.01
Concerns about age	2.04	0.53	< 0.01
Trust in doctors	1.89	0.88	< 0.05
Want to know health condition	2.99	0.53	< 0.01
Concerns about the price of a colonoscopy	-1.64	0.66	< 0.05
Support from family	1.84	1.11	< 0.1
Constant	-1.14	0.36	< 0.01
No. of observations	290		
Pseudo R-squared	0.58		
Prob > Chi-square	<0.01		

Table C.5: Logistic regression result

We next show that if the initial test kit are MLR-feasible, the two-sample dichotomous FIT initial tests are also MLR-feasible. To ensure the two-sample dichotomous FIT is MLR-feasible, it's sufficient if we have $\frac{1-\sigma^2(0|1)}{1-\sigma^2(0|0)} > \frac{\sigma^2(0|1)}{\sigma^2(0|0)}$. The condition is equivalent to $\sigma(0|0) > \sigma(0|1)$. We next show this condition holds when the initial test kit are MLR-feasible by contradiction. Suppose $\frac{\sigma(0|0)}{\sigma(0|1)} \geq 1$, then $\frac{\sigma(1|0)}{\sigma(1|1)} > \frac{\sigma(0|0)}{\sigma(0|1)} \geq 1$. Therefore, $1 = \sigma(1|0) + \sigma(0|0) > \sigma(1|1) + \sigma(0|1) = 1$, which can not hold. Thus, we conclude that the the two-sample dichotomous FIT initial tests are MLR-feasible if the initial test kit are MLR-feasible.

C.5 Other Supplementary Numerical Results

C.5.1 One-sample FIT Design

For one-sample FIT, the features in the two clusters are presented in Table C.6. Since CRC risk serves as an input in the utility model and varies by gender and age, we also present the description of gender in the two clusters.

	Cluster 1	Cluster 2	p-value
The optimal cut-off point, $\mu g/g$	31	39	-
δ-features			
Age	67.37	59.11	< 0.01
Have private insurance	0.67	0.81	< 0.01
Screening knowledge	2.44	2.43	0.77
Colonoscopy knowledge	0.88	0.89	0.55
Frequency of receiving FIT	2.00	1.92	0.11
d-features			
Medical history	0.19	0.18	0.47
Age	0.33	0.27	< 0.01
Trust doctor	0.17	0.19	0.14
Want to know health condition	0.28	0.29	0.58
Price of a colonoscopy	0.22	0.24	0.18
Support from family	0.06	0.07	0.22
Other features			
δ	0.62	0.62	0.68
Risk	0.0023	0.0007	< 0.01
Male	0.81	0.25	< 0.01

Table C.6: Variable description in different clusters of customized dichotomous test when $\tau = 0.003$ and $L = 2$ (two-sample FIT)

The optimal design of three-type customized dichotomous test (i.e., $L = 3$) with the adjusting term τ equals 0.006 is presented in Table C.7.

For the interpretable clustering dichotomous test, the optimal cut-off value and the expected number of positive results and follow-ups are presented in Table C.8.

C.5.2 Two-sample FIT Design

We present the optimal cut-off point and other related performance metrics for the two-sample FIT design in the following. The findings and insights under different

	Cluster 1	Cluster 2	Cluster 3	Total	Cluster 1$^{\&}$	Cluster 2$^{\&}$	Cluster 3$^{\&}$	Total$^{\&}$	Current practice
Number of individuals*	884.60	3137.88	5977.52	10000	884.60	3137.88	5977.52	10000	10000
Expected number of CRC patients*	3.01	6.17	5.05	14.23	3.01	6.17	5.05	14.23	14.23
The optimal cut-off point, $\mu g/g$	28	32	40	-	-	-	-	33	20
Sensitivity, %	78.24	76.68	73.86	-	-	-	-	76.30	81.77
Specificity, %	98.68	99.21	99.70	-	-	-	-	99.30	96.22
Expected number of positives*	13.96	29.57	21.43	64.96	8.47	26.62	45.63	80.72	389.37
Expected number of positives from CRC patients*	2.35	4.74	3.73	10.82	2.30	4.71	3.85	10.86	11.64
Expected number of positives from healthy individuals*	11.61	24.83	17.70	54.14	6.17	21.91	41.78	69.86	377.73
Expected colonoscopy demand*	13.96	29.55	21.31	64.82	8.47	26.62	43.50	78.59	227.01
Expected number of follow-ups from CRC patients*	2.35	4.73	3.73	10.81	2.30	4.71	3.72	10.73	7.21
Expected number of follow-ups from healthy individuals*	11.61	24.82	17.58	54.01	6.17	21.91	39.78	67.86	219.80

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.
& The results are from the universal test when the adjusting term τ equals 0.

Table C.7: The optimal customized dichotomous test when $\tau = 0.006$ and $L = 3$ (one-sample FIT)

	Cluster 1	Cluster 2	Total	Total$^{\&}$	Current practice
Number of individuals*	3782.91	6217.09	10000	10000	10000
Expected number of CRC patients*	8.74	5.49	14.23	14.23	14.23
The optimal cut-off point, $\mu g/g$	31	39	-	33	20
Sensitivity, %	77.06	74.20	-	76.30	81.77
Specificity, %	99.10	99.67	-	99.30	96.22
Expected number of positives*	40.68	24.84	65.52	80.72	389.37
Expected number of positives from CRC patients*	6.74	4.07	10.81	10.86	11.64
Expected number of positives from healthy individuals*	33.94	20.77	54.71	69.86	377.73
Expected colonoscopy demand*	40.63	24.69	65.32	78.59	227.01
Expected number of follow-ups from CRC patients*	6.73	4.07	10.80	10.73	7.21
Expected number of follow-ups from healthy individuals*	33.90	20.62	54.52	67.86	219.80

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.
& The results are from the universal test when the adjusting term τ equals 0.

Table C.8: The optimal interpretable clustering dichotomous test when $\tau = 0.007$ (one-sample FIT)

screening protocols are the same as those in the one-sample FIT; therefore, we don't discuss them in detail.

Universal dichotomous test

As shown in Table C.9, the cut-off point that induces the highest follow-ups from CRC patients is 39 $\mu g/g$, which is considerably different from the practice (20 $\mu g/g$). In addition, it is higher than the optimal cut-off value (33 $\mu g/g$) in one-sample FIT (Table C.9). This is because that in the two-sample test, a positive result is reported once the hemoglobin concentration in any sample is greater than the predefined cut-off value. Thus, compared with one-sample test, it comes with a higher false-positive rate and lower false-negative rate under the same cut-off value.

Customized dichotomous test We further explore the performance of pro-

Adjusting term (τ)	-1	-0.05	-0.02	-0.01	0	0.005	0.01	0.05	1
The optimal cut-off point, $\mu g/g$	13	17	32	35	39	45	50	65	90
Sensitivity, %	97.91	97.21	94.56	94.04	93.34	92.31	91.46	88.98	85.05
Specificity, %	80.78	88.89	98.42	98.91	99.33	99.67	99.82	99.97	100.00
Expected number of positives *	1933.41	1123.21	171.19	122.08	79.97	45.66	31.11	15.95	12.33
Expected number of positives from CRC patients*	13.93	13.83	13.46	13.38	13.28	13.14	13.02	12.66	12.11
Expected number of positives from healthy individuals*	1919.48	1109.38	157.73	108.70	66.69	32.52	18.09	3.29	0.22
Expected colonoscopy demand*	392.39	387.11	151.01	115.82	78.99	45.53	31.07	15.95	12.33
Expected number of follow-ups from CRC patients*	3.35	5.22	12.52	13.04	13.23	13.12	13.02	12.66	12.11
Expected number of follow-ups from healthy individuals*	389.04	381.89	138.49	102.78	65.76	32.41	18.05	3.29	0.22

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.

Table C.9: The optimal universal dichotomous test (two-sample FIT)

moting customized dischotomous test to two subpopulations (i.e, $L = 2$). The results are presented in Table C.10 with τ equals 0.0035.

	Cluster 1	Cluster 2	Total	Cluster 1$^{\&}$	Cluster 2$^{\&}$	Total$^{\&}$	Current practice
Number of individuals*	4022.48	5977.52	10000	4022.48	5977.52	10000	10000
Expected number of CRC patients*	9.18	5.05	14.23	9.18	5.05	14.23	14.23
The optimal cut-off point, $\mu g/g$	37	45	-	-	-	39	20
Sensitivity, %	93.69	92.31	-	-	-	93.34	96.68
Specificity, %	99.15	99.67	-	-	-	99.33	92.58
Expected number of positives*	42.79	24.11	66.90	35.37	44.60	79.97	754.93
Expected number of positives from CRC patients*	8.61	4.66	13.27	8.57	4.71	13.28	13.76
Expected number of positives from healthy individuals*	34.18	19.45	53.63	26.80	39.89	66.69	741.17
Expected colonoscopy demand*	42.78	24.00	66.78	35.38	43.61	78.99	336.17
Expected number of follow-ups from CRC patients*	8.60	4.66	13.26	8.58	4.65	13.23	6.47
Expected number of follow-ups from healthy individuals*	34.18	19.34	53.52	26.80	38.96	65.76	329.70

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.
$^{\&}$ The results are from the universal test when the adjusting term τ equals 0.

Table C.10: The optimal customized dichotomous test when $\tau = 0.0035$ and $L = 2$ (two-sample FIT)

Interpretable clustering dichotomous test Similarly, we apply the interpretable decision tree to predict individuals' clusters generated from the customized test design result. The decision tree obtained is exacly the same as the one in on-sample FIT case.

We derive the optimal cut-off value and the other metrics with the adjusting term τ equals 0.0035 in Table C.11.

C.6 Model Extension

In this section, we extend our model to study both colorectal polyps and CRC detection.

	Cluster 1	Cluster 2	Total	Total$^{\&}$	Current practice
Number of individuals*	3782.91	6217.19	10000	10000	10000
Expected number of CRC patients*	8.74	5.49	14.23	14.23	14.23
The optimal cut-off point, $\mu g/g$	37	45	-	39	20
Sensitivity, %	93.69	92.31	-	93.34	96.68
Specificity, %	99.15	99.67	-	99.33	92.58
Expected number of positives*	40.34	25.30	65.64	79.97	754.93
Expected number of positives from CRC patients*	8.19	5.07	13.26	13.28	13.76
Expected number of positives from healthy individuals*	32.15	20.23	52.38	66.69	741.17
Expected colonoscopy demand*	40.33	25.19	65.31	78.99	336.17
Expected number of follow-ups from CRC patients*	8.19	5.07	13.25	13.23	6.47
Expected number of follow-ups from healthy individuals*	32.14	20.12	52.26	65.76	329.70

* The population base is assumed to be 10,000, of which 14.23 individuals have CRC and 9985.77 do not.

& The results are from the universal test when the adjusting term τ equals 0.

Table C.11: The optimal decision-tree-based clustering when $\tau = 0.0035$ (two-sample FIT)

Initial test: Let $\hat{s}$ denote individuals' state $\hat{s} \in \hat{S} = \{0, 1, 2\}$. Where $\hat{s} = 1$ represents a heath state with colorectal polyps and $\hat{s} = 2$ indicates a heath state with CRC. ζ is used to denote f-Hb concentration level tested from FIT. Let $\hat{H}_0(\hat{h}_0)$, $\hat{H}_1(\hat{h}_1)$ and $\hat{H}_2(\hat{h}_2)$ denote the cumulative distribution function (probability density function) of f-Hb concentration for healthy individuals, individuals with colorectal polyps and CRC, respectively. The range of f-Hb concentration is $[\underline{\zeta}, \overline{\zeta}]$. We assume that $\hat{h}_0(\zeta), \hat{h}_1(\zeta), \hat{h}_2(\zeta)$ and are continuous in ζ for $\zeta \in [\underline{\zeta}, \overline{\zeta}]$.

A test design $(T, \mathcal{C}_T)$ gives rise to $T + 1$ test outcomes in set $\Gamma^{\mathcal{C}} = \{0, 1, 2, ..., T\}$. We define $\sigma^{\mathcal{C}_T}(t|\hat{s})$ as the likelihood of receiving test outcome t if individual's state is $\hat{s}$, such that for $\hat{s} \in \{0, 1, 2\}$, $\sigma^{\mathcal{C}_T}(0|\hat{s}) = \hat{H}_{\hat{s}}(c_1)$, $\sigma^{\mathcal{C}_T}(t|\hat{s}) = \hat{H}_{\hat{s}}(c_{t+1}) - \hat{H}_{\hat{s}}(c_t)$ if $t \in \{1, 2, ..., T - 1\}$ and $\sigma^{\mathcal{C}_T}(T|\hat{s}) = 1 - \hat{H}_{\hat{s}}(c_T)$. In real practice, the initial cancer screening test should possess a property that individuals with CRC are more likely to receive severe test outcomes than individuals with polyps. Similarly, we define an initial test that possesses this nice property as a "MLR-feasible" initial test.

Property 2. *An initial test $(T, \mathcal{C}_T)$ is MLR-feasible if $\frac{\sigma^{\mathcal{C}_T}(t|2)}{\sigma^{\mathcal{C}_T}(t|1)}$ and $\frac{\sigma^{\mathcal{C}_T}(t|1)}{\sigma^{\mathcal{C}_T}(t|0)}$ are increasing in t, $t \in \Gamma^{\mathcal{C}}$.*

Indiviudal's follow-up problem: For each individual i, the prior risk of developing polyps and CRC are denoted by p_i^1 and p_i^2, where $i \in \{1, 2, ..., N\}$. After a participant obtains a test outcome $t \in \Gamma$, the risk of having polyps or CRC is updated to $p_i^{s1}(t)$ and $p_i^{s2}(t)$, following Bayesian updating process. The total probability of receiving test outcome t becomes $\hat{\lambda}_i(t) = \sigma(t|0)(1 - p_i^1 - p_i^2) + \sigma(t|1)p_i^1 + \sigma(t|2)p_i^2$, $t \in \Gamma, i \in \{1, 2, ..., N\}$. We have

$$p_i^{s1}(t) = \frac{\sigma(t|1)p_i^1}{\lambda_i(t)} = \frac{p_i^1}{1 + \frac{\sigma(t|2)}{\sigma(t|1)}p_i^2 + \frac{\sigma(t|0)}{\sigma(t|1)}(1 - p_i^1 - p_i^2)}, \quad i \in \{1, 2, ...N\},\ t \in \Gamma;$$

$$p_i^{s2}(t) = \frac{\sigma(t|2)p_i^2}{\lambda_i(t)} = \frac{p_i^2}{1 + \frac{\sigma(t|1)}{\sigma(t|2)}p_i^1 + \frac{\sigma(t|0)}{\sigma(t|2)}(1 - p_i^1 - p_i^2)}, \quad i \in \{1, 2, ...N\},\ t \in \Gamma.$$

According to Property 2, given an "MLR-feasible" initial test, $p_i^{s2}(t)$ is always increasing in t, while $p_i^{s1}(t)$ may not have the same trend. Intuitively speaking, with the increase of t, the posterior risk of having polyps will firstly increase. However, as t becomes larger, individuals will have a higher risk of having CRC, and the posterior risk of having polyps will decrease since the posterior risk of having cancer will dominate.

We use $a_i(t) \in \{0, 1\}$ as individual i 's follow-up decision, and $u_i(\hat{s}_i, a_i(t))$ as the individual's utility function. Let $\pi_i^{s1}(t)$ and $\pi_i^{s2}(t)$ represent individual's subjective belief of having polyps and CRC after receiving a test outcome t from the initial test. The total expected utility of participant i under action $a_i(t)$ is given as follows.

$$\begin{aligned} U_i(a_i(t) = 0) =& \delta_i E[u_i(s_i, a_i(t) = 0)|(\pi_i^{s1}(t), \pi_i^{s2}(t))] + (1 - \delta_i)E[u_i(s_i, a_i(t) = 0)|(p_i^{s1}(t), p_i^{s2}(t))] + \epsilon_i^0, \\ U_i(a_i(t) = 1) =& \delta_i E[u_i(s_i, a_i(t) = 1)|(\pi_i^{s1}(t), \pi_i^{s2}(t))] + (1 - \delta_i)E[u_i(s_i, a_i(t) = 1)|(p_i^{s1}(t), p_i^{s2}(t))] \\ & - E[d_i(s_i)|(\pi_i^{s1}(t), \pi_i^{s2}(t))] + \epsilon_i^1. \end{aligned}$$

We assume that indiviudals will not follow up if they receive the best test outcome. Thus, the conditions which decide individual i's follow-up behavior can be written as follows:

$$a_i(t) = \begin{cases} 1 & \text{if } t \neq 0 \text{ and } U_i(a_i(t) = 1) > U_i(a_i(t) = 0), \\ 0 & \text{if } t = 0 \text{ or } U_i(a_i(t) = 0) \geq U_i(a_i(t) = 1). \end{cases}$$

Similar to the base model, we assume that for any individual i with test outcome t, individual's subjective belief $\pi_i^{s1}(t)$ ($\pi_i^{s2}(t)$) is a function of $p_i^{s1}(t)$ ($p_i^{s2}(t)$). Thus, the probability of follow-up becomes a function of $p_i^{s1}(t)$ and $p_i^{s2}(t)$ if $t \neq 0$. We use $W\big(p_i^{s1}(t), p_i^{s2}(t)\big)$ to denote this function. Therefore, we have the following result:

$$f_i(t) = \begin{cases} 0 & \text{if } t = 0, \\ W\big(p_i^{s1}(t), p_i^{s2}(t)\big) & \text{otherwise.} \end{cases} \tag{C.24}$$

Health system test design problem: Given $\mathcal{C}_T$ is the selected set of cut-off points with the corresponding test outcome set $\Gamma^{\mathcal{C}_T}$, we formulate the health system's problem as follows.

$$\begin{aligned} \max_{T,\mathcal{C}_T} \quad & \sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T}} [f_i(t)\sigma^{\mathcal{C}_T}(t|2)p_i^2 + \tau_1 f_i(t)\sigma^{\mathcal{C}_T}(t|1)p_i^1 - \tau_0 f_i(t)\sigma^{\mathcal{C}_T}(t|0)(1 - p_i^1 - p_i^2)] \\ s.t. \quad & f_i(t) \text{ satisfying } (C.24) \end{aligned} \tag{C.25}$$

Specifically, the first term in the objective function represents the follow-up rate from individuals with CRC. It is controversial whether the screening policy should recommend individuals with polyps to follow up with a colonoscopy or not. Western countries remove all colorectal polyps, except for rectosigmoid hyperplastic polyps ≤ 5 mm in size. However, in Asian countries, the treatment strategy for colorectal serrated polyps is still not established (Sano et al., 2020). In addition, Vleugels et al. (2017) showed that the estimated progression rate of small (6-9 mm) colorectal polyps to advanced adenoma or CRC is very low, so individuals with diminutive and small polyps may experience potential "harm" of polypectomy in the colonoscopy. Therefore, we introduce a multiplier $\tau_1 \in [0, 1]$ in front of the second term , which indicates the percentage of such individuals that the health system encourages to follow up. In addition, we use τ_0 to denote the "cost" of a colonoscopy demand from healthy individuals.

We first explore the optimal structure of the initial tests in the compliance maximization case ($\tau_1 = 1$ and $\tau_0 = -1$) and effectiveness maximization case ($\tau_1 \in [0, 1]$ and $\tau_0 = 0$), respectively. We have the following results.

Theorem C.2. *For the compliance maximization case, if $W\big(p_i^{s1}(t), p_i^{s2}(t)\big)$ is concave in $p_i^{s1}(t)$ and $p_i^{s2}(t)$, a dichotomous initial test with cut-off point value $\underline{\zeta}$ is optimal.*

Proof. Proof of Theorem C.2. We prove this theorem by two steps: (1) we first prove that when $W(\cdot)$ is concave in $p_i^{s1}(t)$ and $p_i^{s2}(t)$, a dichotomous test is optimal; (2) we then prove that the optimal cut-off point value is $\underline{\zeta}$.

Let $(T, \mathcal{C}_T)$ be an initial test with more than one cut-off point, where $T > 1$ and $\mathcal{C}_T \equiv \{c_1, c_2, .., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$. The corresponding test outcome set is $\Gamma^{\mathcal{C}_T} = \{0, 1, 2, ..., T\}$ and the likelihood of receiving test outcome t under state $\hat{s}_i$ is $\sigma^{\mathcal{C}_T}(t|\hat{s}_i)$, $t \in \Gamma^{\mathcal{C}_T}, \hat{s}_i \in \hat{S}$. Let $\lambda_i^{\mathcal{C}_T}(t)$ denote the total probability of receiving test outcome t, i.e., $\hat{\lambda}_i^{\mathcal{C}_T}(t) = \sigma^{\mathcal{C}_T}(t|0)(1 - p_i^1 - p_i^2) + \sigma^{\mathcal{C}_T}(t|1)p_i^1 + \sigma^{\mathcal{C}_T}(t|2)p_i^2$. Then for any individual i, the expected probability of follow-up under $\tau_1 = 1$ and $\tau_0 = -1$ can be written as

$$\sum_{t \in \Gamma^{\mathcal{C}_T}} \hat{\lambda}_i^{\mathcal{C}_T}(t) f_i(t) = \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) W\big(p_i^{s1}(t), p_i^{s2}(t)\big). \tag{C.26}$$

We next propose another initial test which contains only one cut-off point c_1. That is $T = 1$, $\mathcal{C}_1 = \{c_1\}$ and $\Gamma^{\mathcal{C}_1} = \{0, 1\}$. By construction, we have $\sigma^{\mathcal{C}_1}(0|\hat{s}_i) = \sigma^{\mathcal{C}_T}(0|\hat{s}_i)$ and $\sigma^{\mathcal{C}_1}(1|\hat{s}_i) = \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \sigma^{\mathcal{C}_T}(t|\hat{s}_i)$. Similarly, for individual i, the probability of receiving test outcome 0 or 1 becomes $\hat{\lambda}_i^{\mathcal{C}_1}(0) = \hat{\lambda}_i^{\mathcal{C}_T}(0)$, $\hat{\lambda}_i^{\mathcal{C}_1}(1) = \sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)$. Let $\hat{f}_i(t)$ denote individual i's follow-up probability under $(1, \Gamma^{\mathcal{C}_1})$. Then, the expected probability of follow-up for individual i is

$$\begin{aligned}
\hat{\lambda}_i^{\mathcal{C}_1}(1)\hat{f}_i(1) &= \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) W(\hat{p}_i^{s1}(1), \hat{p}_i^{s2}(1)) \qquad \text{(C.27)} \\
&= \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) W \left(\frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) p_i^{s1}(t)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)}, \frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) p_i^{s2}(t)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)} \right).
\end{aligned}$$

Since $W(\cdot)$ is concave in $p_i^{s1}(t)$ and $p_i^{s2}(t)$, following by the concavity, we have

$$W \left(\frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) p_i^{s1}(t)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)}, \frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) p_i^{s2}(t)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)} \right) \geq \frac{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t) W\Big(p_i^{s1}(t), p_i^{s2}(t)\Big)}{\sum\limits_{t \in \Gamma^{\mathcal{C}_T} \backslash \{0\}} \hat{\lambda}_i^{\mathcal{C}_T}(t)}.$$

Therefore, the initial test design $(1, \mathcal{C}_1)$ indeed induces an equal or higher probability of following up than $(T, \mathcal{C}_T)$for any individual i $(i \in \{1, 2, ..., N\})$. A dichotomous test is optimal.

To prove (2), consider a dichotomous test with cut-off point $\underline{\zeta}$ and another with cut-off point $c^{'}$ where $\underline{\zeta} < c^{'} \leq \bar{\zeta}$.

Suppose $\underline{\zeta}$ induces two test outcomes denoted by $0_{\underline{\zeta}}$ and $1_{\underline{\zeta}}$, we have $\sigma(1_{\underline{\zeta}}|\hat{s}_i) = 1 - \hat{H}_{\hat{s}_i}(\underline{\zeta})$ and $\hat{\lambda}_i(1_{\underline{\zeta}}) = \sigma(1_{\underline{\zeta}}|0)(1 - p_i^1 - p_i^2) + \sigma(1_{\underline{\zeta}}|1)p_i^1 + \sigma(1_{\underline{\zeta}}|2)p_i^2$. Another dichotomous test with cut-off point $c^{'}$ have two test outcomes denoted by $0_{c^{'}}$ and $1_{c^{'}}$, and $\sigma(1_{c^{'}}|\hat{s}_i) = 1 - \hat{H}_{\hat{s}_i}(c^{'})$. For the sake of proof, we introduce an "invisible test outcome 0^1" and define $\sigma(0^1|\hat{s}_i) = \hat{H}_{\hat{s}_i}(c^{'}) - \hat{H}_{\hat{s}_i}(\underline{\zeta})$. So we can obtain the following relationships: $\sigma(1_{\underline{\zeta}}|\hat{s}_i) = \sigma(1_{c'}|\hat{s}_i) + \sigma(0^1|\hat{s}_i)$, $\hat{\lambda}_i(1_{\underline{\zeta}}) = \hat{\lambda}_i(1_{c'}) + \hat{\lambda}_i(0^1)$, $p_i^{s1}(1_{\underline{\zeta}}) = \frac{\sigma(1_{\underline{\zeta}}|1)p_i^1}{\hat{\lambda}_i(1_{\underline{\zeta}})} = \frac{\hat{\lambda}_i(1_{c^{'}})p_i^{s1}(1_{c^{'}}) + \hat{\lambda}_i(0^1)p_i^{s1}(0^1)}{\hat{\lambda}_i(1_{c^{'}}) + \hat{\lambda}_i(0^1)}$ and $p_i^{s2}(1_{\underline{\zeta}}) = \frac{\sigma(1_{\underline{\zeta}}|2)p_i^2}{\hat{\lambda}_i(1_{\underline{\zeta}})} = \frac{\hat{\lambda}_i(1_{c^{'}})p_i^{s2}(1_{c^{'}}) + \hat{\lambda}_i(0^1)p_i^{s2}(0^1)}{\hat{\lambda}_i(1_{c^{'}}) + \hat{\lambda}_i(0^1)}$.

The expected follow-up probability for individual i under $\underline{\zeta}$ or $c^{'}$ can be expressed by $\hat{\lambda}_i(1_{\underline{\zeta}})W\big[p_i^{s1}(1_{\underline{\zeta}}), p_i^{s2}(1_{\underline{\zeta}})\big]$ and $\hat{\lambda}_i(1_{c'})W\big[p_i^{s1}(1_{c'}), p_i^{s2}(1_{c'})\big]$. According to the concavity of $W(\cdot)$, we have

$$\begin{aligned}
W\big[p_i^{s1}(1_{\underline{\zeta}}), p_i^{s2}(1_{\underline{\zeta}})\big] &= W\Big[\frac{\hat{\lambda}_i(1_{c'})p_i^{s1}(1_{c'}) + \hat{\lambda}_i(0^1)p_i^{s1}(0^1)}{\hat{\lambda}_i(1_{c'}) + \hat{\lambda}_i(0^1)}, \frac{\hat{\lambda}_i(1_{c'})p_i^{s2}(1_{c'}) + \hat{\lambda}_i(0^1)p_i^{s2}(0^1)}{\hat{\lambda}_i(1_{c'}) + \hat{\lambda}_i(0^1)}\Big] \\
&\geq \frac{\hat{\lambda}_i(1_{c'})W\big[p_i^{s1}(1_{c'}), p_i^{s2}(1_{c'})\big]}{\hat{\lambda}_i(1_{c'}) + \hat{\lambda}_i(0^1)} + \frac{\hat{\lambda}_i(0^1)W\big[p_i^{s1}(0^1), p_i^{s2}(0^1)\big]}{\hat{\lambda}_i(1_{c'}) + \hat{\lambda}_i(0^1)} \\
&= \frac{\hat{\lambda}_i(1_{c'})W\big[p_i^{s1}(1_{c'}), p_i^{s2}(1_{c'})\big] + \hat{\lambda}_i(0^1)W\big[p_i^{s1}(0^1), p_i^{s2}(0^1)\big]}{\hat{\lambda}_i(1_{\underline{\zeta}})}.
\end{aligned}$$

Therefore, we can establish that the follow-up probability of any individual i under the test with cut-off point $\underline{\zeta}$ is higher than that under the test with cut-off point c' via the following.

$$\begin{aligned}
\sum_{i=1}^{N} \hat{\lambda}_i(1_{\underline{\zeta}})W\big[p_i^{s1}(1_{\underline{\zeta}}), p_i^{s2}(1_{\underline{\zeta}})\big] &\geq \sum_{i=1}^{N} \hat{\lambda}_i(1_{c'})W\big[p_i^{s1}(1_{c'}), p_i^{s2}(1_{c'})\big] + \sum_{i=1}^{N} \hat{\lambda}_i(0^1)W\big[p_i^{s1}(0^1), p_i^{s2}(0^1)\big] \\
&> \sum_{i=1}^{N} \hat{\lambda}_i(1_{c'})W\big[p_i^{s1}(1_{c'}), p_i^{s2}(1_{c'})\big].
\end{aligned}$$

We can conclude that the optimal cut-off point is $\underline{\zeta}$. □

Theorem C.3. *For the effectiveness maximization case, if $W\big(p_i^{s1}(t), p_i^{s2}(t)\big)$ is convex in $p_i^{s1}(t)$ and $p_i^{s2}(t)$ and nondecreasing in t, it's optimal to adopt the continuous initial test which directly report individuals' f-Hb values.*

Proof. Proof of Theorem C.3. We prove this theorem via two steps: (1) We first show that for any initial test with a finite number of cut-off points, the objective value under $\tau_1 \in (0,1)$ and $\tau_0 = 0$ will not decrease by arbitrarily adding one more cut-off point from $[\underline{\zeta}, \bar{\zeta}]$. (2) We prove that if we uniformly add infinitely many cut-off points, the objective value converges to that of the continuous test.

Given an initial test $(T, \mathcal{C}_T)$ with T cut-off points, where $\mathcal{C}_T = \{c_1, c_2, ..., c_T | c_t \in [\underline{\zeta}, \bar{\zeta}], c_1 < c_2 < ... < c_T\}$, the test outcome set is denoted as $\Gamma^{\mathcal{C}_T} = \{0, 1, ..., T\}$ and the likelihood of receiving test outcome t given state $\hat{s}_i$ is $\sigma(t|\hat{s}_i)$, $t \in \Gamma^{\mathcal{C}_T}$, $\hat{s}_i \in \hat{S}$. The probability of receiving test outcome t for an individual i is $\lambda_i(t)$, $t \in \Gamma^{\mathcal{C}_T}$.

The overall expected follow-up probability from individuals with CRC or polys equals

$$\sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T}} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1) p_i^1] f_i(t) = \sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T} \setminus 0} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1) p_i^1] W\left[p_i^{s1}(t), p_i^{s2}(t)\right]. \tag{C.28}$$

Suppose we arbitrarily add one cut-off point, $c' \in [\underline{\zeta}, \bar{\zeta}]$ that is not in $\mathcal{C}_T$. Adding one more cut-off point will split one test outcome, say k, to two, denoted by k_1 and k_2. For any individual i, We denote the likelihood of receiving test outcome k_1 and k_2 given state $\hat{s}_i$ as $\hat{\sigma}(j|\hat{s}_i), j = k_1, k_2$. By construction, $\sigma(k|\hat{s}_i) = \hat{\sigma}(k_1|\hat{s}_i) + \hat{\sigma}(k_2|\hat{s}_i)$. For the new initial test, the overall follow-up probability from patients with CRC or polyps is

$$\sum_{i=1}^{N} \sum_{t \in \cup \Gamma^{\mathcal{C}_T} \setminus \{k\}} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1) p_i^1] f_i(t) + \sum_{j \in \{k_1, k_2\}} [\hat{\sigma}(j|2)p_i^2 + \tau_1 \hat{\sigma}(j|1) p_i^1] f_i(j). \tag{C.29}$$

We next compare the values of (C.28) and (C.29) in the following two cases.

Case 1: If $k = 0$:

$$(C.29) = \sum_{i=1}^{N} \sum_{t \in \cup \Gamma^{\mathcal{C}_T} \setminus \{0\}} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1)p_i^1] W\left[p_i^{s1}(t), p_i^{s2}(t)\right] + \sum_{j \in \{k_1,k_2\}} [\hat{\sigma}(j|2)p_i^2 + \tau_1 \hat{\sigma}(j|1)p_i^1] f_i(j) \geq (C.2$$

Case 2: If $k \neq 0$, by construction, $k_1 > 0$ and $k_2 > 0$:

$$\begin{aligned}(C.29) &= \sum_{i=1}^{N} \sum_{t \in \cup \Gamma^{\mathcal{C}_T} \setminus \{0\}} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1)p_i^1] W\left[p_i^{s1}(t), p_i^{s2}(t)\right] + \sum_{j \in \{k_1,k_2\}} [\hat{\sigma}(j|2)p_i^2 + \tau_1 \hat{\sigma}(j|1)p_i^1] f_i(j) \\ &= \sum_{i=1}^{N} \sum_{t \in \cup \Gamma^{\mathcal{C}_T} \setminus \{0\}} [\sigma(t|2)p_i^2 + \tau_1 \sigma(t|1)p_i^1] W\left[p_i^{s1}(t), p_i^{s2}(t)\right] + \sum_{j \in \{k_1,k_2\}} [\hat{\sigma}(j|2)p_i^2 + \tau_1 \hat{\sigma}(j|1)p_i^1] W\left[\hat{p}_i^{s1}(j), \hat{p}\right.\end{aligned}$$

Where $\hat{p}_i^{s1}(j)$ and $\hat{p}_i^{s2}(j)$ are the posterior risk of having polyps and CRC given the test outcome $j (\in \{k_1, k_2\})$ under the new initial test. To show $(C.29) \geq (C.28)$, it is sufficient to prove

$$\sum_{j \in \{k_1,k_2\}} [\hat{\sigma}(j|2)p_i^2 + \tau_1 \hat{\sigma}(j|1)p_i^1] W\left[\hat{p}_i^{s1}(j), \hat{p}_i^{s2}(j)\right] \geq [\sigma(k|2)p_i^2 + \tau_1 \sigma(k|1)p_i^1] W\left[p_i^{s1}(k), p_i^{s2}(k)\right]. \tag{C.30}$$

Note that $\sigma(k|\hat{s}_i) = \hat{\sigma}(k_1|\hat{s}_i) + \hat{\sigma}(k_2|\hat{s}_i)$. Let the total probability of receiving test outcome j for an individual i under the new test be $\hat{\lambda}_i(j)$. Then, $\lambda_i(k) = \sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)$. Thus, $p_i^{s1}(k) = \frac{\sum_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^{s1}(j)}{\sum_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)}$ and $p_i^{s2}(k) = \frac{\sum_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^{s2}(j)}{\sum_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)}$. To prove Inequality (C.30), it is equivalent to show that

$$\frac{\sum\limits_{j \in \{k_1,k_2\}} [\tau_1 \hat{\sigma}(j|1)p_i^1 + \hat{\sigma}(j|2)p_i^2] W\left(\hat{p}_i^{s1}(j), \hat{p}_i^{s2}(j)\right)}{\sum\limits_{j \in \{k_1,k_2\}} \tau_1 \hat{\sigma}(j|1)p_i^1 + \hat{\sigma}(j|2)p_i^2} \geq W\left(\frac{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^{s1}(j)}{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)}, \frac{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)\hat{p}_i^{s2}(j)}{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)} \right). \tag{C.31}$$

Because $W(\cdot)$ is non-decreasing, and it is easy to check that $\frac{\tau_1\hat{\sigma}(j|1)p_i^{s1} + \hat{\sigma}(j|2)p_i^{s2}}{(1-\tau_1)\hat{\sigma}(j|1)p_i^{s1} + \hat{\sigma}(j|0)(1-p_i^{s1}-p_i^{s2})}$ is increasing in j. Then following from Corollary C.1, we have

$$\frac{\sum\limits_{j \in \{k_1,k_2\}} [\tau_1 \hat{\sigma}(j|1)p_i^1 + \hat{\sigma}(j|2)p_i^2] W\left(\hat{p}_i^{s1}(j), \hat{p}_i^{s2}(j)\right)}{\sum\limits_{j \in \{k_1,k_2\}} \tau_1 \hat{\sigma}(j|1)p_i^1 + \hat{\sigma}(j|2)p_i^2} \geq \frac{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j) W\left(\hat{p}_i^{s1}(j), \hat{p}_i^{s2}(j)\right)}{\sum\limits_{j \in \{k_1,k_2\}} \hat{\lambda}_i(j)}. \tag{C.32}$$

Following from the convexity of $W(\cdot)$, we have

$$\frac{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)W\big(\hat{p}_i^{s1}(j),\hat{p}_i^{s2}(j)\big)}{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)} \geq W\left(\frac{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)\hat{p}_i^{s1}(j)}{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)},\frac{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)\hat{p}_i^{s2}(j)}{\sum\limits_{j\in\{k_1,k_2\}}\hat{\lambda}_i(j)}\right). \tag{C.33}$$

Thus, according to above two inequalities, we conclude Inequality (C.31) holds. Thus, it's always optimal to add one more cut-off point to the original set of cut-off points.

Before proceeding to prove (2), we first introduce the Bayesian updating process and the property of the continuous test: under the continuous test, an individual receive a test outcome $\zeta \in [\underline{\zeta},\bar{\zeta}]$, which is the exact f-Hb concentration. For any individual i with test outcome ζ, the posterior risk of having polyps or CRC is denoted by $\check{p}_i^{s1}(\zeta)$ and $\check{p}_i^{s2}(\zeta)$, where:

$$\check{p}_i^{s1}(\zeta) = \frac{\hat{h}_1(\zeta)p_i^1}{\hat{h}_2(\zeta)p_i^2+\hat{h}_1(\zeta)p_i^1+\hat{h}_0(\zeta)(1-p_i^1-p_i^2)}, \quad i\in\{1,2,...N\},\ \zeta\in[\underline{\zeta},\bar{\zeta}];$$

$$\check{p}_i^{s2}(\zeta) = \frac{\hat{h}_2(\zeta)p_i^2}{\hat{h}_2(\zeta)p_i^2+\hat{h}_1(\zeta)p_i^1+\hat{h}_0(\zeta)(1-p_i^1-p_i^2)}, \quad i\in\{1,2,...N\},\ \zeta\in[\underline{\zeta},\bar{\zeta}].$$

Let $\check{f}_i(\zeta)$ denote the follow-up probability for individual i when adopting the continuous test, where $\check{f}_i(\zeta) = W(\check{p}_i^{s1}(\zeta),\check{p}_i^{s2}(\zeta))$. Given the range of possible fecal hemoglobin concentration $[\underline{\zeta},\bar{\zeta}]$, the overall follow-up probability for all individuals with CRC or polyps can be written as:

$$\sum_{i=1}^{N}\int_{\underline{\zeta}}^{\bar{\zeta}}\check{f}_i(\zeta)[\hat{h}_2(\zeta)p_i^2+\tau_1\hat{h}_1(\zeta)p_i^1]d\zeta \tag{C.34}$$

$$=\sum_{i=1}^{N}\lim_{J\to\infty}\sum_{j=0}^{J}\check{f}_i(\underline{\zeta}+\frac{(\bar{\zeta}-\underline{\zeta})j}{J})[\hat{h}_2(\underline{\zeta}+\frac{(\bar{\zeta}-\underline{\zeta})j}{J})p_i^2+\tau_1\hat{h}_1(\underline{\zeta}+\frac{(\bar{\zeta}-\underline{\zeta})j}{J})p_i^1]\cdot\frac{(\bar{\zeta}-\underline{\zeta})j}{J}$$

Next, we will show that in ordinal CRC tests, if policymakers uniformly choose infinitely many cut-off points from $[\underline{\zeta},\bar{\zeta}]$, the overall adjusted follow-up probability is the same as the value under continuous tests.

Define an ordinal test $(T,\mathcal{C}_T)$ with T cut-off points and test outcome set $\Gamma^{\mathcal{C}_T} =$

$\{0, 1, 2, ...T-1\}$, where

$$\mathcal{C}_T \equiv \{c_1, c_2, .., c_T | c_1 = \underline{\zeta}, c_{t+1} = c_t + \frac{\bar{\zeta} - \underline{\zeta}}{T-1}, \forall t = 1, ..., T-1\}.$$

Let $\Delta c(T) = \frac{\bar{\zeta}-\underline{\zeta}}{T-1}$. Then, $c_{t+1} = c_t + \Delta c(T) = \underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1}, \forall t = 1, ..., T-1$. The overall follow-up probability for individuals with CRC or polyps under this ordinal test is

$$\sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T}} f_i(t)[\sigma^{\mathcal{C}_T}(t|2)p_i^2 + \tau_1 \sigma^{\mathcal{C}_T}(t|1)p_i^1] = \sum_{i=1}^{N} \sum_{t \in \Gamma^{\mathcal{C}_T} \backslash 0} W\Big(p_i^{s1}(t), p_i^{s2}(t)\Big)[\sigma^{\mathcal{C}_T}(t|2)p_i^2 + \tau_1 \sigma^{\mathcal{C}_T}(t|1)p_i^1]. \tag{C.35}$$

Since for $t \in \Gamma^{\mathcal{C}_T} \backslash 0$ and $\hat{s} \in \{0, 1, 2\}$, $\sigma^{\mathcal{C}_T}(t|\hat{s}) = \hat{H}_{\hat{s}}(c_{t+1}) - \hat{H}_{\hat{s}}(c_t) = \hat{H}_{\hat{s}}(c_t + \Delta c(T)) - \hat{H}_{\hat{s}}(c_t)$. Based on the definition of $\Delta c(T)$, we have $\lim_{T\to\infty} \sigma^{\mathcal{C}_T}(t|\hat{s}) = \hat{h}_{\hat{s}}(c_t)\Delta c(T)$. Thus, when T goes to infinity, we have

$$\lim_{T\to\infty} p_i^{s1}(t) = \frac{\hat{h}_1(c_t)p_i^1}{\hat{h}_2(c_t)p_i^2 + \hat{h}_1(c_t)p_i^1 + \hat{h}_0(c_t)(1 - p_i^1 - p_i^2)} = \check{p}_i^{s1}(c_t),$$

$$\lim_{T\to\infty} p_i^{s2}(t) = \frac{\hat{h}_2(c_t)p_i^2}{\hat{h}_2(c_t)p_i^2 + \hat{h}_1(c_t)p_i^1 + \hat{h}_0(c_t)(1 - p_i^1 - p_i^2)} = \check{p}_i^{s2}(c_t).$$

Because $W(\cdot)$ is continuous on $\mathbb{R}^2$, Equation (C.35) can be rewritten as:

$$\begin{aligned}
(C.35) =& \sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=1}^{T-1} W\Big(\check{p}_i^{s1}(c_t), \check{p}_i^{s2}(c_t)\Big)[\hat{h}_2(c_t)p_i^2 + \tau_1 \hat{h}_1(c_t)p_i^1] \cdot \Delta c(T) \\
=& \sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=0}^{T-1} \check{f}_i(c_t)[\hat{h}_2(c_t)p_i^2 + \tau_1 \hat{h}_1(c_t)p_i^1] \cdot \Delta c(T) \\
=& \sum_{i=1}^{N} \lim_{T\to\infty} \sum_{t=0}^{T-1} \check{f}_i(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1})[\hat{h}_2(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1})p_i^2 + \tau_1 \hat{h}_1(\underline{\zeta} + \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1})p_i^1] \cdot \frac{(\bar{\zeta}-\underline{\zeta})t}{T-1} = (C.34).
\end{aligned}$$

Thus, if policymakers uniformly choose infinitely many cut-off points from $[\underline{\zeta}, \bar{\zeta}]$, the overall adjusted follow-up probability is the same under a continuous tests. □

C.7 Survey

Purpose of this survey: This survey is to understand your perception of colorectal cancer and its screening guidelines. The information you provide in this survey will be used to understand the current awareness level of colorectal cancer and further help design screening guidelines to improve public health.

Part 1: Background for Colorectal Cancer and Its Screening

1. Colorectal cancer, also known as bowel cancer, colon cancer, or rectal cancer, is any cancer that affects the colon and the rectum. Colorectal cancer can develop from a ”polyp”, a nonspecific term to describe a growth on the inner surface of the colon. This survey aims to understand your awareness and opinions towards colorectal cancer and its screening process.

Which of the following apply to you? Please tick all that apply.

a. I have a medical history of colitis (inflammation of the inner lining of the colon)

b. I have a medical history of polyps

c. I have a medical history of colorectal cancer

d. I have a close relative with colorectal cancer (parents, siblings, or children)

e. I have 2 or more close relatives with colorectal cancer (parents, siblings, or children)

f. I have a close relative with other cancers (parents, siblings, or children)

g. None of the above

2. A fecal occult blood test (FOBT) looks at a sample of your stool (feces) to check for hidden (occult) blood that you can’t see with the naked eye. The Faecal Immunochemical Test (FIT) is an advanced version of FOBT. These tests are preliminary tests that are used to estimate the chance of having polyps or bowel cancer by detecting occult blood in the stool sample. We are interested in your views on the accuracy of these tests.

If 100 people who have colorectal cancer take a FOBT/FIT test, how many do you think will incorrectly be classified as not having colorectal cancer?

3. Now consider the opposite scenario: If 100 people who do not have colorectal cancer take a FOBT/FIT test, how many do you think will incorrectly be classified

as having colorectal cancer?

4. Have you ever done a FOBT/FIT before? If yes, how many times?

   a. No, never

   b. Yes, once in the past

   c. Yes, more than once, but not regularly

   d. Yes, regularly, but less than once a year

   e. Yes, regularly, once a year

5. If Yes in Q4, have you received a positive result (indicating the presence of blood in your sample) from an FOBT/FIT test? If yes, how many times?

   a. No   b. Yes, once   c. Yes, more than once

6. If No in Q5, imagine you received a positive result from an FOBT/FIT test, and the doctor recommended that you have a colonoscopy. Which of the following factors would you consider when deciding whether to follow the doctor's recommendations? Please check all that apply

   a. My medical history

   b. My age

   c. How confident I am about my health condition

   d. How much I trust my doctor

   e. Accuracy of FOBT/FIT

   f. How much I want to know my actual health condition

   g. Embarrassment of doing a colonoscopy

   h. Comfort of a colonoscopy

   i. Price of a colonoscopy

   j. How supportive my family/friends were

k. None of the above

Please check questions Q9-Q16 based on your option in Q6.

7. If Yes in Q5, did you consult any doctors after receiving the positive result? Please check all that apply.

a. No, I did not consult any doctors

b. Yes, A doctor I consulted recommended waiting

c. Yes, A doctor I consulted recommended a follow-up colonoscopy

d. Yes, A doctor I consulted recommended another FOBT/FIT

e. Yes, A doctor I consulted recommended a different test other than FOBT/FIT or colonoscopy

8. If Yes in Q7, which of the following factors did you consider when deciding whether to follow the doctor's recommendations? Please check all that apply

a. My medical history

b. My age

c. How confident I am about my health condition

d. How much I trust my doctor

e. Accuracy of FOBT/FIT

f. How much I want to know my actual health condition

g. Embarrassment of doing a colonoscopy

h. Comfort of a colonoscopy

i. Price of a colonoscopy

j. How supportive my family/friends were

k. None of the above

Please check questions Q9-Q16 based on your option in Q8.

9. If you choose ”My age” in Q6/Q8, which of the following statements best describes how you felt about your age?

a. I thought I was too old for a colonoscopy b. I did not think I was too old for a colonoscopy

10. If you choose ”How confident I am about my health condition” in Q6/Q8, which of the following statements best describes how you felt about your health condition?

a. I was extremely confident that I was healthy

b. I was somewhat confident that I was healthy

c. I was a bit worried

d. I was very pessimistic

11. If you choose ”How much I trust my doctor” in Q6/Q8, which of the following statements best describes how you felt about your doctor?

a. I did not trust the doctor at all

b. I did not trust the doctor very much

c. I was neutral about the doctor

d. I trusted the doctor somewhat

e. I trusted the doctor completely

12. If you choose ”How much I wanted to know my actual health condition” in Q6/Q8, how much did you want to know your actual health condition?

a. I was very eager to know my actual health condition

b. I wanted to know my actual health condition

c. I was afraid to know my actual health condition

13. If you choose ”Embarrassment of doing a colonoscopy” in Q6/Q8, which of the fol-

lowing statements best describes how embarrassed you felt about doing a colonoscopy?

a. I felt very embarrassed of doing a colonoscopy

b. I felt somewhat embarrassed of doing a colonoscopy

c. I didn't feel embarrassed of doing a colonoscopy

14. If you choose "Comfort of a colonoscopy" in Q6/Q8, before the colonoscopy, how comfortable did you expect that the procedure would be?

a. I expected that the colonoscopy would be uncomfortable

b. I expected that the colonoscopy would be manageable

c. I expected that the colonoscopy would be enjoyable

15. If you choose "Price of a colonoscopy" in Q6/Q8, which of the following statements best describes how you felt about the price of the colonoscopy?

a. I thought colonoscopy was costly

b. I thought colonoscopy was affordable

c. I thought colonoscopy was cheap

16. If you choose "How supportive my family/friends were" in Q6/Q8, which of the following statements best describes the support of your family/friends?

a. My family/friends were very supportive of me doing a colonoscopy

b. My family/friends were somewhat supportive of me doing a colonoscopy

c. My family/friends were not supportive of me doing a colonoscopy

d. My family/friends were against me doing a colonoscopy

17. We are interested in your understanding of the facts relating to colorectal cancer. Please indicate whether you think that each of the following statements is True or False.

a. Colorectal cancer is the most diagnosed cancer in Singapore and the second most common cause of cancer-related deaths. Every day, around five Singaporeans are diagnosed with colorectal cancer and two die of it.

b. Even if colorectal cancer is detected at an early stage, it can only be controlled and is unlikely to be cured.

c. If colorectal cancer is detected at stage 1 the survival rates can be as high as 92%, compared to 11% for stage 4.

d. Few instances of colorectal cancer begin as polyps. So even if polyps are removed in a timely manner, patients can still develop colorectal cancer from other sources.

e. Early-stage polyps or colorectal cancer cannot be detected by regular screening.

f. If polyps are detected during a colonoscopy, the patient needs to make another appointment for a procedure to remove the polyps.

g. Colonoscopy is very accurate (almost 100%), but it is very costly and has potential serious side effects.

h. The nationwide Screen for Life (SFL) programme offers affordable and convenient screening for colorectal cancer.

i. Singaporeans and PRs aged 50 and above can obtain FOBT/FIT kits for free from the Singapore Cancer Society (SCS), once a year.

j. FOBT/FIT should be done once every year. If the result is positive, colonoscopy is strongly recommended.

<u>Part 2</u>: General Information

18. What is your age? __________

19. What is your current marital situation?

a. Married b. Single (never married) c. Separated d. Divorced e.

Widowed

20. Do you have any other health insurance plans such as insurance through employer, a business, or health insurance you buy for yourself?

a. Yes b. No

21. Do you own or partly own the house or apartment in which you live?

a. Yes b. No

22. If Yes in Q21, aside from the apartment in which you live, do you own (or co-own) any other real estate, such as residential properties, rental real estate, or land? Please do not include business property, that is any property that is used by a business that you might own.

a. Yes b. No

23. What is your monthly consumption of tobacco?