Optimal Data-driven Methods for Subject Classification in Public Health Screening

Seyedehsaloumeh Sadeghzadeh

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in the partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Industrial and Systems Engineering

Ebru K. Bish
Douglas R. Bish
Xi Chen
Scott J. Zimmerman

May 15, 2019
Blacksburg, Virginia

Keywords: Population Level Disease Screening, Risk-based Testing, Robust Optimization, Newborn Screening, Cystic Fibrosis

Optimal Data-driven Methods for Subject Classification in Public Health Screening

Seyedehsaloumeh Sadeghzadeh

(ACADEMIC ABSTRACT)

Biomarker testing, wherein the concentration of a biochemical marker is measured to predict the presence or absence of a certain binary characteristic (e.g., a disease) in a subject, is an essential component of public health screening. For many diseases, the concentration of disease-related biomarkers may exhibit a wide range, particularly among the disease positive subjects, in part due to variations caused by external and/or subject-specific factors. Further, a subject's actual biomarker concentration is not directly observable by the decision maker (e.g., the tester), who has access only to the test's measurement of the biomarker concentration, which can be noisy. In this setting, the decision maker needs to determine a classification scheme in order to classify each subject as test negative or test positive. However, the inherent variability in biomarker concentrations and the noisy test measurements can increase the likelihood of subject misclassification.

We develop an optimal data-driven framework, which integrates optimization and data analytics methodologies, for subject classification in disease screening, with the aim of minimizing classification errors. In particular, our framework utilizes data analytics methodologies to estimate the posterior disease risk of each subject, based on both subject-specific and external factors, coupled with robust optimization methodologies to derive an optimal robust subject classification scheme, under uncertainty on actual biomarker concentrations. We establish various key structural properties of optimal classification schemes, show that they are easily implementable, and develop key insights and principles for classification schemes in disease screening.

As one application of our framework, we study newborn screening for cystic fibrosis in the United States. Cystic fibrosis is one of the most common genetic diseases in the United States. Early diagnosis of cystic fibrosis can substantially improve health outcomes, while a delayed diagnosis can result in severe symptoms of the disease, including fatality. We demonstrate our framework on a five-year newborn screening data set from the North Carolina State Laboratory of Public Health. Our study underscores the value of optimization-based approaches to subject classification, and show that substantial reductions in classification error can be achieved through the use of the proposed framework over current practices.

Optimal Data-driven Methods for Subject Classification in Public Health Screening

Seyedehsaloumeh Sadeghzadeh

(GENERAL AUDIENCE ABSTRACT)

A biomarker is a measurable characteristic that is used as an indicator of a biological state or condition, such as a disease or disorder. Biomarker testing, where a biochemical marker is used to predict the presence or absence of a disease in a subject, is an essential tool in public health screening. For many diseases, related biomarkers may have a wide range of concentration among subjects, particularly among the disease positive subjects. Furthermore, biomarker levels may fluctuate based on external factors (e.g., temperature, humidity) or subject-specific characteristics (e.g., weight, race, gender). These sources of variability can increase the likelihood of subject misclassification based on a biomarker test.

We develop an optimal data-driven framework, which integrates optimization and data analytics methodologies, for subject classification in disease screening, with the aim of minimizing classification errors. We establish various key structural properties of optimal classification schemes, show that they are easily implementable, and develop key insights and principles for classification schemes in disease screening.

As one application of our framework, we study newborn screening for cystic fibrosis in the United States. Cystic fibrosis is one of the most common genetic diseases in the United States. Early diagnosis of cystic fibrosis can substantially improve health outcomes, while a delayed diagnosis can result in severe symptoms of the disease, including fatality. As a result, newborn screening for cystic fibrosis is conducted throughout the United States. We demonstrate our framework on a five-year newborn screening data set from the North Carolina State Laboratory of Public Health. Our study underscores the value of optimization-based approaches to subject classification, and show that substantial reductions in classification error can be achieved through the use of the proposed framework over current practices.

To Mom and Dad,
who never stopped believing in me.

Acknowledgments

I would like to express my deepest appreciation towards my advisors, Dr. Ebru Bish, and Dr. Douglas Bish, for their continuous support and guidance throughout my Ph.D. studies. I thank Ebru for her encouragement, trust, enthusiasm, extensive knowledge, kindness, and patience. I am forever indebted to her for everything she has taught me. She has been a true mentor for me, on both professional and personal levels, and I will be always inspired by her kindness, hard work, and passionate attitude. I thank Doug for his novel ideas, and for always challenging me to think out of the box. I am grateful for the friendly and intellectually stimulating environment Ebru and Doug create for all their students, where we could do research and challenge ourselves, while enjoying the learning process.

I would also like to extend my gratitude to my committee members, Dr. Xi Chen and Dr. Scott Zimmerman, for their insightful comments and encouragements. I appreciate the time, insight, and support that they gave me throughout this research. Special thanks to Dr. Xi Chen for always being there for me whenever I needed her guidance. I am also grateful to our industrial collaborator, the North Carolina State Laboratory of Public Health, for offering us valuable insights into public health screening practices for cystic fibrosis.

I am thankful to all my friends, in Blacksburg and abroad, who have always supported me, uplifted me, comforted me, and shared unforgettable moments with me. I would like to extend my gratitude to my officemates, Hussein and Ngoc, for their friendship and support, and for all the fun we have had during these years.

Finally, I am forever indebted to my family for their endless support, unconditional love, and sacrifices. Many thanks to my sister, Solaleh, and my brother in law, Sharif, for always being there for me, and for being my source of confidence. At the end, I am thankful to my parents for teaching me the values and strength of character that have made me who I am. To my parents, Saleh and Maryam, I dedicate this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction, Motivation, and Research Overview

This chapter is organized as follows. Section 1.1 provides the motivation and objectives of this study, while Section 1.2 provides an overview of the research.

1.1 Motivation

Biochemical marker (*biomarker*) is a measurable characteristic that is used as an indicator of some biological state or condition, such as a disease or an infection. Examples of biomarkers include immunoreactive trypsinogen (IRT) for detection of cystic fibrosis [43, 72], and natriuretic peptides, particularly BNP and its amino-terminal co-metabolite, NT-proBNP, for detection, diagnosis, and evaluation of the severity of heart failure [71]. Biomarker testing is utilized for many purposes, including the diagnosis, monitoring, and management of many diseases, such as genetic disorders (e.g., cystic fibrosis) [43, 72], cardiovascular diseases [71], Alzheimer's disease [26], asthma [70], neurological diseases [49], and different types of cancer [11, 13, 59]. In this study, we focus on using biomarkers in population level disease screening (e.g., newborn screening for genetic diseases), where a *threshold* on the biomarker measurement needs to be set to classify subjects into groups of *test positives* and *test negatives*.

Although biomarker testing greatly facilitates the diagnosis process of many diseases, designing an effective biomarker testing and subject classification policy remains a challenging decision. This is because, in many cases, there are some factors (besides the existence or absence of a disease) that can affect the related biomarker's concentration level or measurement (i.e., the test reading). These factors include external factors, such as seasonality (i.e., temperature and humidity), or testing kit calibration, and/or subject-specific factors, such as the subject's weight, gender, or race [18, 43, 49, 55, 59, 72]. As a result, the biomarker concentration level of disease positive and disease negative populations may overlap. This increases the likelihood of subject misclassification, i.e., classifying a disease negative subject as test positive (a *false positive* classification), or classifying a disease positive subject as test negative (a *false negative* classification).

Both false negative and false positive cases lead to negative consequences. False negative cases result in delayed diagnoses, and may lead to severe symptoms of the disease, leading to treatment complications, including fatality, and to higher healthcare expenditures. False positive cases, on the other hand, may be referred for unnecessary, and typically expensive, further testing, and may lead to stress and anxiety in the subjects. Thus, for many diseases, the consequences of a false negative classification are more severe, and the primary goal is to identify the disease positive subjects as accurately as possible. Our goal is to devise optimal biomarker testing and classification policies so as to minimize the consequences of subject misclassification, which are represented in terms of false negative and false positive classification costs.

In particular, we determine an optimal biomarker testing and classification policy, informed by potentially noisy biomarker measurements, when the actual biomarker concentration levels are unobservable. Because the decision maker does not have perfect information about the actual biomarker levels of subjects, we formulate a novel robust optimization model that requires only an uncertainty set around the biomarker level of each subject. We also study an expectation-based optimization model that uses an expected value of each

subject's biomarker level. We characterize various structural properties of optimal subject classification policies for both the robust and the expectation-based models. We also demonstrate the effectiveness of the proposed policies through a realistic case study on newborn screening for cystic fibrosis (CF) in the United States. CF is one of the most prevalent and life-threatening genetic diseases in the United States, with an approximate prevalence rate of 1 in 3,700 newborns [12, 58]. Early diagnosis of CF, through newborn screening, substantially improves long-term growth and prevents severe symptoms of the disease [19, 24, 43], while a delayed diagnosis may lead to complications or even fatality [30, 46, 63]. Therefore, newborn screening for CF is conducted throughout the United States. Although the IRT (the biomarker used in CF newborn screening) classification policy, used to classify newborns as test positive or test negative, has a high impact on the sensitivity and specificity of the overall CF screening process, there are no guidelines on how to set the IRT threshold value to minimize classification errors. We identify various external and newborn-specific factors that substantially affect the IRT level of newborns, based on a five-year data set from the North Carolina State Laboratory of Public Health (NCSLPH). Then, we use optimal data-driven approaches, which integrate stochastic and robust optimization with data analytics methodologies, and show that our optimization-based classification policies, which are easily implementable, substantially improve the classification accuracy for the IRT test, thus reduce the expected cost of misclassification over current practices.

1.2 Research Overview

We first study the problem of determining an optimal classification policy for biomarkers used in population level disease screening under noisy biomarker measurement. This research is detailed in Chapter 2 of this dissertation. We formulate this decision problem considering robust and expectation-based models, and obtain key structural properties of the optimal solutions. Our proposed models take into account biomarker variations due to both external

and subject-specific factors. We show that optimal classification policies are risk-based threshold policies, and we integrate the optimal classification with data-driven approaches to estimate the disease risk for each subject. Then, we use our models in a case study, and generate optimal state-level biomarker classification policies for the IRT test, which is a biomarker test that measures the concentration of IRT in the blood, and is commonly used for CF newborn screening, and show that our optimization-based policies outperform current IRT classification policies.

Then, in Chapters 3 and 4, we exclusively focus on newborn screening for CF. This study is in collaboration with the NCSLPH. All states start the CF screening process with the IRT test [35]. On average, CF positive newborns have elevated IRT levels [42, 43, 72]. We show that current IRT classification policies, i.e., proportion-based policies, which classify a daily fixed percentage of newborns tested each day as test positive, and concentration-based policies, which classify newborns with IRT levels greater than or equal to a fixed IRT threshold as test positive, are sub-optimal, mainly because they fail to account for variations in IRT concentration levels due to external factors, such as testing kit calibration, and seasonality, as well as newborn-specific factors, such as birth weight, race, and gender [18, 43, 55, 72], in a rigorous manner. Moreover, current IRT classification policies are not designed to take equity issues into consideration, i.e., they result in significantly different probability of false positive and false negative for different races.

In Chapter 3, we study the effect of one of the main external factors, namely seasonality, on IRT concentration levels, and show that on average, cold weather substantially increases the IRT level. Proportion-based policies heuristically take into account the effect of external factors, such as seasonality (while concentration-based policies do not). They do so by considering the testing population on each day, i.e., newborns who are screened on each day, and classify a specific percentage of newborns with elevated IRT measurements as test positive. Consequently, the performance of proportion-based policies depends highly on the size of the testing population each day. As a result, we show that proportion-based policies

may lead to a high number of misclassifications on days when the testing population is very small. We propose an adjusted policy that takes into account external factors, while reducing the sample size error.

Then, in Chapter 4, we extend our analysis to identify all external and newborn-specific factors that have a major effect on the IRT level, and show that the current IRT classification policies can be improved, both in terms of accuracy and equity, by incorporating these key factors into the design of an IRT classification policy. Specifically, we develop a data-driven adjusted IRT classification policy, and show that it can substantially improve CF newborn screening and classification outcomes.

Finally, Chapter 5 presents a summary of this research. To facilitate the presentation, all proofs, along with some tables and derivations, are relegated to the Appendix.

Chapter 2

Optimal Data-driven Policies for Disease Screening under Noisy Biomarker Measurement

2.1 Introduction and Motivation

A biomarker is a measurable characteristic that is used as an indicator of some biological state or condition, such as a disease or disorder (we use the term "disease" to refer to all such conditions). Biomarker testing plays an integral role in screening, diagnosis, monitoring, and management of many diseases, including genetic diseases such as cystic fibrosis [43, 72], cardiovascular diseases [71], Alzheimer's disease [26], asthma [70], neurological diseases [49], and various types of cancer [11, 13]. As an example of a biomarker, consider cystic fibrosis, which often leads to elevated immunoreactive trypsinogen (IRT) levels; therefore, cystic fibrosis screening in the United States (US) typically includes a biomarker test that measures the IRT concentration (the *IRT test*) [43, 72]. In this paper, our focus is on using biomarkers in population level screening of non-infectious diseases. (Our models apply to non-infectious diseases, because we do not model disease transmission among subjects, which

is an important source of transmission for infectious diseases.)

Biomarker testing offers a low cost and a convenient option for screening large populations, and hence, is commonly used for screening purposes. However, biomarker tests may not be perfectly reliable, and as a result, designing an effective biomarker screening policy becomes challenging. What complicates policy design is that for many diseases, the concentration level of the related biomarkers may have a wide range, particularly among the disease positive subjects. This may occur due to subject-specific characteristics, such as weight, race, gender [18, 43, 49, 55, 72], level of disease progression, or other medical conditions of the subject [18, 25, 44]. As a result, the range of biomarker concentrations in disease positive and disease negative populations may overlap. Further, the test's biomarker reading (measurement) may differ from the subject's true biomarker level, which is often not directly observable [78, 84], due to perturbations caused by external factors. As an example, IRT readings in both cystic fibrosis positive and cystic fibrosis negative subjects can be altered by outside temperature and humidity, and calibration of the testing measurement kit [18, 43, 55, 72]. These natural variations and perturbations in biomarker concentrations increase the likelihood of subject *misclassification*, i.e., a disease negative subject classified as test positive (a *false positive classification*), or a disease positive subject classified as test negative (a *false negative classification*). False negative cases experience delayed diagnoses, which may lead to poor health outcomes, and/or an increase in healthcare expenditures. False positive cases, on the other hand, may be sent for further testing that is unnecessary. In particular, to improve the accuracy of screening, subjects testing positive in a biomarker test may undergo further, and typically more expensive, tests that have higher sensitivity and specificity. For instance, in newborn screening for cystic fibrosis, subjects testing positive in the initial biomarker (IRT) test are sent for further testing, including a genetic test and the diagnostic sweat chloride test, depending on the state's policy [2]. For many diseases, the consequence of a false negative classification is more severe than that of a false positive

classification. Our goal is to devise an optimal biomarker screening policy so as to minimize the consequences of subject misclassification, which are represented in terms of false negative and false positive classification costs.

A number of papers develop optimal policies for different types of cancer screening, including breast cancer screening (e.g. [7, 8, 9, 10, 64]) and prostate cancer screening (e.g., [13, 16, 56, 74, 85]), as well as for other screening purposes, such as childhood obesity [80]. The objective of these papers is to maximize a utility function, and the focus is on sequential screening policies, while we consider a one-time screening policy. Other studies investigate how to set biomarker thresholds in pooled testing, which involves combining specimens (e.g., blood samples) from multiple subjects into a pool and testing the entire pool with one test (e.g., [50, 62, 77, 78]). We focus on settings where testing is performed on an individual basis, e.g., the biomarker level of every subject is measured.

A stream of research investigates how to set a single decision threshold for screening or diagnostic purposes, based on the receiver operation characteristic (ROC) curve [53]. The threshold can correspond to a risk threshold [14, 31, 75], or a biomarker threshold [32, 34, 65, 68, 69, 76], such that all subjects having a disease risk or biomarker concentration above a certain threshold are classified as test positive, and all others are classified as test negative. A number of papers study methods to estimate the sensitivity and specificity of a test at a number of biomarker thresholds. These works then determine the "best" biomarker threshold, among the set of thresholds considered, that yields the highest weighted sum of the sensitivity and specificity of the test (e.g., [68, 69]). In particular, the aforementioned papers assume that biomarker distributions for disease positive and disease negative populations are known, but their parameters are uncertain, and utilize a Bayesian framework to update the distributions of those parameters [68, 69]. As opposed to this, we study distribution-free approaches through robust optimization models. Further, we propose an *optimization* framework for biomarker threshold selection, and consider the potential perturbations in

biomarker concentrations or readings due to external or subject-specific factors.

There are a number of studies that determine optimal thresholds that maximize a utility function, which assigns a utility to all possible outcomes (e.g., [23, 40, 52]). For example, Deneef *et al.* [23] assess the tester's utility by considering the trade-off between the number of false positives and number of true positives, and characterize diagnostic threshold policies within an expected utility framework. Pauker *et al.* [52] consider the options of administering treatment, ordering a diagnostic test, and withholding treatment, and determine optimal thresholds on the subject's estimated disease risk, considering a two-threshold policy, so as to maximize the expected utility.

This paper's contribution is to determine an optimal data-driven screening policy for non-infectious diseases that is informed by noisy, and possibly correlated, biomarker readings, and other subject-specific attributes (e.g., weight, race, gender), when the true biomarker concentrations are unobservable. In particular, we explore expectation-based and robust formulations of this decision problem, and characterize various structural properties of optimal screening policies. Our models are generic, and apply also in settings where the distributions of biomarker concentrations in disease positive and disease negative populations are unknown. We demonstrate the effectiveness of the proposed data-driven policies through a case study on newborn screening for cystic fibrosis in North Carolina, using a five-year data set from the North Carolina State Laboratory of Public Health. Cystic fibrosis is one of the most prevalent genetic diseases, and newborn screening for cystic fibrosis is performed throughout the US. While the IRT biomarker test is used in the newborn cystic fibrosis screening process of all fifty states, each state determines its own screening policy, and there are no guidelines on how a state should customize its biomarker screening policy, considering unique state-level inputs. The proposed mathematical models, complemented by regression analyses on the North Carolina data set, produce state-wide optimal policies that consider the demographics and climate of the state (important inputs for cystic fibrosis screening),

and that are easily implementable. Our case study indicates that these optimal policies can substantially increase the classification accuracy for cystic fibrosis screening over current practices.

The remainder of this paper is organized as follows. Section 2.2 presents the notation and the decision problem. Then, Section 2.3 provides the expectation-based and robust formulations, and Section 2.4 derives important structural properties of optimal policies. Section 2.5 studies the *price of robustness* and the *price of expectation-based optimization*, which correspond to the respective deviation of the expected misclassification cost produced by each model from the minimum possible expected misclassification cost, i.e., when the true biomarker concentrations are perfectly observable. Section 2.6 discusses our case study of cystic fibrosis newborn screening program in North Carolina. Finally, Section 2.7 summarizes our findings and provides directions for future research. To facilitate the presentation, all proofs, and some tables and derivations are relegated to the Appendix.

2.2 The Notation and the Decision Problem

In this section, we present the notation and the decision problem. Throughout, we denote vectors by an arrow; and random variables and their realization in upper-case and lower-case letters, respectively. We use the notation that $(X)^+ = max\{X, 0\}$. The terms *positive* and *negative* refer to both a subject's true disease status (true positive or true negative), and classification outcome (test positive or test negative).

In each period, the lab receives a set, Ω, of subjects to be screened for and classified as positive or negative for a certain non-infectious disease. Subjects testing positive in the biomarker test can be sent for further testing, depending on the setting. Screening involves a test that measures the concentration of a disease-related biomarker. While subjects with the disease tend to have elevated biomarker levels, disease negative subjects may also have elevated biomarker levels (or test readings above normal levels) due to subject-specific

attributes (e.g., weight, race, gender), external factors (e.g., temperature, humidity), or testing error. Hence, subject j, with a true biomarker level, Y_j, is a true positive for the disease with some probability (*risk*) $P_j(Y_j)$, which is non-decreasing in Y_j, i.e., the higher the biomarker level, the higher the probability that the subject has the disease. Then, given a biomarker level, Y_j, the true positivity status of subject j for the disease follows a Bernoulli distribution with a probability of $P_j(Y_j)$, i.e., $D_j(Y_j) \sim Bernoulli(P_j(Y_j))$, with the random variable D_j attaining a value of 1 if the subject is true positive, and a value of 0, otherwise. To simplify the subsequent notation, we denote the true risk of subject j by P_j.

On the testing side, the true biomarker level, Y_j, $j \in \Omega$ (hence the true risk, P_j), is not observable; and the biomarker test may provide a noisy reading, which we denote by $\widetilde{Y}_j$. Further, as discussed above, the biomarker reading vector in each testing period, denoted by $\overrightarrow{\widetilde{Y}}$, may be correlated because of the possibility of *common* perturbations in biomarker measurements in each period, due to external factors that may affect the reading for each subject in a similar way (see the discussion and case study in Section 2.6). Thus, our model can take into account both a common perturbation in biomarker measurements for all subjects tested within the same period, as well as independent perturbations due to subject-specific characteristics or independent measurement errors.

Let $\overrightarrow{\Theta}_j$ denote the values of a set of attributes for subject j that are known to influence biomarker levels, e.g., weight, race, and gender. After observing the biomarker reading vector $\overrightarrow{\widetilde{y}}$ and subject attribute vector $\overrightarrow{\theta}$ in each period, the tester: (1) derives point estimates for the true biomarker level, i.e., $\widehat{y}_j = h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}})$; and disease risk, i.e., $\widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j)$, for each subject $j \in \Omega$, via some estimation functions $h(.)$ and $g(.)$; (2) constructs an uncertainty set around the true risk vector, $\overrightarrow{P}$, given by $S(\overrightarrow{P}) = \left(\left[\underline{p}_j, \overline{p}_j \right] \right)_{j \in \Omega}$; and (3) classifies each subject as test positive or test negative (see Table 2.1 for the notation). Note that the width of the uncertainty set on $\overrightarrow{P}$, i.e., the "budget of uncertainty" [15], can be adjusted to reflect varying levels of confidence around the random variables, as discussed subsequently.

We make no assumptions on the functional forms of $h(\overrightarrow{\theta}, \overrightarrow{\widetilde{y}})$ and $g(\widehat{y}, \widetilde{y})$, and our approach is distribution-free, that is, our models do not require the distributions of biomarker levels in disease positive and disease negative populations. Function $h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}})$, which is used to estimate a subject's true biomarker level, by removing the common perturbation and subject-specific perturbations (due to the subject's specific attributes) from the subject's biomarker reading, depends on the testing period's biomarker reading vector, $\overrightarrow{\widetilde{y}}$, and subject-specific attributes, $\overrightarrow{\theta_j}$. Therefore, we refer to $\widehat{y}_j = h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}})$ as the "processed" biomarker level for each subject $j \in \Omega$, i.e., with perturbations removed. Then, function $g(\widehat{y}, \widetilde{y})$, which is used to estimate a subject's disease risk, depends only on the subject's processed and raw reading levels, i.e., $\widehat{y}$ and $\widetilde{y}$, respectively.

Table 2.1: Random variables, point estimates, and uncertainty sets

Random variables (unobservable)	Measurements	Point estimates	Uncertainty sets
$\overrightarrow{Y} = (Y_j)_{j\in\Omega}$ (true biomarker level vector) $\overrightarrow{P} = (P_j(Y_j))_{j\in\Omega}$ (true risk vector)	$\overrightarrow{\widetilde{y}} = (\widetilde{y}_j)_{j\in\Omega}$ (biomarker reading vector)	$\overrightarrow{\widehat{y}} = (\widehat{y}_j)_{j\in\Omega} = (h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}}))_{j\in\Omega}$ $\overrightarrow{\widehat{p}} = (\widehat{p}_j)_{j\in\Omega} = (g(\widehat{y}_j, \widetilde{y}_j))_{j\in\Omega}$	$S(\overrightarrow{P}) = \left([\underline{p}_j, \overline{p}_j]\right)_{j\in\Omega}$

Remark 1. *Various methods can be employed to derive the support of the risk vector,* $S(\overrightarrow{P}) = \left([\underline{p}_j, \overline{p}_j]\right)_{j\in\Omega}$*; for instance, by constructing an uncertainty set around* $\overrightarrow{\widetilde{Y}}$*, given by* $S(\overrightarrow{\widetilde{Y}}) = \left([\underline{y}_j, \overline{y}_j]\right)_{j\in\Omega}$*, which translates into* $S(\overrightarrow{P}) = \left([g(\underline{y}_j, \widetilde{y}_j), g(\overline{y}_j, \widetilde{y}_j)]\right)_{j\in\Omega}$*; or by letting* $\underline{p}_j = inf_{g(.)\in G(.)}\{g(\widehat{y}_j, \widetilde{y}_j)\}$ *and* $\overline{p}_j = sup_{g(.)\in G(.)}\{g(\widehat{y}_j, \widetilde{y}_j)\}$*, where* $G(.)$ *is the set of all possible risk estimation functions,* $g(.)$*.*

In this setting, subject misclassification is possible, because the true biomarker level, Y_j (hence, the true risk, P_j), is not observable, and moreover, even if Y_j were observed, the true disease status would still not be observable (i.e., D_j is a random variable). Consequently, a true positive subject can be falsely classified as negative (i.e., a false negative classification), or a true negative subject can be falsely classified as positive (i.e., a false positive classification).

Then the tester's decision problem is how to classify each subject in set Ω as test positive

versus test negative for the disease, based on the biomarker reading vector, $\overrightarrow{\widetilde{y}}$, and the subject-specific attribute vector, $\overrightarrow{\theta_j}$, $j \in \Omega$, which provide an estimated risk vector, $\overrightarrow{\widehat{p}} = (\widehat{p}_j)_{j\in\Omega}$, so as to minimize a function of the misclassification cost in each period. Thus, the decision variable set is a binary vector, $\overrightarrow{x} = (x_j)_{j\in\Omega}$, where x_j attains a value of 1 if subject j is classified as test positive, and a value of 0, otherwise. Then, subject j, $\forall j$, will be a false positive if $\{x_j = 1, D_j = 0\} \Leftrightarrow \{x_j(1 - D_j) = 1\}$, and a false negative if $\{x_j = 0, D_j = 1\} \Leftrightarrow \{(1 - x_j)D_j = 1\}$. Letting c_{FN} and c_{FP} respectively denote the per subject cost of a false negative classification and a false positive classification, the total misclassification cost, for a given classification vector $\overrightarrow{x}$, can be expressed as:

$$C(\overrightarrow{x}) = \sum_{j\in\Omega} \Big[c_{FN}(1 - x_j)D_j + c_{FP}x_j(1 - D_j)\Big].$$

To simplify the notation, we omit the arguments in parentheses when clear from the context.

2.3 Mathematical Formulations of the Decision Problem

In this section, we provide two formulations of the decision problem under uncertainty on the true subject risk, $P_j(Y_j)$, $j \in \Omega$: (i) an expectation-based optimization model (**EM**), and (ii) a robust optimization model (**RM**).

In the expectation-based optimization model, the tester classifies each subject in set Ω as test positive or test negative based on an estimated disease risk vector $\overrightarrow{\widehat{p}}$, so as to minimize the *perceived* expected misclassification cost in each period. In doing so, the tester *assumes* that $E[D_j|\,\overrightarrow{\widehat{p}}] = \widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j)$, which is not necessarily the case (see Section 2.5 for discussion of the price of expectation-based optimization, i.e., the deviation of the **EM** optimal cost from the minimum possible expected misclassification cost corresponding to the

true risk vector, $\overrightarrow{p}$).

Problem EM:

$$\begin{aligned} minimize_{\overrightarrow{x}=(x_j)_{j\in\Omega}} E\left[C(\overrightarrow{x})|\,\overrightarrow{\widehat{p}}\right] &= E_{\overrightarrow{D}}\left[\left(\sum_{j\in\Omega}\left[c_{FN}(1-x_j)D_j + c_{FP}\,x_j(1-D_j)\right]\right)|\,\overrightarrow{\widehat{p}}\right] \\ &= \sum_{j\in\Omega}\left[c_{FN}(1-x_j)E\left[D_j|\,\overrightarrow{\widehat{p}}\right] + c_{FP}\,x_j E\left[(1-D_j)|\,\overrightarrow{\widehat{p}}\right]\right] \\ &= \sum_{j\in\Omega}\left[c_{FN}(1-x_j)\widehat{p}_j + c_{FP}\,x_j(1-\widehat{p}_j)\right] \end{aligned}$$

$$subject\,to \qquad x_j\ binary, \forall j \in \Omega.$$

Observe that the **EM** objective function is additively separable in each x_j, $j \in \Omega$, given $\overrightarrow{\widehat{p}}$. This is because each $\widehat{p}_j$, $j \in \Omega$, is a function only of the subject's measured biomarker level, $\widetilde{y}_j$, and estimated (processed) biomarker level, $\widehat{y}_j$, which is derived by removing the common perturbation term and subject-specific perturbations from the biomarker readings in each period, via the $h(.)$ function (see Table 2.1).

Since perfect information on subject disease risk is not available to the tester, the optimal value of the expectation-based model may deviate from the minimum possible expected misclassification cost. Hence, in the following, we also provide a distribution-free approach, via a robust optimization model, that requires only an uncertainty set around the disease risk. In the robust optimization model, the tester classifies each subject in set Ω as test positive or test negative based on the uncertainty set around $\overrightarrow{P}$, i.e., $S(\overrightarrow{P}) \equiv \left([\underline{p}_j, \overline{p}_j]\right)_{j\in\Omega}$ (see Table 2.1). The objective is to minimize the maximum *Regret*, where *Regret* represents the cost of not acting optimally due to the unobservability of the true risk vector, $\overrightarrow{P}$, that

is, for any classification, $\overrightarrow{x}$, and any possible risk vector realization, $\overrightarrow{p} \in S(\overrightarrow{P})$, we have:

$$\begin{aligned} Regret(\overrightarrow{x}, \overrightarrow{p}) &\equiv E\Big[C(\overrightarrow{x})|\overrightarrow{p}\Big] - E\Big[C(\overrightarrow{x}^*(\overrightarrow{p}))|\overrightarrow{p}\Big] \\ &= \sum_{j\in\Omega}\Big[c_{FN}(1-x_j)p_j + c_{FP}\,x_j(1-p_j)\Big] - \sum_{j\in\Omega}\Big[c_{FN}(1-x_j^*(\overrightarrow{p}))p_j + c_{FP}\,x_j^*(\overrightarrow{p})(1-p_j)\Big] \\ &= \sum_{j\in\Omega} Regret(x_j, p_j), \end{aligned} \tag{2.1}$$

where $\overrightarrow{x}^*(\overrightarrow{p})$ is the optimal solution to the deterministic problem in which $\overrightarrow{p}$ is known, i.e., the solution to **EM** when $\overrightarrow{p}$ is replaced by $\overrightarrow{p}$.

In other words, *Regret* is the "additional" misclassification cost that is incurred due to imperfect information; in our context, imperfect information on the disease risk of each subject. Mini-max *Regret* type objectives are used for various decision problems under uncertainty (e.g., [3, 6, 28, 48, 54, 61, 83]), mainly because the mini-max *Regret* objective is less conservative than traditional objective functions of robust formulations, such as the mini-max objective that minimizes the cost of the worst-case scenario [54]. The robust formulation of finding a classification, $\overrightarrow{x}$, that minimizes the maximum *Regret* over all possible realizations of the random vector, $\overrightarrow{P}$, then follows:

Problem RM:

$$\begin{aligned} &minimize_{\overrightarrow{x}=(x_j)_{j\in\Omega}}\Big\{max_{\overrightarrow{p}\in S(\overrightarrow{P})}\{Regret(\overrightarrow{x}, \overrightarrow{p})\}\Big\} \\ &subject\ to \qquad x_j\ binary, \forall j \in \Omega. \end{aligned} \tag{2.2}$$

The maximum *Regret* value for each $\overrightarrow{x}$ needs to be determined over the sample space of $\overrightarrow{P}$, $S(\overrightarrow{P})$, which is uncountable. In what follows, we study structural properties of the *Regret* function to develop an effective algorithm for **RM**.

We use the superscripts E and R to denote the expressions that respectively correspond to **EM** and **RM**, and use the superscript $*$ to denote an optimal solution to each problem, e.g., $\overrightarrow{x}^{*E}$ and $\overrightarrow{x}^{*R}$, respectively.

2.4 Structural Properties of EM and RM Optimal Solutions

We develop key structural properties of optimal **EM** and **RM** solutions in Section 2.4.1, and discuss the link between the objectives of minimizing the misclassification cost and maximizing the test efficacy (Appendix A.2), so as to relate our optimization models to those studied in the literature. In order to facilitate the presentation, all proofs are relegated to the Appendix.

2.4.1 Structural Properties of EM and RM Optimal Solutions

We first characterize the structural properties of an optimal **EM** solution.

Theorem 1. *Given a risk estimate vector, $\overrightarrow{\widehat{p}}$, an optimal **EM** solution follows a risk-based threshold policy, that is, for each subject $j \in \Omega$, an optimal classification is given by:*

$$x_j^{*E} = \begin{cases} 1, & \text{if } \widehat{p}_j \geq p_{th}^{*E} \\ 0, & \text{if } \widehat{p}_j < p_{th}^{*E} \end{cases},$$

where $p_{th}^{*E} = \frac{c_{FP}}{c_{FN}+c_{FP}}$.

The *risk-based threshold* policy prescribed in Theorem 1 depends on a threshold, p_{th}^{*E}, on the probability of positivity (risk) of a subject, and the threshold is a function of the misclassification cost parameters only. Theorem 1 leads to an equivalent formulation of **EM** in which the binary decision vector, $\overrightarrow{x}$, is replaced by a single threshold value.

Corollary 1. *An equivalent formulation for **EM** follows:*

***Problem EM**:*

$$minimize_{p_{th}\in[0,1]} E\left[C(p_{th})|\,\overrightarrow{\widehat{p}}\right] = c_{FN} \sum_{j\in\Omega:\widehat{p}_j<p_{th}} \widehat{p}_j + c_{FP} \sum_{j\in\Omega:\widehat{p}_j\geq p_{th}} (1-\widehat{p}_j), \tag{2.3}$$

with an optimal solution given by $p_{th}^{*E} = \frac{c_{FP}}{c_{FN}+c_{FP}}$.

Next we analyze the robust formulation, **RM**. For this purpose, we first characterize the structural properties of the *Regret* function.

Lemma 1. *For any given classification outcome for subject* j, x_j, $j \in \Omega$, *the maximum Regret function can be characterized as follows:*

$$max_{p_j \in [\underline{p}_j, \overline{p}_j]} \Big\{ Regret(x_j, p_j) \Big\} = \begin{cases} \Big(\overline{p}_j (c_{FN} + c_{FP}) - c_{FP} \Big)^+, & if\ x_j = 0 \\ \Big(c_{FP} - \underline{p}_j (c_{FP} + c_{FN}) \Big)^+, & if\ x_j = 1 \end{cases}.$$

Lemma 1 allows us to reformulate **RM** as a tractable optimization problem.

Corollary 2. *An equivalent formulation for* ***RM*** *follows:*

Problem RM*:*

$$minimize_{\overrightarrow{x}=(x_j)_{j\in\Omega}} \sum_{j\in\Omega} \Big(max_{p_j \in [\underline{p}_j, \overline{p}_j]} \Big\{ Regret(x_j, p_j) \Big\} \Big) =$$

$$\sum_{j\in\Omega} \Big[(1 - x_j) \Big(\overline{p}_j (c_{FN} + c_{FP}) - c_{FP} \Big)^+ + x_j \Big(c_{FP} - \underline{p}_j (c_{FP} + c_{FN}) \Big)^+ \Big]$$

$$subject\ to \qquad x_j\ binary, \forall j \in \Omega.$$

Theorem 2. *Given an uncertainty set around the true risk vector,* $S(\overrightarrow{P}) = \Big([\underline{p}_j, \overline{p}_j] \Big)_{j\in\Omega}$, *an optimal* ***RM*** *solution follows a risk-based threshold policy, that is, for each subject* $j \in \Omega$, *an optimal classification is given by:*

$$x_j^{*R} = \begin{cases} 1, & if\ \frac{\overline{p}_j + \underline{p}_j}{2} \geq p_{th}^{*E} \\ 0, & if\ \frac{\overline{p}_j + \underline{p}_j}{2} < p_{th}^{*E} \end{cases},$$

where $p_{th}^{*E} = \frac{c_{FP}}{c_{FN}+c_{FP}}$.

Based on Theorems 1 and 2, optimal solutions to both **EM** and **RM** can be expressed in terms of a risk threshold, p_{th}^{*E}, which is compared with each subject's estimated risk, $\widehat{p}_j$,

to obtain an optimal **EM** solution, and with the average of the lower and upper bounds on the subject's risk, $\frac{\overline{p}_j+\underline{p}_j}{2}$, to obtain an optimal **RM** solution. Therefore, both policies are *risk-based threshold* policies.

In Appendix A.2, we also provide an equivalent formulation of the decision problem so as to link the **EM** objective, i.e., the minimization of the expected misclassification cost, to an objective function commonly considered in the literature, i.e., the maximization of a weighted sum of test sensitivity and specificity (e.g., [32, 42, 68, 69, 81]). This reformulation proves to be especially useful when accurately estimating the subject misclassification costs, i.e., c_{FP} and c_{FN}, is difficult. This is often the case, because c_{FN}, the cost of a false negative, represents the cost of a missed diagnosis, i.e., the cost of poor health outcomes resulting from a missed or delayed diagnosis, including fatality; and c_{FP}, the cost of a false positive, depends on the entire screening process, i.e., the cost and accuracy of further tests conducted if the subject is classified as a positive by the biomarker test. In particular, this reformulation allows the tester to define target levels for test sensitivity and specificity, rather than specify misclassification costs.

2.5 Comparison of RM and EM Solutions, and Risk Estimation

In this section, we derive analytical expressions on the price of robustness and price of expectation-based optimization, provide some examples of the risk estimation function $g(.)$, and discuss further properties of **EM** and **RM**.

2.5.1 Price of Robustness and Price of Expectation-based Optimization

The **RM** solution, by relying solely on an uncertainty set around the disease risk, provides a robust solution that may be sub-optimal for minimizing the expected misclassification cost. On the other hand, the **EM** solution, by relying on a point estimate, $\overrightarrow{\widehat{p}}$, of the true risk vector, $\overrightarrow{P}$, may also deviate from the solution that achieves the minimum possible expected misclassification cost, i.e., $\overrightarrow{x}^*(\overrightarrow{p})$, with expected cost, $E[C(\overrightarrow{x}^{*E}(\overrightarrow{p}))]$. Then, an important policy question is which of these models, **RM** or **EM**, would perform better for designing a biomarker screening policy. In order to answer this question, in what follows, we study two related performance measures: the *price of robustness* (Π^R), and the *price of expectation-based optimization* (Π^E), which respectively correspond to the deviation of the **RM** and **EM** optimal solution values from the minimum expected misclassification cost, when the true disease risk vector, $\overrightarrow{p}$, is perfectly observable, that is,

$$\Pi^R(\overrightarrow{p}) \equiv E[C(\overrightarrow{x}^{*R})] - E[C(\overrightarrow{x}^{*E}(\overrightarrow{p}))], \quad \textit{and} \quad \Pi^E(\overrightarrow{p}) \equiv E[C(\overrightarrow{x}^{*E}(\overrightarrow{\widehat{p}}))] - E[C(\overrightarrow{x}^{*E}(\overrightarrow{p}))].$$

Thus, higher values of Π^R and Π^E respectively indicate that **RM** and **EM** solutions deviate further from the minimum possible expected misclassification cost.

Theorem 3. *For a risk vector realization $\overrightarrow{p}$, the price of robustness, $\Pi^R(\overrightarrow{p})$, and the price of expectation-based optimization, $\Pi^E(\overrightarrow{p})$, can be expressed as follows:*

$$\begin{aligned}
\Pi^R(\overrightarrow{p}) &= \sum_{j\in\Omega:\, \substack{p_j \geq p^{*E}_{th}, \\ \frac{p_j+\overline{p}_j}{2} < p^{*E}_{th}}} \Big[p_j(c_{FP}+c_{FN}) - c_{FP}\Big] + \sum_{j\in\Omega:\, \substack{p_j < p^{*E}_{th}, \\ \frac{p_j+\overline{p}_j}{2} \geq p^{*E}_{th}}} \Big[c_{FP} - p_j(c_{FP}+c_{FN})\Big], \\
\Pi^E(\overrightarrow{p}) &= \sum_{j\in\Omega:\, \substack{p_j \geq p^{*E}_{th}, \\ \widehat{p}_j < p^{*E}_{th}}} \Big[p_j(c_{FP}+c_{FN}) - c_{FP}\Big] + \sum_{j\in\Omega:\, \substack{p_j < p^{*E}_{th}, \\ \widehat{p}_j \geq p^{*E}_{th}}} \Big[c_{FP} - p_j(c_{FP}+c_{FN})\Big].
\end{aligned}$$

Corollary 3. *For a risk vector realization* $\overrightarrow{p}$*, we have the following:*

$$\Pi^E(\overrightarrow{p}) - \Pi^R(\overrightarrow{p}) = \sum_{\substack{j\in\Omega:\ \widehat{p}_j<p^{*E}_{th},\\ \frac{\underline{p}_j+\overline{p}_j}{2}\geq p^{*E}_{th}}} \Big[p_j(c_{FN}+c_{FP}) - c_{FP}\Big] + \sum_{\substack{j\in\Omega:\ \widehat{p}_j\geq p^{*E}_{th},\\ \frac{\underline{p}_j+\overline{p}_j}{2}< p^{*E}_{th}}} \Big[c_{FP} - p_j(c_{FN}+c_{FP})\Big].$$

Corollary 4. *If* $\widehat{p}_j = \frac{\overline{p}_j+\underline{p}_j}{2}$*,* $\forall j \in \Omega$*, then* $\overrightarrow{x}^{*E} = \overrightarrow{x}^{*R}$*, and hence, the price of robustness and the price of expectation-based optimization are equal, i.e.,* $\Pi^R(\overrightarrow{p}) = \Pi^E(\overrightarrow{p})$*,* $\forall \overrightarrow{p} \in S(\overrightarrow{P})$*.*

In Section 2.6, we show, via a numerical study, that under different conditions, each of **RM** or **EM** could be a better choice for the tester for minimizing the expected misclassification cost.

2.5.2 Risk Estimation Function

In this section, we study how the risk estimation function $g(.)$, which maps each subject's biomarker reading to their disease risk, i.e., $g(\widehat{y},\widetilde{y}) = \widehat{p}$ (see Table 2.1), impacts the price of robustness and the price of expectation-based optimization. Recall that in our setting, of noisy biomarker readings, common and subject-specific perturbations, if present, are removed via the $h(.)$ function.

Remark 2. *Given a processed biomarker level,* $\widehat{y}$*, and a biomarker reading,* $\widetilde{y}$*, one can estimate the subject disease risk via, for example, a logistic regression model [5, 37, 60, 82], e.g.,* $g(\widehat{y},\widetilde{y}) = \frac{1}{1+e^{-(a+b(\widetilde{y}-\widehat{y}))}}$*, where* a *and* b *are some constants, and* $b > 0$ *(see Section 2.6).*

Remark 3. *The* $g(.)$ *function derived by the logistic regression model in Remark 2 is non-decreasing in* $\widetilde{y}$*, non-increasing in* $\widehat{y}$*, and is S-shaped in* $\widetilde{y}-\widehat{y}$*, i.e., it is first convex increasing, then concave increasing, and converging to 1 (see Figure 2.1 as an example). This follows*

because letting $z = \widetilde{y} - \widehat{y}$*, where* $z \in (-\infty, \infty)$*, we can write*

$$g(\widehat{y}, \widetilde{y}) = \frac{1}{1 + e^{-(a+b(\widetilde{y}-\widehat{y}))}} = \frac{1}{1 + e^{-(a+b(z))}} \quad \Rightarrow \quad \frac{\partial g(z)}{\partial z} > 0$$

$$\Rightarrow \begin{cases} \frac{\partial^2 g(z)}{\partial z^2} > 0, & \text{if } z < \frac{-a}{b} \\ \frac{\partial^2 g(z)}{\partial z^2} < 0, & \text{if } z > \frac{-a}{b} \end{cases}$$

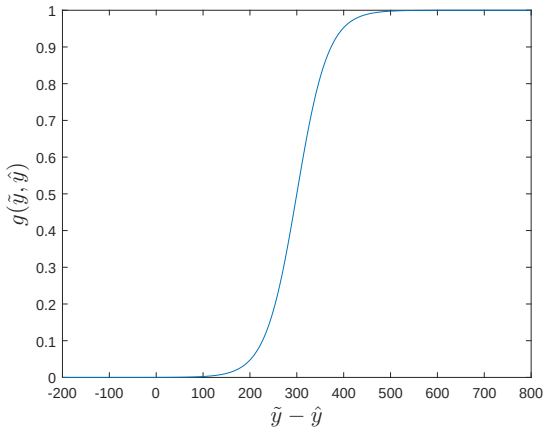

Figure 2.1: $g(.)$ function that corresponds to the logistic regression model in Remark 2, when $a = -9$ and $b = 0.03$.

Therefore, in the following we discuss the implications of S-shaped $g(.)$ functions on our results. Specifically, S-shaped risk estimation functions are less sensitive to perturbations in biomarker readings when the difference, $\widetilde{y} - \widehat{y}$, is very low (e.g., $(\widetilde{y} - \widehat{y}) \in (-\infty, 100]$ in Fig. 2.1), or very high (e.g., $(\widetilde{y} - \widehat{y}) \in [500, +\infty)$ in Fig. 2.1). This implies that the performance of the **RM** solution (i.e., the deviation from the true optimal solution for a given $\overrightarrow{p}$) will vary for the different subjects, depending on their $\widetilde{y} - \widehat{y}$ value. This follows because the uncertainty set around P is likely to be narrower when the value of $\widetilde{y} - \widehat{y}$ is either very low or very high. To illustrate this last point, recall that the uncertainty set of $\overrightarrow{\widetilde{Y}}$, i.e., $S(\overrightarrow{\widetilde{Y}}) = \left([\underline{y}_j, \overline{y}_j]\right)_{j\in\Omega}$, can be used to construct an uncertainty set around $\overrightarrow{P}$, i.e., $S(\overrightarrow{P}) = \left([g(\underline{y}_j, \widetilde{y}_j), g(\overline{y}_j, \widetilde{y}_j)]\right)_{j\in\Omega} = \left([\underline{p}_j, \overline{p}_j]\right)_{j\in\Omega}$ (see Remark 1), and consider the following example.

Example 1. *Consider that three subjects are tested in a given period, with the testing outcomes reported in Table 2.2, and $g(.)$ function given by Remark 2, with $a = -9$ and $b = 0.03$:*

Table 2.2: Testing outcomes for Example 1

Subject	$\widetilde{y}$	$\widehat{y}$	$\widetilde{y} - \widehat{y}$	$[\underline{y}, \overline{y}]$	$[\underline{p}, \overline{p}] = [g(\widetilde{y} - \overline{y}), g(\widetilde{y} - \underline{y})]$	$\frac{p+\overline{p}}{2}$	$max_{P \in [\underline{p}, \overline{p}]} \left\{ \lvert P - \frac{p+\overline{p}}{2} \rvert \right\}$
1	700	800	-100	$[750, 850]$	$(g(-150), g(-50)) = (1.37 \times 10^{-6}, 2.75 \times 10^{-5})$	1.44×10^{-5}	1.31×10^{-5}
2	700	400	300	$[350, 450]$	$(g(250), g(350)) = (1.824 \times 10^{-1}, 8.176 \times 10^{-1})$	5×10^{-1}	3.2×10^{-1}
3	700	100	600	$[50, 150]$	$(g(550), g(650)) = (9.994 \times 10^{-1}, 1.00)$	9.997×10^{-1}	3×10^{-4}

*From Table 2.2, the maximum deviation of the disease risk used in **RM**, i.e., $\frac{p+\overline{p}}{2}$ (Theorem 2), from the true disease risk is at most 3×10^{-4} for subjects 1 and 3, which have the same biomarker reading of 700, but respective biomarker uncertainty sets of $[750, 850]$ and $[50, 150]$ (due to different subject-specific attributes), leading to different values of $\widehat{y}_1$ and $\widehat{y}_2$ (see Table 2.2); and this deviation can be as high as 0.32 for subject 2, with the same biomarker reading of 700, and a biomarker uncertainty set of $[350, 450]$, which translates into a wider risk uncertainty set of $(0.1824, 0.8176)$. Hence, the optimal **RM** classification for subject 2 may not be highly reliable.*

Remark 4 provides another example of the risk estimation function, $g(.)$, using a Bayesian framework.

Remark 4. *Let $\widehat{Y}_+$ and $\widehat{Y}_-$ respectively denote the processed biomarker level of a random true positive and a random true negative subject, with respective probability density functions (pdf) of $f_{\widehat{Y}_+}$ and $f_{\widehat{Y}_-}$, and let q denote the disease prevalence rate within the population. Then, $g(\widehat{y}) = P(D = 1 | \widehat{Y} = \widehat{y}) = \frac{q f_{\widehat{Y}_+}(\widehat{y})}{q f_{\widehat{Y}_+}(\widehat{y}) + (1-q) f_{\widehat{Y}_-}(\widehat{y})}$.*

Our analysis of realistic distributions of $\widehat{Y}_+$, $\widehat{Y}_-$, and values of the prevalence rate q suggest that the $g(.)$ function in Remark 4 is also an S-shaped function.

Remark 5. *The distribution and parameters of random variables $\widehat{Y}_+$ and $\widehat{Y}_-$ can be estimated by using, for example, training data and Monte Carlo simulation (e.g., [50]), i.e.,*

by assuming different parametric models for $\widehat{Y}_+$ and $\widehat{Y}_-$, generating random samples from these models, and using, for example, the maximum likelihood estimator, to estimate the distribution parameters.

2.6 Case Study: Newborn Screening for Cystic Fibrosis

In this section, we perform a case study of cystic fibrosis (CF) screening for newborns. In the US, every state has a program that screens newborns for a panel of genetic diseases (using dried blood spots routinely obtained from the newborns). With a prevalence rate of approximately 1 in 3,700 newborns in the US [12, 58], CF is one of the most prevalent genetic diseases, and is included in every state's newborn screening panel [2, 51]. Newborn screening for CF allows for early diagnosis, and can substantially improve health outcomes [24, 43]. Newborns with false negative screening results experience a delayed diagnosis, which complicates the treatment process, and may result in poor health outcomes, including severe malnutrition, lung disease, and fatality [30, 63]. On the other hand, false positive screening results cause parental distress and result in further, expensive tests, including genetic tests, and the diagnostic sweat chloride test, which is too expensive for screening purposes and must be performed at a specialized testing facility [2, 18, 22, 41, 42, 73].

Due to the importance of timely results, state laboratories perform IRT tests daily. Newborn screening for CF is performed via a screening process, which refers to the sequence of tests and policies for interpreting their results. While CF screening processes vary between states [2], all states start the CF screening process with a biomarker test that measures the concentration of immunoreactive trypsinogen (IRT) in the blood, i.e., the IRT test [35]. As discussed earlier, newborns with CF tend to have elevated IRT levels [20, 29, 47, 66], although the distributions of IRT concentrations for CF positive and CF negative newborns

have a wide variance and overlap. Consequently, CF classification based on IRT readings is not perfectly reliable, and all states use further, expensive tests for newborns that are classified as IRT test positive, as discussed above [42].

The IRT biomarker fits the modeling assumptions discussed in Section 2.2. Specifically, IRT readings are affected by external factors that are *common* for all newborns tested in the same period, leading to a positive correlation among the test readings, i.e., $\overrightarrow{\widetilde{y}}$. For example, low temperatures tend to increase the IRT levels of *all* newborns, with or without CF [43, 72]. IRT levels are also affected by subject-specific attributes (e.g., birth weight, gender, and race), leading to variations that occur independently for each newborn [18, 43, 55, 72]. In fact, our analysis of a five-year data set of CF newborn screening results in North Carolina confirms, and quantifies, the dependence of IRT levels to the newborn's birth weight, gender, and race, as well as seasonality. Our analysis also indicates that there is a correlation between race and birth weight, and between gender and birth weight, and we incorporate all these factors into a regression model to estimate the probability that a newborn is CF positive, as detailed in Section 2.6.3.

The remainder of this section is organized as follows. In Section 2.6.1, we provide an overview of current IRT screening policies used for CF newborn screening. In Section 2.6.2, we discuss our data sources and calibration. In Section 2.6.3, we model the relationship between IRT readings and true IRT levels through regression analysis. In Section 2.6.4, we perform a case study to compare the proposed optimal policies with current IRT screening policies.

2.6.1 Current IRT Screening Policies

The IRT screening policies currently used in the US fall into the following two classes:

Concentration-based threshold policy (CB) is characterized by an IRT reading threshold, $\widetilde{y}_{th}$, such that the newborn is classified as test positive (i.e., $x = 1$) if $\widetilde{y} \geq \widetilde{y}_{th}$, and is

classified as test negative (i.e., $x = 0$), otherwise. As examples, California uses a **CB** policy with an IRT reading threshold of 62 ng/mL [42], while Washington uses a threshold of 100 ng/mL [72].

Proportion-based threshold policy (PB) is characterized by a proportion r, such that the newborn is classified as test positive (i.e., $x = 1$) if the reading $\widetilde{y}$ is in the top $r\%$ of all IRT readings in a given day, and is classified as test negative (i.e., $x = 0$), otherwise. That is, letting $\widetilde{y}_{(1)} \geq \widetilde{y}_{(2)} \geq ... \geq \widetilde{y}_{(N)}$ represent an ordered set of IRT readings in a random day with N subjects, subjects $(1), (2), ..., (\lceil rN \rceil)$ in the ordered set will be classified as test positive. As examples, Wisconsin and North Carolina use a **PB** policy with a proportion threshold of 4% [2, 43], while Massachusetts uses a **PB** policy with a threshold of 5% [21].

Although the IRT threshold has a large impact on the overall sensitivity and specificity of the CF screening algorithm, there are no nationwide guidelines on how the threshold should be set. Some studies evaluate the performance of a particular threshold policy for the IRT test, in terms of the sensitivity and specificity [42, 43, 55, 72]. Therrell *et al.* [72] state that **PB** outperforms **CB**, especially in regions that experience higher fluctuations in seasonal temperatures, and Kloosterboer *et al.* [43] suggest using **PB** to take into account the impact of common external factors. Observe that under the **PB** policy, the corresponding IRT reading threshold *varies* each day in a random manner.

Both the **CB** threshold, $\widetilde{y}_{th}$, and the **PB** threshold, r, are determined *prior* to observing the IRT readings of each day, and remain constant for all days. Conversely, **EM** and **RM** utilize the IRT readings in a given period to estimate the CF risk for each newborn.

2.6.2 Data Sources and Calibration

We perform a case study based on a data set from the North Carolina State Laboratory of Public Health (NCSLPH), which contains CF newborn screening outcomes for North Carolina over a five-year period, corresponding to 1,359 testing days; and also provides the

IRT test date, gender, race, and birth weight for each newborn tested in the study period, as well as the outcome of the diagnostic sweat chloride test, i.e., the true CF status, for those newborns classified as test positive in screening. The data set contains a small number of newborns with incomplete information, which are removed from the data set, resulting in 569,601 newborns. Following the data set, we consider four racial groups, Caucasian, African American, Hispanic, and Asian, with a total of 107 identified CF positive cases over five years, with IRT readings of the CF positive cases varying between 43.4 ng/mL and 502 ng/mL. Table 4.1 summarizes the demographic characteristics of newborns screened by the NCSLPH during the study period.

Table 2.3: Demographic characteristics of newborns in the NCSLPH data set (five-year period)

Race	*Proportion*	*# Newborns screened*	*# CF cases*	*Average IRT±SD[a] (ng/mL)*	*Average weight±SD[a] (gr)*
Caucasian	58.3%	332,303	94	22.97 ± 0.03	3,287.92 ± 2.76
African American	25.8%	146,646	7	29.26 ± 0.04	2,984.67 ± 4.03
Hispanic	12.7%	72,244	6	22.41 ± 0.06	3,260.10 ± 4.67
Asian	3.2%	18,408	0	21.91 ± 0.11	3,169.61 ± 6.63
Overall	100%	569,601	107	24.48 ± 0.02	3,202.38 ± 0.94

[a] SD denotes the standard deviation around the mean

Our objective is to study the performance of the proposed optimal data-driven policies (**EM** and **RM** policies) over various current IRT screening policies: **(1) CB** policy with IRT reading thresholds of: 55 ng/mL (Georgia), 60 ng/mL (Colorado), 62 ng/mL (California), and 100 ng/mL (Washington) [2, 72]; and **(2) PB** policy with proportion thresholds of: 4% (Florida, North Carolina, Wisconsin [43, 72]), and 5% (Texas, New York, Massachusetts [72]), as well as other possible **CB** and **PB** policies.

As discussed in Section 2.4, it is difficult to estimate the costs of misclassification (c_{FN} and c_{FP}), especially the cost of a false negative, which represents the cost of a missed CF case, i.e., the cost of poor health outcomes resulting from missed or delayed diagnosis. Hence, in our numerical study, we perform a one-way sensitivity analysis on the cost ratio, $k = \frac{c_{FN}}{c_{FP}}$.

We divide the data set into two disjoint sets, *validation data set* (40% of the data set,

corresponding to the first two years) and *training data set* (the remaining 60% of the data set, corresponding to the last three years). The validation data set in our study is relatively large (227,840 newborns), to ensure that it contains a sufficient number of CF positive cases (i.e., 46 identified CF cases) for evaluating the performance of the different IRT policies. However, we do not have reliable data on false negative (i.e., missed CF) cases. Hence, we calibrate our validation data set by randomly adding some CF cases, based on CF prevalence rates for the different races from the literature, so as to match the sensitivity levels reported in the literature; see Appendix C.1. As a result, the existing 46 CF positive cases in the validation data set are augmented by 1.73 ± 0.16 (average $\pm$ SD) CF positive cases based on Monte Carlo simulation, leading to a total of 47.73 ± 0.16 CF cases (Appendix C.1).

2.6.3 The Regression Model

In this section, we develop a two-step regression approach, through the use of $h(.)$ and $g(.)$ functions, to predict the CF risk for each newborn based on their attributes and external factors. The reason for a two-step regression, rather than a single-step binary logistic regression, is that there are certain subject-specific and external factors (e.g., birth weight, gender, and seasonality) that affect *only* the newborn's IRT concentration level, and not their risk of CF. For example, cold weather tends to increase the IRT concentration level in all subjects, but does not alter the CF risk [43]. Thus, while there is no direct correlation between these factors and the CF status, these factors impact our analysis by altering the newborn's IRT level. The proposed two-step regression approach addresses this issue. Specifically, in the first step, it estimates an expected IRT level for each newborn ($\widehat{y}$), based on a linear regression model that considers both external factors and newborn-specific attributes (i.e., through the $h(.)$ function); and in the second step, it estimates the CF risk for each newborn, based on a logistic regression model that considers the discrepancy between their measured IRT reading ($\widetilde{y}$) and the expected IRT level from Step 1 ($\widehat{y}$), (i.e., through the $g(.)$

function). Then, the optimal risk-based threshold policy (**EM** or **RM**) is used to classify the newborn either as a test positive or a test negative (see Appendix C.2 for a comparison of the proposed two-step regression approach with a single-step logistic regression approach).

Our analysis of the data set indicates that birth weight, gender, race, and seasonality each has a significant effect on the IRT level. Moreover, there is correlation between race and birth weight, and between gender and birth weight. We apply the *backward stepwise variable selection* method (e.g., [27]) on the training data set, to select the "best" subset of variables to include in both the first-step linear regression and the second-step logistic regression. Specifically, in the first step, we start with a linear regression that includes all the aforementioned variables, determine its performance (in terms of the root mean squared error (RMSE) [27]), and rank the variables based on their individual impact on the dependent variable (the IRT level). Then we iteratively remove the "least useful" variable (i.e., the variable that is the least statistically significant in each iteration), rank the remaining variables, and repeat this process until all but one variable remains. Finally, we choose the variable set with the best performance (i.e., the lowest RMSE) [27]. The stepwise variable selection method indicates that the following subset of variables should be included in the linear regression: birth weight, gender, race, seasonality, and race-weight correlations for Caucasians, African Americans, and Hispanics (the weight-gender correlation, and the race-weight correlation for Asians are eliminated).

Next, to construct our first-step regression model, we perform a linear regression analysis on the training data set and estimate the dependent variable, i.e., the IRT level for newborn j ($\widehat{Y}_j$), based on the selected variables, where W_j denotes birth weight, R_j denotes race ($R_j^{AF} = 1$ for African American, $R_j^H = 1$ for Hispanic, $R_j^A = 1$ for Asian; and Caucasian is the default value, i.e., $R_j = 0$), G_j denotes gender ($G_j = 1$ if female, and 0 otherwise), and $\bar{\widetilde{y}}_t^R$ denotes the rolling average of IRT readings used in period t to account for seasonality, i.e., the average of all IRT readings over the most recent five testing days. (Our analysis indicates that using a rolling IRT average of five days is sufficient to model seasonality.)

Moreover, we perform a repeated *five-fold cross validation with stratified sampling*, applied to the training data set, to tune the parameters of the linear regression (e.g., [45]). In particular, we randomly partition the training data set into five (almost) equal subsets, with approximately equal proportions of CF positive and CF negative newborns in each subset. Then, we choose one of the subsets to serve as the validation set, and use the remaining four subsets to train the linear regression model. We repeat this process five times, i.e., until each subset is used exactly once as the validation set, and repeat the entire process 10 times, with 10 different random seeds to partition the training data set. The resulting linear regression equation follows:

$$\begin{aligned}
\widehat{y}_j &= h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}}) = E(Y_j | W_j = w_j, R_j^{AF} = r_j^{AF}, R_j^H = r_j^H, R_j^A = r_j^A, G_j = g_j, \widetilde{\widetilde{Y}}_t^R = \widetilde{\widetilde{y}}_t^R) \quad (2.4)\\
&= 1.233 - 8.037 \times 10^{-4}\, w_j + 0.8115\, g_j + 4.525\, r_j^{AF} - 1.228\, r_j^H \\
&- 1.113\, r_j^A + 0.9799\, \widetilde{\widetilde{y}}_t^R + 4.974 \times 10^{-4}\, w_j\, r_j^{AF} + 2.117 \times 10^{-4}\, w_j\, r_j^H, \qquad j \in \Omega,
\end{aligned}$$

with a p-value less than 2.2×10^{-16}. Thus, Eq. (2.4) provides an expected IRT level for each newborn, given their specific attributes, except for their CF status, and external factors.

We note that in general, the $h(.)$ function does not have to be linear; its functional form depends on how subject-specific and external factors impact the biomarker level. For example, a biomarker level may vary over time in a non-linear manner [67]; thus, if the biomarker level is measured over time (as opposed to a one-time measurement, as is done here), and time is one of the selected variables, then the $h(.)$ function will also be non-linear.

In the second step, we consider the difference between the expected IRT level calculated by Eq. (2.4) and the IRT measurement, i.e., $\widetilde{y}_j - \widehat{y}_j$ (see Remark 2) as the statistic of interest, and perform a logistic regression, in which the dependent variable is the CF risk, and the independent variables, selected by the backward stepwise variable selection method discussed above (with the Akaike Information Criterion (AIC) used as the performance metric [45]), include $\widetilde{y}_j - \widehat{y}_j$, r_j^{AF}, r_j^H, and r_j^A. The reason that race remains in the selected variable set is

that it has a two-fold effect on the IRT test: (i) race affects the IRT levels of newborns, i.e., the average IRT level differs significantly among the different races (e.g., African Americans have significantly higher IRT levels than other races, see Table 4.1 and Eq. (4)); and (ii) race affects the CF risk of newborns, i.e., CF prevalence rate differs significantly among the different races, see Table A1. In the logistic regression, we consider the n^{th} root of $(\widetilde{y}_j - \widehat{y}_j)$, i.e., $(\widetilde{y}_j - \widehat{y}_j)^{\frac{1}{n}}$, because this functional form provides an S-shaped function with respect to $(\widetilde{y}_j - \widehat{y}_j)$, which is less sensitive to very high or very low values of the difference between the measured and the expected IRT. To find the "best" value of n, we perform a grid search (e.g., [17]) on $n \in \{1, 3, 5, \cdots, 97, 99\}$, i.e., only the odd values of n, because the difference can be negative; and find that the best value of n is 3. Finally, we perform a repeated five-fold cross validation with stratified sampling, applied to the training data set, to tune the parameters of the logistic regression. The resulting logistic regression equation follows:

$$\widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j) = E(D_j | \widetilde{y}_j - \widehat{y}_j) = \frac{1}{1 + e^{(13.30609 - 2.04352\,(\widetilde{y}_j - \widehat{y}_j)^{\frac{1}{3}} + 2.44\,r_j^{AF} + 0.96496\,r_j^{H} + 14.55\,r_j^{A})}}, \quad j \in \Omega \tag{2.5}$$

with a p-value less than 2.2×10^{-16}.

We next use this two-step regression model, in conjunction with each optimization model (**EM** or **RM**), on the validation data set, to compare their performance with current IRT policies. To this end, we use the linear regression and logistic regression in Eq.s (2.4) and (2.5) to derive an estimated CF risk (to be used by the **EM** policy), as well as an uncertainty set around it (to be used by the **RM** policy) for each newborn in the validation data set. Specifically, for the **EM** policy, we calculate $\widehat{y}_j = h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}})$ (via Eq. (2.4)) and $\widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j)$ (via Eq. (2.5)), while for the **RM** policy, we calculate the 95% CI around $\widehat{y}_j$, given by $\underline{y}_j$ and $\overline{y}_j$, leading to $\underline{p}_j = g(\underline{y}_j, \widetilde{y}_j)$ and $\overline{p}_j = g(\overline{y}_j, \widetilde{y}_j)$ (via Eq. (2.5)). We then respectively compare $\widehat{p}_j$ and $\frac{\overline{p}_j + \underline{p}_j}{2}$ with the optimal risk thresholds for the **EM** and **RM** policies (Theorems 1 and 2), and classify each newborn as test positive or test negative. Then, we compute the number of false negatives and false positives, based on the true CF status of each newborn,

for all newborns in the validation data set, for each simulation replication (used solely to generate additional CF cases, as described in Section 6.2 and Appendix C.1), leading to the total expected misclassification cost, and derive the sensitivity and specificity of each policy.

2.6.4 Case Study Results

In this section, we discuss the case study results on the validation data set, which contains 227,840 subjects. We compare the **EM** and **RM** policies with various current IRT policies of **CB**: 55 ng/mL, 60 ng/mL, 62 ng/mL, and 100 ng/mL; **PB**: 4% and 5%. Moreover, we consider additional **PB** (1%-6%) and **CB** (45 ng/mL-65 ng/mL) policies, to find the "best" **PB** and **CB** policy.

Our results are based on 400 Monte Carlo simulation replications, which are solely used to randomly generate additional CF positive cases that are likely missed under the current IRT policy used in North Carolina, as discussed in Section 2.6.2. Recall that there are 46 CF positive cases in the validation data set, and 1.73 ± 0.16 (average $\pm$ SD) CF positive cases are added based on simulation, leading to 47.73 ± 0.16 CF cases.

Tables 2.4-2.5 and A3-A6 (Appendix C.1) report the results of our case study for the validation data set, including the number of false positives, false negatives, and the misclassification cost for 227,840 newborns over the two-year period. Specifically, Table 2.4 reports the average number of false negatives and false positives, and the average sensitivity (the ratio of CF positive newborns who are classified as test positive by the IRT test to all CF positive newborns), and specificity (the ratio of CF negative newborns who are classified as test negative by the IRT test to all CF negative newborns) for various **EM**, **RM**, and current **CB** and **PB** policies, while Tables A3-A6 report these performance measures and the misclassification cost for a larger set of **PB** (1%, 1.5%, $\cdots$, 5.5%, 6%) and **CB** (45 ng/mL, 46 n/mL, $\cdots$, 65 ng/mL) policies over 400 simulation replications. Finally, Table

2.5 reports the average misclassification cost of all newborns, for each value of $k = \frac{c_{FN}}{c_{FP}}$ and for all policies considered, including the best **PB** and **CB** policy from Tables A3 and A5. We note that all costs are reported in terms of the cost ratio, k, i.e., assuming unit cost for c_{FP}, and are sufficient for our purposes of comparing the different policies. If one is interested in the actual misclassification cost, then each cost term needs to be multiplied by the cost of a false positive, i.e., c_{FP}, which represents the additional expected cost of testing if the IRT test outcome is positive (e.g., the genetic test, followed by the sweat chloride test if the genetic test indicates CF). The state's entire screening policy (i.e., sequence of tests and rules) affects the value of c_{FP}, and hence, the final misclassification cost value depends on the policy of each state.

The optimal threshold values, hence the resulting IRT sensitivity and specificity levels, for both **EM** and **RM** are dictated by the value of parameter k, i.e., as k increases, both **EM** and **RM** have higher sensitivity but lower specificity (see Appendix A.2). To study this aspect, Fig. 2.2 plots the sensitivity and specificity of the various IRT screening policies considered, as well as those of the **EM** and **RM** policies for the different values of k reported in Tables 2.4, A3, and A5. Fig. 2.2 indicates that for each given sensitivity (specificity) level, both **EM** and **RM** provide a higher specificity (sensitivity) level than **PB** and **CB** policies. For example, the **PB** 4% policy provides a sensitivity of 94.28% and a specificity of 95.89%, while the **EM** and **RM** policies (for $k = 2,000$) provide both higher sensitivity (95.98% and 95.56%, respectively) and higher specificity (96.58% and 96.79%, respectively).

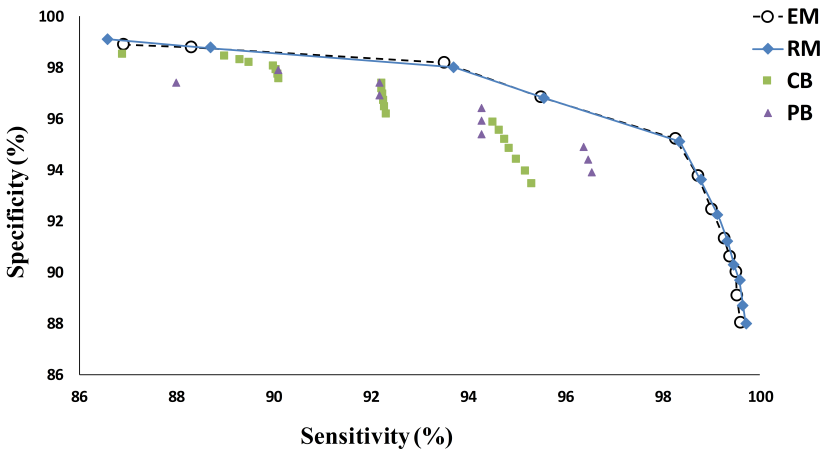

Figure 2.2: Sensitivity versus specificity of various IRT screening policies

Table 2.4: Performance of various IRT screening policies (Validation data set)

	Policy	*False negatives (95% half width)*	*False positives*	*Sensitivity*	*Specificity*
	CB (55)	3.70 (0.16)	6,996	92.25%	96.93%
	CB (60)	4.75 (0.16)	4,858	90.05%	97.87%
	CB (62)	5.01 (0.16)	4,213	89.50%	98.15%
	CB (100)	13.73 (0.16)	513	71.23%	99.77%
	PB (4%)	2.73 (0.16)	9,350	94.28%	95.89%
	PB (5%)	1.73 (0.16)	11,636	96.37%	94.89%
k=1,000	**EM**	5.58 (0.16)	3,248	88.31%	98.57%
	RM	5.55 (0.16)	3,294	88.37%	98.55%
k=2,000	**EM**	2.16 (0.14)	7,168	95.48%	96.85%
	RM	2.12 (0.14)	7,303	95.56%	96.79%
k=3,000	**EM**	0.83 (0.11)	10,864	98.26%	95.23%
	RM	0.79 (0.11)	11,133	98.34%	95.11%
k=4,000	**EM**	0.61 (0.10)	14,189	98.72%	93.77%
	RM	0.58 (0.10)	14,548	98.78%	93.61%
k=5,000	**EM**	0.48 (0.08)	17,197	98.99%	92.45%
	RM	0.42 (0.08)	17,674	99.12%	92.24%
k=6,000	**EM**	0.36 (0.07)	19,817	99.25%	91.30%
	RM	0.32 (0.07)	20,091	99.33%	91.18%
k=7,000	**EM**	0.30 (0.07)	21,459	99.37%	90.58%
	RM	0.26 (0.07)	22,172	99.46%	90.27%
k=8,000	**EM**	0.24 (0.06)	22,811	99.50%	89.99%
	RM	0.20 (0.06)	23,606	99.58%	89.64%
k=9,000	**EM**	0.23 (0.05)	24,913	99.52%	89.06%
	RM	0.17 (0.05)	25,781	99.64%	88.68%
k=10,000	**EM**	0.19 (0.04)	26,875	99.60%	88.20%
	RM	0.14 (0.04)	27,448	99.71%	87.95%

Table 2.5: Average misclassification cost for various IRT screening policies, in terms of $k = \frac{c_{FN}}{c_{FP}}$ (Validation data set)

	k=1,000	*k=2,000*	*k=4,000*	*k=6,000*	*k=8,000*	*k=10,000*
CB (55)	10,696	14,396	21,796	29,196	36,596	43,996
CB (60)	9,608	14,358	23,858	33,358	42,858	52,358
CB (62)	9,223	14,233	24,253	34,273	44,293	54,313
CB (100)	14,243	27,973	55,433	82,893	110,353	137,813
PB (4%)	12,080	14,810	20,270	25,730	31,190	36,650
PB (5%)	13,366	15,096	18,556	22,016	25,476	28,936
Best **CB**	8,925	14,064	19,965	25,205	30,445	35,685
(from Table A4)	(**CB** 64)	(**CB** 61)	(**CB** 51)	(**CB** 51)	(**CB** 51)	(**CB** 51)
Best **PB**	9,391	13,387	19,125	22,016	25,476	28,936
(from Table A6)	(**PB** 1.5%)	(**PB** 2.5%)	(**PB** 3.5%)	(**PB** 5%)	(**PB** 5%)	(**PB** 5%)
EM	8,828	11,488	16,629	21,977	24,731	28,775
RM	8,844	11,543	16,868	22,011	25,206	27,448

In order to estimate the potential reduction in the misclassification cost as a result of the proposed **EM** and **RM** policies, we next estimate the cost of a false positive per newborn (c_{FP}) as \$68.74 per newborn, based on a newborn screening process consisting of a post-IRT genetic test (mutation panel test with a panel of 23 CF-related mutations) and the diagnostic sweat chloride test [79].[1] Then, for the case of $k = 4,000$, for example, the **EM** policy decreases the misclassification cost by at least $\$68.74 \times (20,270 - 16,629) = \$250,282$, while the **RM** policy decreases it by at least $\$68.74 \times (20,270 - 16,868) = \$233,853$ for a two-year period, in comparison to North Carolina's current IRT policy of **PB** with a threshold of 4%, see Table 2.5. Moreover, in Table 2.5, we can observe that under different conditions (i,e,. different values of k), each of the **EM** or **RM** policies can perform better than the other.

In summary, our case study indicates that the proposed **EM** and **RM** policies outperform the current IRT screening policies, and any **PB** or **CB** policy in general. From a practical perspective, it is also important to note that **EM** and **RM** policies are easily implementable (they are no more difficult to implement than the current policies), and provide a great

[1] Some states, including North Carolina, test for more mutations (e.g., North Carolina uses a mutation panel of 139 CF-related mutations), and some other states, such as California, use a two-tier genetic test (mutation panel and sequencing). In these cases, the expected post-IRT cost is likely higher than \$68.74.

level of flexibility by allowing the tester to customize the state's screening policy considering state-level inputs (e.g., demographics and climate), along with sensitivity and specificity targets. This is especially important for CF screening, because environmental and demographic characteristics can substantially differ among the states. The proposed methods and policies use these characteristics as inputs (via the $g(.)$ and $h(.)$ functions that can be fit based on a training data set from the state), allowing the screening policy to be customized for each state in an optimal manner.

2.7 Conclusions and Future Research Directions

We analyze the problem of determining an optimal biomarker testing and subject classification policy for non-infectious diseases under uncertainty on true biomarker levels, due to random perturbations caused by external and/or subject-specific factors. We study both expectation-based and robust formulations to minimize a function of the misclassification cost, derive key structural properties of optimal policies, and show that they follow risk-based threshold policies. Our case study on newborn screening for cystic fibrosis in North Carolina indicates that the proposed policies can substantially decrease the expected misclassification cost for the IRT test over current IRT screening policies for newborn screening for cystic fibrosis.

An important limitation of this work is the presence of missing data on false negative cases in the North Carolina data set, which was used in our case study of Section 2.6. We do not have reliable data on the false negative cases for cystic fibrosis, and therefore, we had to use simulation and data from the literature to randomly generate additional false negative cases. We did this because the sensitivity of some current IRT screening policies were higher in the data set compared to those reported in the literature, and also the prevalence rates of cystic fibrosis for some races were lower in the data set than those reported in the literature. Adding the few additional CF positive cases made the results better match the literature.

An important extension of this work is to determine optimal classification policies based on dynamic progression of biomarker levels over time. Biomarkers have many other uses, such as risk classification [4], monitoring the progression of a disease [57], or evaluating the effectiveness of a specific treatment [39]. In many of these cases, the biomarker value, by itself, is not necessarily the best criterion for decision-making; rather criteria reflecting the dynamic progression of the biomarker over time may be more accurate. Another important future direction is to determine optimal biomarker classification policies for infectious diseases where disease transmission is possible among subjects, i.e., the disease positivity status of subjects may be correlated.

We hope that this study motivates practitioners to consider using risk-based biomarker threshold policies, which can take into account biomarker perturbations due to both external and subject-specific factors, as well as establishing tracking systems to reliably detect false negative cases over time.

Chapter 3

The Effect of Seasonality on the Immunoreactive Trypsinogen Test in Newborn Screening for Cystic Fibrosis

3.1 Introduction

Newborn screening (NBS) has been a major public health success in the United States (US), saving lives and preventing disability in thousands of newborns every year [1]. Each year, millions of newborns in the US are screened shortly after birth, for a number of genetic conditions that are treatable, but not clinically evident in the newborn. Our focus is on newborn screening for cystic fibrosis (CF), which has a prevalence of approximately 1 in 3,700 newborns in the US [12, 58], and is included in the screening panel for all fifty states and the District of Columbia [2]. Newborn screening for CF allows for early diagnosis of the disease, and can substantially improve health outcomes. Newborns diagnosed with CF through newborn screening have improved nutritional status and growth, and require less

hospitalization [19, 24, 43], whereas a delayed diagnosis can result in severe malnutrition, lung disease, or even fatality [30, 46, 63]. However, there is a possibility of misclassification, i.e., classifying a newborn with CF as test negative (i.e., a false negative), or classifying a newborn without CF as test positive (i.e., a false positive). A false negative outcome leads to a delayed diagnosis or a missed case, whereas a false positive outcome leads to unnecessary, further tests and parental distress. Therefore, the CF newborn screening scheme (i.e., the test sequence and classification rules) should be designed considering the trade-off between, and the consequences of, false negative and false positive outcomes.

While CF newborn screening schemes differ among states, the first screening test used by all states is a low-cost biomarker test that measures the concentration (level) of immunoreactive trypsinogen (IRT) (ng/mL) in blood [35], i.e., the IRT test. Newborns with CF tend to have elevated IRT levels [18, 20, 29, 47, 66]. Dried blood spots (DBS) from newborns are collected soon after birth, and sent to state laboratories on a daily basis. State laboratories then measure the IRT value of each newborn using their DBS, and classify each newborn as test negative or test positive for the IRT test. Newborns with positive IRT test results are referred for further testing, which typically includes a series of tests, such as another IRT test, a genetic test (GT), or the diagnostic sweat chloride test (SCT); the specific testing sequence and classification rules depend on the state's NBS scheme. On the other hand, newborns with negative IRT test results are classified as CF negative, and the testing is terminated.

Two types of classification policies are currently in use by different states for the IRT test. These policies rely on a given threshold (cutoff point) to classify each newborn as test positive or test negative for the IRT test:

- Proportion-based (floating IRT threshold) policy: This policy is characterized by some percentile, x. Newborns with IRT values in the top x^{th} percentile in each testing day are classified as test positive, and all other newborns are classified as test negative.

Therefore, this policy can be equivalently described by an IRT concentration threshold that corresponds to the $(100 - x)^{th}$ percentile of the IRT values of all newborns tested in each testing day. As a result, the IRT concentration threshold (IRT threshold) potentially changes each day, i.e., it is a floating threshold. Examples of states using a proportion-based policy include North Carolina, Wisconsin, and Florida, each with a percentile of 4% [43, 72], and Texas and New York, each with a percentile of 5% [33, 72].

- Concentration-based (fixed IRT threshold) policy: This policy is characterized by some IRT concentration threshold, x (ng/mL), which remains fixed over all testing days. Newborns with IRT values greater than or equal to the given IRT concentration threshold are classified as test positive, and all other newborns are classified as test negative. Examples of states using a concentration-based policy include Georgia, with a threshold of 55 ng/mL [72], Colorado, with a threshold of 60 ng/mL [72], California, with a threshold of 62 ng/mL [42], and Washington, with a threshold of 100 ng/mL [72].

A number of studies show that IRT values can fluctuate based on seasonality [20, 35, 43, 66, 72]; more specifically, IRT values are generally higher in the winter than the summer. Thus, these studies indicate that proportion-based policies can improve classification accuracy in CF newborn screening in the presence of seasonality. Because seasonality theoretically affects the IRT values of all newborns in a similar manner, classifying a fixed percent of newborns as test positive each day (thus using a floating daily IRT threshold) is expected to reduce classification errors. This is in contrast with concentration-based policies, which use a fixed IRT threshold each day.

In general, both proportion-based and concentration-based policies have certain limitations. By relying on a fixed IRT threshold, a concentration-based policy cannot take into account the impact of external factors, such as seasonality, on IRT values. On the other hand, by using a floating IRT threshold, a proportion-based policy considers seasonality to

some extent [20, 35], but its performance depends on the size of the daily testing population, and hence, is subject to sample size error, i.e., for days with small testing populations, a proportion-based policy may lead to high classification errors.

There are no guidelines on which IRT classification policy should be selected (i.e., proportion-based or concentration-based), or what x^{th} percentile or IRT concentration threshold should be used. In this study, we analyze and quantify the effect of seasonality on the IRT values of newborns, based on a five-year CF newborn screening data set from North Carolina, and discuss the key limitations of concentration-based and proportion-based policies. Based on this analysis, we propose an adjusted IRT policy, which takes into account the effect of seasonality, but does not suffer from the limitations of current policies.

3.2 Method

We use a data set from the North Carolina State Laboratory of Public Health (NCSLPH), which consists of 646,782 newborns screened in North Carolina as part of their NBS program, over a five-year period (February 1, 2013 to February 1, 2018), corresponding to 1,357 testing days. The data set includes, for each newborn, the demographics information, including the gender, race, and birth weight; details about the IRT test, including the test date, IRT value, and classification outcome; and the classification outcomes for the GT and SCT, if applicable.

On a daily basis, the NCSLPH receives DBS of newborns, measures their IRT values (in ng/mL), and classifies each newborn using a proportion-based policy. During the study period, the NCSLPH's proportion-based policy used a percentile of 5% from February 1, 2013 to February 1, 2015, and a percentile of 4% from February 1, 2015 to February 1, 2018. Among the 646,782 newborns in the data set, 29,783 (4.6%) were classified as IRT test positive and sent to the GT. Newborns with at least one CF-causing mutation detected in GT were referred to the SCT; and the SCT identified 126 CF cases (0.02%) (with minimum

and maximum IRT values of 43.4 ng/mL and 502 ng/mL, respectively), who are referred for treatment. The SCT also identified 1,428 CF carriers (0.22%) (i.e., newborns with only one CF-causing mutation), who are asymptomatic and hence do not require treatment. The size of the daily testing population (sample size) varied between 5 and 1,311; see Figure 3.1 for a histogram of the sample sizes for 1,357 screening days.

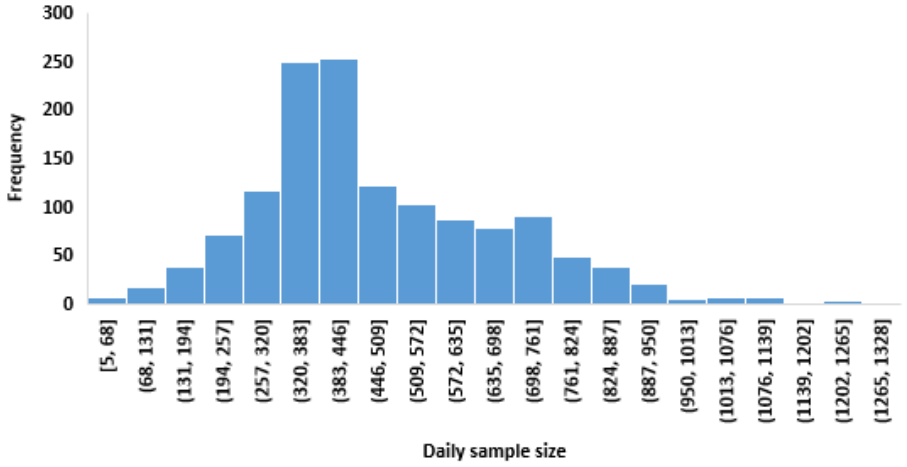

Figure 3.1: Histogram of daily sample sizes

Following Kloosterboer *et al.* [43], we divide each year into four periods: winter (December, January, and February), spring (March, April, and May), summer (June, July, and August), and fall (September, October, and November). Table 3.1 provides a summary of the CF NBS results by season. For each day, we compute the IRT concentration threshold for the 4% proportion-based policy, which corresponds to the 96^{th} percentile of IRT values for the day.

Table 3.1: Summary of CF NBS results by season for the North Carolina data set

	Winter	*Spring*	*Summer*	*Fall*	*Overall*
Number (%) of newborns	157,717 (24.39%)	156,581 (24.21%)	168,662 (26.08%)	163,709 (25.32%)	646,782
Number (%) of identified CF cases	36 (0.023%)	22 (0.014%)	24 (0.014%)	44 (0.027)	126 (0.020%)
Number (%) of IRT test positives	7,390 (4.72%)	6,454 (4.12%)	7,859 (4.67%)	8,080 (4.94%)	29,783 (4.60%)

3.3 Results

Proportion-based policies are used due to their ability to adjust the daily IRT threshold, in response to fluctuations in IRT values caused by common, external factors, such as seasonality. In this section, we first quantify the effect of seasonality and discuss some key properties of daily IRT threshold values in proportion-based policies, followed by a discussion of the limitations of proportion-based policies.

Table 3.2 reports the seasonal mean IRT and the seasonal mean 96^{th} percentile for each of the four seasons over the five-year study period. Specifically, the summer mean IRT is 23.9 ± 0.04 (mean $\pm$ standard deviation around the mean (SD)) versus 24.9 ± 0.04 for the winter; and the summer mean 96^{th} percentile is 50.5 ± 0.20 versus 52.8 ± 0.27 for the winter. These numbers support that there is a seasonality effect on the IRT values; in particular, both the mean IRT and the mean 96^{th} percentile are higher in the winter compared to the summer, and the differences are statistically significant. Further, the percent of identified CF cases is higher in the fall and winter, compared to spring and summer (0.027% and 0.023%, versus 0.014%, see Table 3.1), due in part to the seasonality effect. Kloosterboer *et al.* [43] reports a mean 95^{th} percentile of 56.8 ± 0.02 in the summer, versus 60.8 ± 0.02 in the winter in Wisconsin over a ten-year period. In our data set, the mean 95^{th} percentile is 47.6 ± 0.19 in the summer, versus 49.9 ± 0.22 in the winter. Thus, while the mean IRT values in the summer and winter are somewhat lower in our data set than those reported in Kloosterboer *et al.* [43], we also find significant seasonality effect in North Carolina, much like Wisconsin.

The seasonality effect is also evident in Figure 3.2, which plots the daily IRT mean and the daily 96^{th} percentile for each day of the study period, as well as the overall IRT mean and the overall 96^{th} percentile over the entire study period of 1,357 days. In general, the variability and range of the daily 96^{th} percentile are higher than those for the daily IRT mean, because the daily 96^{th} percentile is more sensitive to small sample sizes (Figure 3.2 and Table 3.2). Next we discuss the implications of these properties on the effectiveness of

proportion-based policies.

Table 3.2: The seasonal mean IRT and mean 96^{th} percentile for the North Carolina data set

	Winter	*Spring*	*Summer*	*Fall*	*Overall*
Year	Mean IRT ± SD (min, max)				
2013	NA	26.4 ± 0.08 (21.7, 27.6)	23.7 ± 0.09 (17.6, 30.1)	23.7 ± 0.08 (20.6, 26.5)	24.3 ± 0.08 (17.6, 30.1)
2014	24.2 ± 0.08 (21.8, 29.1)	23.2 ± 0.08 (20.3, 26.5)	23.9 ± 0.08 (21.4, 26.4)	25.1 ± 0.08 (22.6, 27.9)	24.2 ± 0.08 (20.3, 29.1)
2015	25.8 ± 0.10 (22.3, 34.8)	25.1 ± 0.08 (23.3, 27.6)	24.8 ± 0.09 (22.4, 29.5)	25.2 ± 0.08 (23.1, 27.1)	25.2 ± 0.08 (22.3, 34.8)
2016	25.7 ± 0.08 (12.4, 29.2)	26.2 ± 0.08 (22.5, 30.3)	23.8 ± 0.08 (21.4, 26.3)	24.5 ± 0.08 (22.2, 28.4)	24.9 ± 0.09 (12.4, 30.3)
2017	23.7 ± 0.08 (21.4, 26.3)	23.6 ± 0.08 (20.5, 28.7)	23.6 ± 0.08 (21.5, 26.8)	24.8 ± 0.08 (21.8, 27.3)	24.1 ± 0.07 (20.5, 28.7)
2018	24.4 ± 0.08 (21.4, 27.5)	NA	NA	NA	24.4 ± 0.08 (21.4 - 27.5)
Overall	24.9 ± 0.04 (12.4, 34.8)	24.6 ± 0.04 (20.3, 30.3)	23.9 ± 0.04 (17.6, 30.1)	24.7 ± 0.04 (20.6, 28.4)	24.5 ± 0.04 (12.4, 34.8)
Year	Mean 96^{th} percentile ± SD (min, max)				
2013	NA	54.5 ± 0.40 (45.4, 63.9)	50.0 ± 0.50 (40.9, 65.0)	49.0 ± 0.40 (41.8, 56.4)	51.6 ± 0.40 (40.9, 65.0)
2014	51.4 ± 0.51 (44.3, 59.2)	48.9 ± 0.40 (42.7, 56.9)	50.3 ± 0.41 (43.8, 61.6)	52.9 ± 0.40 (44.5, 62.9)	51.1 ± 0.40 (42.7, 62.9)
2015	55.2 ± 0.47 (47.7, 64.3)	52.8 ± 0.40 (46.5, 61.4)	51.8 ± 0.47 (43.6, 59.4)	52.3 ± 0.40 (43.7, 59.4)	52.8 ± 0.40 (43.6, 64.3)
2016	54.5 ± 0.71 (20.1, 66.4)	54.9 ± 0.40 (48.1, 67.2)	50.3 ± 0.39 (39.8, 60.1)	51.2 ± 0.40 (43.1, 63.8)	52.4 ± 0.40 (20.1, 67.2)
2017	49.4 ± 0.30 (42.3, 58.2)	50.0 ± 0.40 (41.4, 63.6)	49.9 ± 0.40 (41.7, 62.2)	52.5 ± 0.40 (43.5, 67.4)	50.7 ± 0.40 (41.4, 67.4)
2018	51.5 ± 0.40 (42.3, 67.4)	NA	NA	NA	51.5 ± 0.40 (42.3, 67.4)
Overall	52.8 ± 0.27 (20.1, 67.4)	52.1 ± 0.25 (41.4, 67.2)	50.5 ± 0.20 (39.8, 65.0)	51.6 ± 0.25 (41.8, 67.4)	51.7 ± 0.20 (20.1, 67.4)

The effectiveness of proportion-based policies, in mitigating against seasonality-based fluctuations, can diminish significantly, increasing classification errors, especially on days

with small sample sizes, because small samples may not provide a good representation of the general population. Table 3.2 indicates that the 96^{th} percentile, which represents the 4% daily IRT (concentration) threshold in the proportion-based policy, has a wide range (20.1 ng/mL to 67.4 ng/mL) and a coefficient of variation of 4.31. The maximum daily IRT threshold value (67.4 ng/mL) occurs in a day with a sample of 215, and an IRT mean of 27.05 ng/mL , while the minimum value (20.1 ng/mL) occurs in a day with a sample of only ten, and an IRT mean of 12.38 ng/mL. The daily IRT mean also fluctuates (12.38 ng/mL to 34.82 ng/mL, Table 2), however, its range is much narrower and its coefficient of variation (1.49) is much lower than those for the 96^{th} percentile, see Figure 3.2.

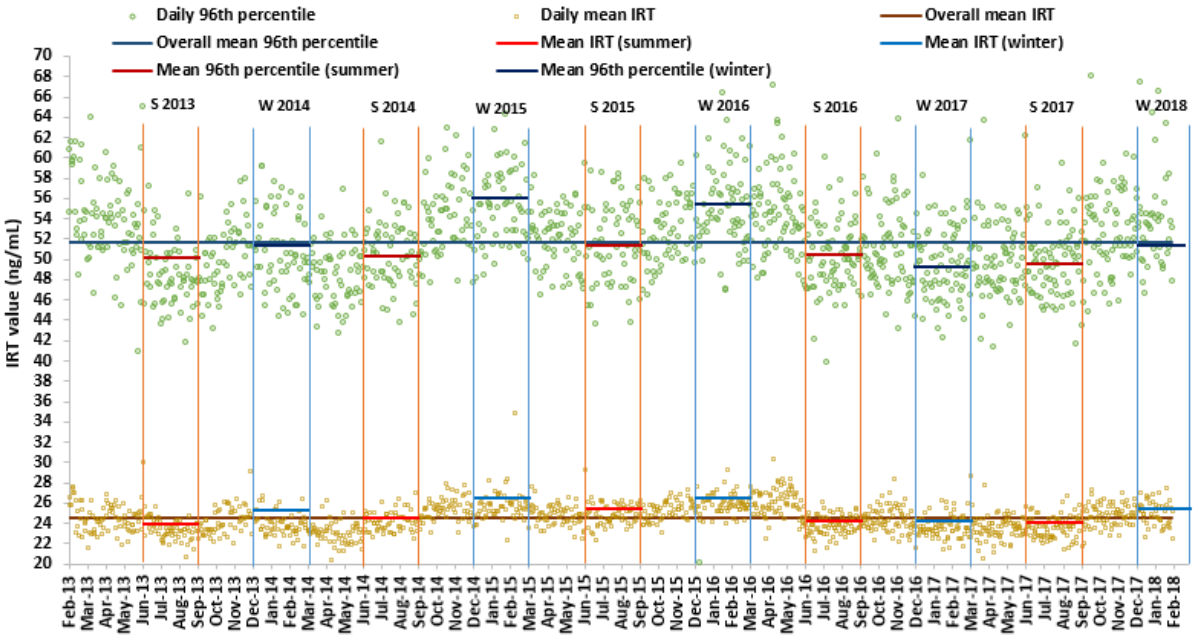

Figure 3.2: The daily mean IRT and 96^{th} percentile, the seasonal mean IRT and mean 96^{th} percentile, and the overall mean IRT and mean 96^{th} percentile for the North Carolina data set (S and W respectively denote the summer and winter seasons)

Next, we discuss how small samples may lead to an increase in the likelihood of a false negative outcome in proportion-based policies. Figure 3.3 depicts the histogram of the IRT values of all 126 newborns identified as CF positive in the data set. As discussed above, the maximum daily 96^{th} percentile in the data set is 67.4 ng/mL (Table 3.2), and out of the 126 CF positive newborns, ten cases have IRT values less than 67.4 ng/mL (Figure

3.3). Thus, these ten CF positive cases would have been missed, had they been tested on the particular day with an IRT threshold of 67.4 ng/mL. (In fact, one of these CF cases, i.e., with an IRT value of 43.4 ng/mL, was detected under the NCSLPH's old proportion-based policy of 5%, and would have been missed under a 4% proportion-based policy.) To further analyze this aspect, Table 3.3 reports some important characteristics of these ten CF positive newborns, including the likelihood of a false negative outcome, which represents the proportion of testing days with a daily IRT threshold value that is greater than the IRT value of the CF positive newborn, i.e., the proportion of testing days in which the particular CF positive newborn would have been classified as test negative.

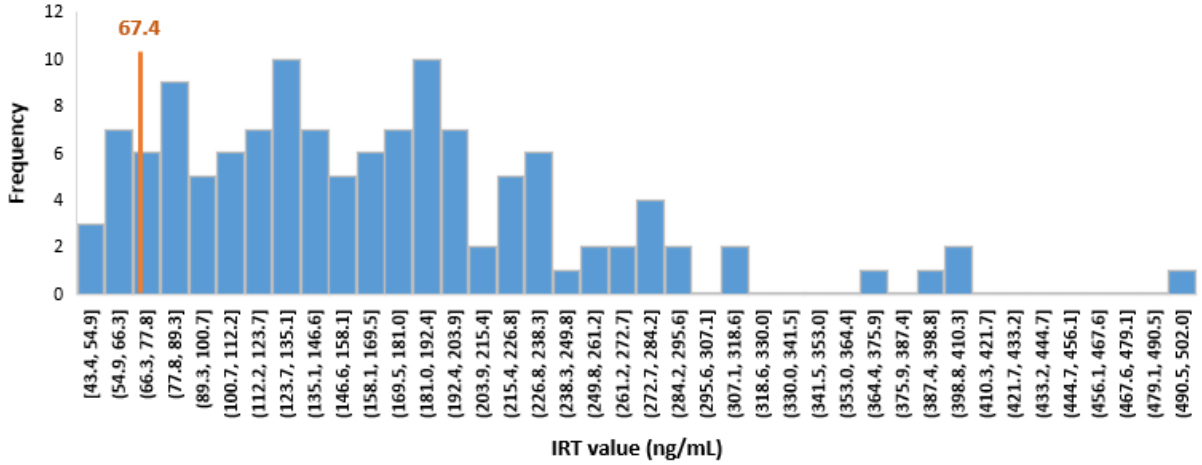

Figure 3.3: Histogram of IRT values for all identified CF positive newborns in the North Carolina data set

Table 3.3: Characteristics of CF positive newborns with IRT values less than 67.4 ng/mL, which is the maximum daily 96^{th} percentile, in the North Carolina's data set

IRT (ng/mL)	*Season* (year)	*Sample size*	*Daily IRT mean (ng/mL)*	*Daily* 96^{th} *percentile (ng/mL)*	*Likelihood of a false negative outcome (%)*
57.1	Fall (2013)	584	24.27	50.5	145 (10.68%)
43.4	Fall (2013)	198	25.25	46	1,342 (98.82%)
59.6	Winter (2013)	1,311	24.23	51.1	59 (4.35%)
51.8	Spring (2014)	296	22.45	44.3	636 (46.87%)
64.1	Summer (2014)	687	24.34	53.5	10 (0.74%)
57.8	Summer (2015)	545	25.73	55.2	112 (8.25%)
64.9	Winter (2016)	401	25.27	55.3	7 (0.52%)
63.2	Spring (2016)	543	26.5	55.3	17 (1.25%)
50.8	Spring (2017)	700	22.82	50.3	775 (57.11%)
61.6	Winter (2018)	423	23.42	49.0	28 (2.06%)

Although nine of the ten CF positive newborns with IRT values less than 67.4 ng/mL would have been identified by the 4% proportion-based policy (except for the subject with an IRT value of 43.4 ng/mL, which was identified only under the 5% policy), the likelihood of a false negative outcome for each of these ten CF cases is quite high (Table 3.3). For example, the CF positive newborn with an IRT value of 51.8 ng/mL (the fourth subject in Table 3.3) was identified as IRT test positive under the 4% proportion-based policy, but there is a likelihood of around 47% that this newborn would have been missed, had the newborn be tested on a different testing day. For the CF positive newborn with IRT value of 43.3 ng/mL (the second subject in Table 3.3), this likelihood is around 99%. In both of these cases, the sample sizes are relatively small, i.e., less than 300 (198 and 296, respectively), especially considering that only around 14% of testing days in the data set have sample sizes less than 300.

3.4 Discussion

Current IRT classification policies, namely the proportion-based and concentration-based policies, have major limitations. The concentration-based policy relies on a fixed IRT concentration threshold, and in doing so, fails to take into account the impact of external factors, such as seasonality, on IRT values. On the other hand, the proportion-based policy can consider external factors, but is subject to sample size error, and may also suffer from the fact that the daily IRT concentration threshold is more variable than the daily IRT mean.

In the following, we propose a novel policy, which we refer to as the *adjusted IRT threshold policy*, to mitigate the negative effects of small sample sizes. We do this by using a base threshold, and adjusting it up or down based on a forecast of the effect of seasonality on IRT values. In particular, we forecast the effect of seasonality using a Simple Moving Average (SMA) Method [38], where we find the average of the IRT values of the last n days, including the current day, for some given n, which we refer to as the moving IRT mean (n). Our forecast is based on the difference between the moving IRT mean (n) and the overall IRT mean, where the latter can be derived using, for example, historical data (e.g., see Figure 3.4). The adjusted threshold then follows:

Adjusted threshold = Base threshold + [Moving IRT mean (n) - Overall IRT mean].

For illustrative purposes, we use the North Carolina data set, and set the base threshold to the overall 96^{th} percentile, which equals 51.7 ng/mL, and the overall IRT mean to 24.5 ng/mL (see Figure 3.4). For demonstration, we select an n value of 20; thus the SMA is calculated by averaging the IRT values for the current day and the previous 19 testing days. This process provides us with moving IRT means that correspond to a large number of newborns, i.e., in the range of 7,467 and 12,281 newborns (i.e., number tested in 20 days).

As an example of how the adjusted IRT threshold policy works, consider a day with a moving IRT mean of 26.5 ng/mL (i.e., 2 ng/mL higher than the overall IRT mean). Then, the adjusted threshold for that day will be 53.7 ng/mL (also 2 ng/mL higher than the

base threshold), and all newborns with IRT values greater than or equal to 53.7 ng/mL will be classified as IRT test positive. (Note that in the special case without seasonality, the adjusted threshold would essentially equal the base threshold.) Figure 3.4 depicts the adjusted threshold as well as the daily IRT threshold (i.e., the 96^{th} percentile) for the 4% proportion-based policy. The proposed adjusted IRT policy leads to daily IRT thresholds in the range of 49.8 ng/mL and 55.9 ng/mL, and with a coefficient of variation of 0.94; hence, it is more stable than the current proportion-based policy, and it takes into account external factors, i.e., in days with higher IRT mean, it has a higher IRT threshold, and vice versa.

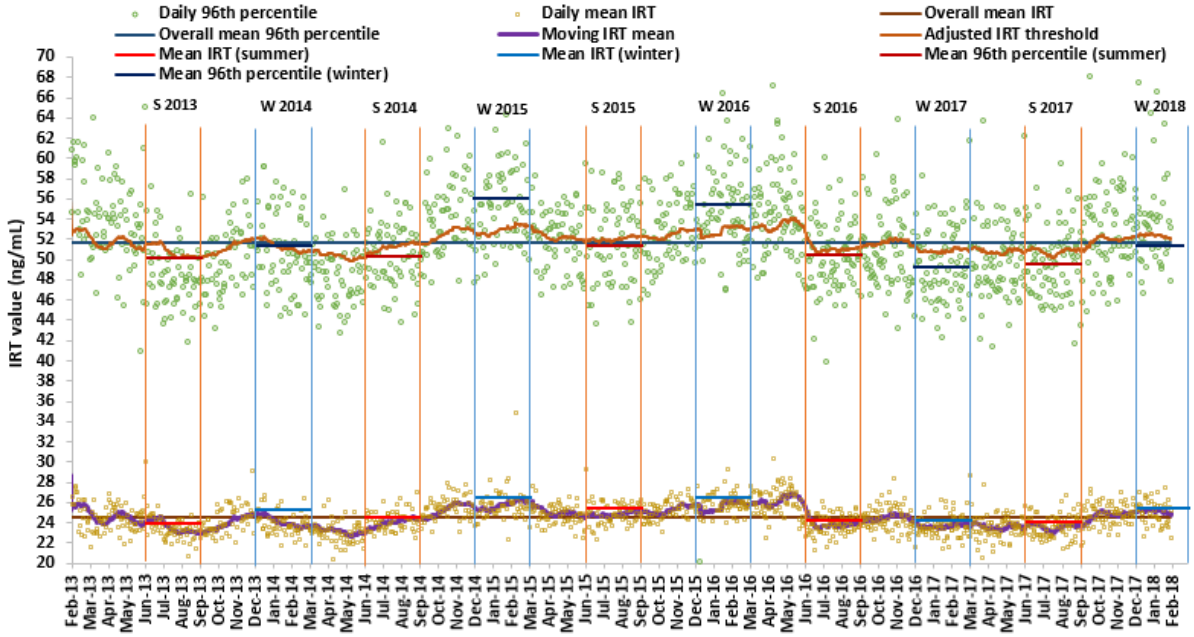

Figure 3.4: The adjusted IRT policy

3.5 Conclusion

The current IRT classification policies have certain limitations that may affect the accuracy of classification. In particular, the concentration-based policy does not take into account the impact of external factors, such as seasonality, on the IRT values of newborns. On the other hand, the proportion-based policy relies on a daily sample, and hence, is subject

to sample size error. We propose an adjusted IRT policy that takes into account external factors, while reducing the sample size error. We show that the proposed adjusted policy is substantially more reliable than North Carolina's current IRT classification policy, i.e., proportion-based policy with a 4% threshold. We hope that the proposed method provides a standard IRT classification policy that can be customized for each state, based on the effect of seasonality in the state (and other potential factors), to improve IRT classification outcomes in newborn screening for CF.

Chapter 4

A Data-driven Policy to Improve Newborn Screening for Cystic Fibrosis

4.1 Introduction

Cystic fibrosis (CF) is one of the most common life-threatening genetic diseases in the United States (US), with a prevalence rate of approximately 1 in 3,700 newborns [12, 58]. Early detection of CF can substantially improve health outcomes [19, 24, 43], while a delayed diagnosis can result in severe symptoms of the disease, including fatality [30, 46, 63]. Therefore, newborn screening for CF is conducted throughout the US [72].

In order to perform CF newborn screening, dried blood spots (DBS) from newborns are collected, and sent to state laboratories shortly after birth [33], to measure the level of immunoreactive trypsinogen (IRT) in the DBS [35]. Newborns with CF tend to have higher IRT values [18, 20, 29, 47, 66]. IRT test is very common in CF screening programs in the US, because of its low-cost. However, it results in a high number of false positive cases (newborns without CF who are classified as test positive), and hence, newborns with positive IRT test

results are referred for further tests, which could be another IRT test, a genetic test (GT), or a sweat chloride test (SCT), which is a diagnostic test, based on the state's CF screening scheme [43, 72].

There are two types of IRT classification policies used in different states of the US: *proportion-based (floating IRT threshold) policy*, and *concentration-based (fixed IRT threshold) policy*. In a proportion-based policy, there is a fixed percentile (typically 4% or 5%), and newborns with IRT values in the top percentile are classified as test positive and referred for further testing, while all other newborns are classified as test negative. In a concentration-based policy, there is a fixed IRT threshold (which varies between 55 ng/mL and 100 ng/mL among the states), and newborns with IRT values greater than or equal to the threshold are classified as test positive.

There are several external and newborn-specific factors that can affect the IRT value of newborns. External factors include seasonality (i.e., temperature and humidity), and testing kit calibration [20, 35, 43, 66, 72], while newborn-specific factors include newborns' birth weight, race, and gender [18, 35, 42, 43, 72]. For example, it has been shown that cold weather can increase the IRT level of newborns [43, 72]. These sources of variability make IRT values unreliable, and increase the likelihood of misclassification (i.e., false positive and false negative results).

Current IRT classification policies, both proportion-based and concentration-based, fail to take into account these factors that can influence the IRT value, in a rigorous manner. Concentration-based policies, by relying on a fixed IRT threshold, cannot take into account any external or newborn-specific factor that may affect the IRT value. Proportion-based policies, by relying on a daily percentile instead of a fixed IRT threshold, can heuristically take into account the external factors [20, 35, 43, 66], but do not take into account newborn-specific factors. Moreover, their performance depends on the daily sample, and hence, is subject to sample size error, i.e., for days with small sample sizes, proportion-based policies

may lead to high classification errors.

There are no guidelines on how an IRT classification policy should be selected (e.g., proportion-based or concentration-based), and what IRT threshold value to use. For example, North Carolina, Wisconsin, and Florida use a proportion-based policy with a 4% threshold [43, 72], while Texas and New York use a proportion-based policy with a 5% threshold [33, 72]. For concentration-based policies, there is also a wide range of IRT thresholds, from 55 ng/mL to 100 ng/mL. For example, Georgia uses 55 ng/mL [72], Colorado uses 60 ng/mL [72], California uses 62 ng/mL [42], and Washington uses 100 ng/mL as the IRT threshold [72].

Another aspect that should be considered in designing a CF newborn screening scheme, is to ensure equity for newborns from all racial groups, i.e., the probability of a false negative and a false positive should be approximately the same for all races. Race plays an important role in both the IRT value of newborns (e.g., African American newborns have a higher IRT mean, in comparison to other races) and their probability of having CF (i.e., CF prevalence rate substantially differs among different racial groups), and this needs to be considered in designing an effective and equitable CF screening scheme.

In this study, we analyze North Carolina's CF newborn screening scheme, with a focus on the IRT test, identify the factors that significantly affect the IRT value, propose a data-driven policy to increase the accuracy and equity of the IRT test, and show that such an approach offers an improved performance, both in terms of accuracy and equity, in comparison to North Carolina's current IRT classification policy.

4.2 Method

We use a data set from the North Carolina State Laboratory of Public Health (NCSLPH), which consists of 646,782 newborns screened in North Carolina as part of their newborn screening program, over a five-year period (February 1, 2013 to February 1, 2018),

corresponding to 1,357 testing days. The data set includes, for each newborn, the demographics information, including the gender, race, and birth weight; details about the IRT test, including the test date, IRT value, and classification outcome; and the classification outcomes for the GT and SCT, if applicable. Table 4.1 shows the demographic characteristics for newborns screened by the NCSLPH during the five-year period.

Table 4.1: Demographic characteristics for newborns screened between February 1, 2013 and February 1, 2018 in North Carolina

Characteristic	*Number (%) of newborns*
Gender listed	641,319 (99%)
Male	329,652 (51.4%)
Female	311,667 (48.6%)
Birth weight listed	646,604 (99.8%)
$<$ 2,500 gr	77,378 (11.9%)
$\geq$ 2,500 gr	569,226 (88.1%)
Race/ethnicity listed	621,766 (96.0%)
Caucasian	363,539 (56.1%)
African American	148,491 (22.9%)
Hispanic	72,388 (11.2%)
Asian	18,650 (2.9%)
Multi-racial	11,090 (1.7%)
American Indian	6,638 (1.0%)
Native HI/Pac Isl	970 (0.1%)

During the study period, the NCSLPH's proportion-based policy used a percentile of 5% from February 1, 2013 to February 1, 2015, and a percentile of 4% from February 1, 2015 to February 1, 2018. Among the 646,782 newborns in the data set, 29,783 (4.6%) were classified as IRT test positive and sent to the GT. Newborns with at least one CF-causing mutation detected in GT were referred to the SCT; and the SCT identified 126 CF cases (0.02%) (with minimum and maximum IRT values of 43.4 ng/mL and 502 ng/mL, respectively), who are referred for treatment. The SCT also identified 1,428 CF carriers (0.22%) (i.e., newborns with only one CF-causing mutation), who are asymptomatic and hence do not

require treatment.

As mentioned in Section 4.1, there are some external and newborn-specific factors that affect the IRT value of newborns. In the following, we list these factors and their effect on the IRT value.

Seasonality

Cold weather can substantially increase the IRT value of newborns [20, 35, 43, 66, 72]. We show that in the five-year period, both the mean IRT and the mean 96^{th} percentile are significantly higher in winter in comparison to summer, with respective values of 24.9 ± 0.04 and 52.8 ± 0.27 in winter, and 23.9 ± 0.04 and 50.5 ± 0.20 in summer (see Chapter 3 for more details).

Birth weight

Birth weight can affect the IRT value, with lower weight newborns having higher IRT values [43, 72]. In the North Carolina's five-year data set, the mean IRT for normal weight newborns (i.e., newborns with birth weight greater than or equal to 2,500 gr) is 24.3 ± 0.02, while for low weight newborns (i.e., newborns with birth weight less than 2,500 gr), it is 25.9 ± 0.07. Our analysis shows that the mean IRT for low weight newborns is significantly higher than those with normal birth weight.

Gender

Gender can affect the IRT value of newborns, with females having significantly higher IRT values than males. In the North Carolina's five-year data set, the mean IRT for female newborns is 25.0 ± 0.03, while for male newborns, it is 24.0 ± 0.03. Our analysis shows that the mean IRT is significantly higher in females. Therefore, although in the data set, the total number of males is higher than females in the IRT test (329,652 males in comparison to 311,667 females), more females are sent to the GT, due to elevated IRT values (15,286 females in comparison to 14,497 males).

Race

Race plays an important role in the performance of the IRT test in CF newborn screening in the US. It has a two-fold effect: (i) The IRT value can differ significantly among racial groups, e.g., African Americans have significantly higher IRT values (see Table 4.2), in comparison to other races; and (ii) The CF prevalence rate can differ significantly among racial groups, e.g., African Americans have significantly lower CF prevalence rate, in comparison to some other races, e.g., Caucasians.

Table 4.2 presents a summary of CF screening results for each racial group in the North Carolina data set. The Table also displays the mean IRT for each racial group. We show that except for the difference between Hispanics and Asians, the differences between the mean IRT for the racial groups in Table 4.2 are significant. Where applicable, a two-tailed t-test with an α level of 0.05 is used to determine statistical significance.

Table 4.2: Summary of CF screening results for different racial groups

Race	*Proportion*	*# Screened*	*Mean IRT±SD*[a]	*# GT*[b]	*# No mutations*[c]	*# SCT*[d]	*# CF*[e]	*CF prevalence rate*
Caucasian	56.22%	363,539	22.90 ± 0.02	10,702	9,488	1,982	98	1:3,710
African American	22.96%	148,491	29.24 ± 0.04	13,564	12,879	910	7	1:21,213
Hispanic	11.19%	72,388	22.41 ± 0.05	2,175	2,044	279	6	1:12,064
Asian	2.88%	18,650	21.91 ± 0.09	570	539	87	0	0
Multi-racial	1.71%	11,090	24.27 ± 0.13	472	434	65	2	1:5,545
American Indian	1.03%	6,638	25.12 ± 0.24	279	262	36	2	1:3,319
Native HI/Pacific Isl	0.15%	970	20.25 ± 0.91	21	2	5	0	0
Others	3.86%	24,903	24.54 ± 0.10	2,000	1,509	0	0	0
Overall	100%	646,669	24.49 ± 0.02	29,783	27,157	3,364	126	1:5,133

[a] SD denotes the standard deviation around the mean
[b] The number of newborns who are sent to the genetic test
[c] The number of newborns with no CF-causing mutations identified in the genetic test
[d] The number of newborns who are sent to the sweat chloride test
[e] The number of newborns who are diagnosed with CF

To illustrate some important differences between racial groups, consider the two largest groups in the data set, i.e., Caucasians and African Americans. African American newborns, with a mean IRT of 29.24±0.04, have the highest mean IRT among the races, while Caucasian newborns have a mean IRT of 22.90 ± 0.02, which is significantly lower. Figure 4.1 plots a

histogram of IRT values (with percentage instead of absolute values) for both Caucasians and African Americans. It can be observed that African Americans tend to have elevated IRT values.

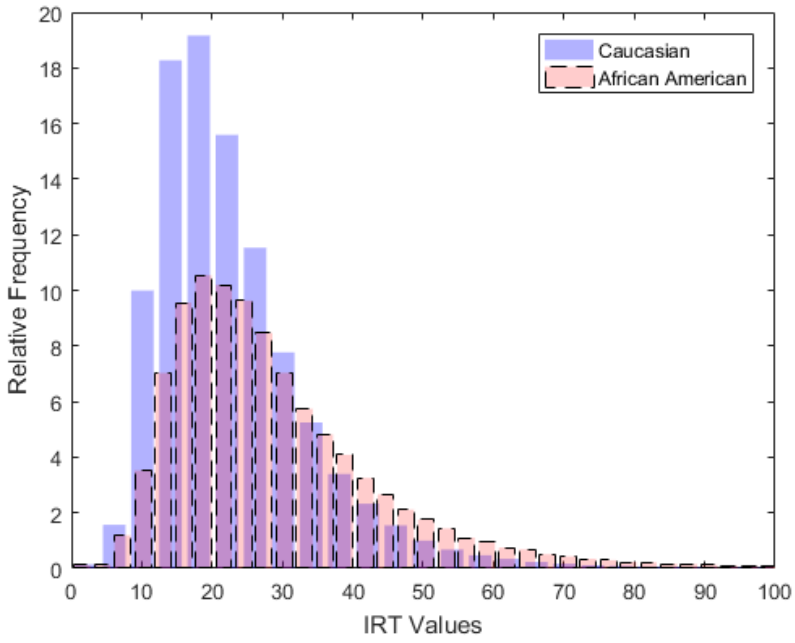

Figure 4.1: IRT relative frequency (%) in Caucasians and African Americans

Additionally, the difference in CF prevalence rates between African Americans and Caucasians is quite large; African Americans have a relatively low CF prevalence rate of 1 in 21,213 newborns, while Caucasians have a much higher CF prevalence rate of 1 in 3,710 newborns. Hence, while African Americans tend to have elevated IRT values, they have a much lower likelihood of having CF. Hence, not considering race as a factor in CF screening schemes (which is the case in current CF screening schemes), can lead to many false positive results for African American newborns, and decreases the likelihood of higher-risk groups, e.g., Caucasians to be sent for further tests.

Based on North Carolina's current IRT classification policy, i.e., a proportion-based policy with a 4% threshold, while African American newborns represent 23.96% of newborns in the data set, they represent 48.9% of the newborns who are sent to the GT. Conversely,

Caucasian newborns represent 56.22% of the newborns in the data set, yet they represent only 38.5% of the newborns sent to the GT.

4.3 Results

Table 4.3 shows the performance of various current IRT classification policies, in terms of the number of false negative and false positive cases, sensitivity and specificity. We note that we do not have reliable data on the false negative cases of North Carolina's five-year data set, and hence, our analysis is based on the identified CF cases through North Carolina's IRT classification policy.

Table 4.3: Performance of various IRT classification policies

Policy	*FN*[a]	*FP*[b]	*Sensitivity*	*Specificity*
proportion-based 5%	0	32,978	100%	94.90%
proportion-based 4%	1	26,519	99.21%	95.90%
Concentration-based 50 (ng/mL)	1	29,672	99.21%	95.41%
Concentration-based 55 (ng/mL)	3	20,433	97.62%	96.84%
Concentration-based 60 (ng/mL)	6	14,367	95.24%	97.78%
Concentration-based 62 (ng/mL)	7	12,558	94.44%	97.06%
Concentration-based 100 (ng/mL)	30	1,588	76.19%	99.75%

[a] Number of false negative cases
[b] Number of false positive cases

As mentioned in Section 4.1, the proportion-based and the concentration-based policies do not take into account external and newborn-specific factors. In the following, we propose an *adjusted IRT classification policy* that takes into account all factors that affect the IRT values. In this policy, we use a two-step regression analysis to compute the CF risk for each newborn (based on newborns' specific attributes, such as birth weight, gender, and race). Then, instead of sending 4% of newborns with the highest IRT values (which is the North Carolina's current IRT classification policy), 4% of newborns with the highest **CF**

risk are sent to the GT. We show that the adjusted 4% policy has both higher sensitivity and specificity, and can substantially increase the equity in the CF newborn screening.

We split the data set into two disjoint subsets, the training data set (60% of the data set) and the validation data set (40% of the data set). Further, we use the earliest dates in the data set for validation, as this was performed with a 5% proportion-based policy, and allows us to see the ramifications of the reduced threshold, i.e., 4%. We perform a two-step regression analysis, consisting of a linear and a logistic regression, on the training data set. We find that birth weight, gender, and race can significantly affect IRT values, and there are correlations between race and birth weight, and between gender and birth weight, and add these correlations to our regression model. Throughout, W represents newborns' birth weight, G represents the gender (it obtains value of one if the newborn is a female, and zero otherwise), X_{AF}, X_H, X_A, X_M, X_{AI}, X_N are binary variables, attaining value of one if newborn's race is respectively African American, Hispanic, Asian, Multi-racial, American Indian, or Native HI/Pac Isl. Moreover, D represents the CF status of newborns, with $D = 1$ meaning that the newborn has CF, and $D = 0$ otherwise.

We perform the backward stepwise variable selection to select the "best" subset of variables. For the linear regression model, weight, races, gender, and weight-race correlation for African Americans, Caucasians, and Hispanics are selected. Then, we perform a five-fold cross validation with stratified sampling to tune the parameters of our model. The relationship between the expected IRT value, denoted by Y, and the factors that significantly affect this value is as follows (Equation (4.1)):

$$\begin{aligned}\widehat{y} = E(Y|w, g, r_{AF}, r_H, r_A, r_{AI}, r_M, r_N) &= 25.99 - 9.694 \times 10^{-4}\, w \\ &+ 0.8109\, g + 4.426\, r_{AF} - 2.062\, r_H \\ &- 1.388\, r_A + 1.130\, r_M + 1.804\, r_{AI} - 2.151\, r_N \\ &+ 4.525 \times 10^{-4}\, w\, r_{AF} + 3.845 \times 10^{-4}\, w\, r_H, \qquad j \in \Omega,\end{aligned} \tag{4.1}$$

For the logistic regression, we consider the difference between the measured IRT value, denoted by $\widetilde{y}$, and the expected IRT value computed by Equation (4.1), $\widehat{y}$, as one factor that can be used to estimate the CF risk, and perform the backward stepwise variable selection to select the subset of variables. Caucasian, African American, Hispanic, and American Indian races are selected, along with the IRT difference, $\widetilde{Y} - \widehat{Y}$. The relationship between the CF risk, and the factors that significantly affect this value is as follows (Equation (4.2)):

$$E(D|\widetilde{y} - \widehat{y}) = \frac{1}{1 + e^{(8.826 - 0.0154\,(\widetilde{y} - \widehat{y}) + 0.344\,r_{AF} + 0.117\,r_H + 14.09\,r_A - 0.1185\,r_I)}}, \quad j \in \Omega, \tag{4.2}$$

with a p-value less than 2.2×10^{-16} for both the linear and the logistic regression models.

Table 4.4 compares the performance of the adjusted 4% IRT classification policy with the current 4% policy for the validation data set. The adjusted 4% IRT classification policy improves the efficiency of the IRT test, by referring more newborns from high risk populations (e.g., Caucasians), and less newborns from low risk populations (e.g., African-American) to the GT.

Table 4.4: Performance comparison of the adjusted 4% IRT classification policy with the current 4% IRT classification policy used in North Carolina, for the validation data set

	Adjusted 4% policy		Current 4% policy	
Race	*Sent to the GT (%)*[a]	*False negative*	*Sent to the GT (%)*	*False negative*
Caucasian	8,029 (5.19%)	0	4,428 (2.86%)	1
African American	1,219 (2.02%)	0	5,036 (8.34%)	0
Hispanic	793 (2.65%)	0	804 (2.69%)	0
Asian	0 (0%)	0	185 (2.62%)	0
Multi-racial	227 (6.26%)	0	172 (4.74%)	0
American Indian	268 (10.25%)	0	71 (2.72%)	0
Native HI/Pac Isl	8 (2.59%)	0	9 (2.91%)	0
Overall	10,544	0	10,705	1

[a] Number of newborns sent to the GT, divided by the total number of newborns in each racial group

It can be observed that the adjusted 4% policy substantially decreases the number of

African American newborns (by 76%), and substantially increases the number of Caucasian newborns (by 81%) who are sent to the GT. This improves the classification, since African American population has a very low CF prevalence rate (1 in 21,213), in comparison to Caucasian population (1 in 3,710). Moreover, the current 4% policy results in a false negative case in the validation data set (this CF positive case was found under North Carolina's previous IRT classification policy, i.e., 5% policy). We observe that this false negative case is a male Caucasian with a birth weight of 4,140 gr, and measured IRT value of 43.4 ng/mL. This newborn was screened in a day with a floating IRT threshold of 46 ng/mL (under the 4% proportion-based policy). The IRT value of this newborn, i.e., 43.4 ng/mL is relatively high, but not for a newborn with CF (the mean IRT value for identified CF positive cases is 164 ng/mL in the data set). However, considering the newborn's specific attributes (male, Caucasian, relatively high birth weight), this IRT value is considered high and hence, can be identified by the proposed adjusted 4% IRT classification policy.

Figure 4.2 shows the percentage of newborns referred to the GT from each racial group. The order of racial groups presented in the Figure is from the group with the highest CF prevalence rate, i.e., American Indian, to the group with the lowest CF prevalence rate, i.e., Native HI/Pac Isl and Asian. One can observe that the adjusted 4% policy, by taking into account newborn-specific attributes, substantially improves the classification, in terms of classifying a higher percentage of newborns from high risk groups as test positive, and sending them to the GT.

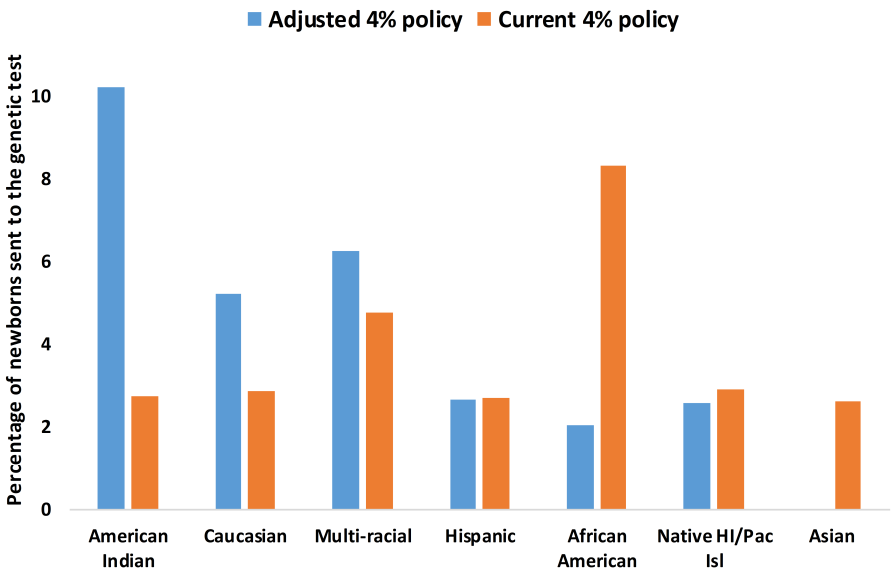

Figure 4.2: Percentage of newborns sent to the genetic test from each racial group

4.4 Discussion

Current IRT classification policies fail to take into account newborn-specific attributes, such as birth weight, race, and gender, which affect IRT values. For example, we show that, African American newborns have the highest mean IRT, while their CF prevalence rate is low (1 in 21,213). This leads to a large number of false positive cases among African American newborns, under the current IRT classification policy used in North Carolina (a proportion-based policy with a threshold of 4%). The proposed adjusted IRT classification policy, which takes into account important newborn-specific factors, modifies the population sent to the genetic test, i.e., sends more newborns with high risk.

We note that we do not have reliable data on the false negative cases of North Carolina's five-year data set, and hence, our analysis is based on the identified CF cases through North Carolina's IRT classification policy.

As mentioned in Section 4.1, some states use concentration-based IRT classification policies. The proposed adjusted IRT classification policy is not limited to proportion-based policies, and a similar adjustment can be used to improve the classification accuracy of

concentration-based IRT classification policies.

4.5 Conclusion

The proposed adjusted IRT classification policy, which is easily implementable, takes into account external and newborn-specific factors that affect the IRT values, and improves the efficiency and equity of the CF newborn screening, by referring more newborns from higher risk groups for further testings.

Chapter 5

Summary

Biomarker testing plays an important role in the diagnosis, monitoring, and management of many diseases. For many of these diseases, the actual biomarker level is unknown, due to random perturbations caused by external and/or subject-specific factors. Incorporating these factors into the design of a subject classification policy is important, and can substantially improve the accuracy of classification. We analyze the problem of determining an optimal biomarker testing and classification policy under uncertainty on the actual biomarker levels of the subjects. We formulate and study an expectation-based formulation, as well as a robust formulation that only requires an uncertainty set around the biomarker value, to minimize a function of the misclassification cost, derive key structural properties of both problems, and show that their optimal solutions follow risk-based threshold policies. We integrate the optimal risk-based policy with data analytics methodologies to estimate the disease risk for each subject. This integrated risk-based threshold policy takes into account non-disease factors that affect the biomarker level, and substantially improves the classification outcomes.

In our case study, on newborn screening for cystic fibrosis in the United States, we identify the key external and newborn-specific factors that affect the biomarker level (i.e., IRT level), based on a five-year data set from the North Carolina State Laboratory of Public Health. We implement the proposed optimization models on the data set, and show that substantial

reductions in classification errors can be achieved, over current IRT classification policies used for cystic fibrosis newborn screening. Moreover, we show that taking into account newborn-specific factors can improve both the equity of IRT classification outcomes among races (i.e., it results in approximately the same probability of false negative and false positive outcomes for different races), and the accuracy, since race plays an important role in both the biomarker levels and the prevalence rate of cystic fibrosis in newborns.

An important limitation of this work is a lack of reliable information on false negative cases in the North Carolina data set for cystic fibrosis newborn screening. Obtaining information about the false negative cases of North Carolina's IRT classification policy during the five-year study period could improve the results. Moreover, information about potential external factors that can affect IRT measurement (e.g., testing kit calibration) could also improve the analysis.

We hope that this study motivates practitioners to utilize data-driven optimization and risk prediction models for subject classification based on biomarker testing. We also hope that this work motivates researchers to investigate optimal data-driven biomarker testing designs for different purposes in healthcare, i.e., monitoring and management of diseases, and assessing the effectiveness of a treatment.

Bibliography

[1] Centers for Disease Control and Prevention. https://www.cdc.gov/, accessed May 2019.

[2] Cystic Fibrosis Foundation. https://www.cff.org/, accessed April 2018.

[3] Aissi, H., Bazgan, C., and Vanderpooten, D. (2006). Approximating min-max (regret) versions of some polynomial problems. In *International Computing and Combinatorics Conference*, pages 428–438. Springer.

[4] Atrash, S., Robinson, M. M., Aneralla, A., Brown, T., Friend, R., Sprouse, C., Ndiaye, A., Zhang, Q., Lipford, E. H., Block, J. G., et al. (2017). Validation of dynamic biomarker-based risk progression model for smoldering multiple myeloma. *Blood*, 130:1779.

[5] Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., and Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology*, 66(4):398–407.

[6] Averbakh, I. (2004). Minmax regret linear resource allocation problems. *Operations Research Letters*, 32(2):174–180.

[7] 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.

[8] Ayer, T., Alagoz, O., Stout, N. K., and Burnside, E. S. (2015). Heterogeneity in women's

adherence and its role in optimal breast cancer screening policies. *Management Science*, 62(5):1339–1362.

[9] 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.

[10] Ayvaci, M. U. S., Ahsen, M. E., Raghunathan, S., and Gharibi, Z. (2017). Timing the use of breast cancer risk information in biopsy decision-making. *Production and Operations Management*, 26(7):1333–1358.

[11] Bacus, S. and Spector, N. (2007). Biomarkers in cancer. US Patent App. 10/568,251.

[12] Baker, M. W., Atkins, A. E., Cordovado, S. K., Hendrix, M., Earley, M. C., and Farrell, P. M. (2016). Improving newborn screening for cystic fibrosis using next-generation sequencing technology: a technical feasibility study. *Genetics in Medicine*, 18(3):231.

[13] Barnett, C. L., Tomlins, S. A., Underwood, D. J., Wei, J. T., Morgan, T. M., Montie, J. E., and Denton, B. T. (2017). Two-stage biomarker protocols for improving the precision of early detection of prostate cancer. *Medical Decision Making*, 37(7):815–826.

[14] Berezin, A. E., Kremzer, A. A., Martovitskaya, Y. V., Berezina, T. A., and Samura, T. A. (2015). The utility of biomarker risk prediction score in patients with chronic heart failure. *Clinical Hypertension*, 22(1):3.

[15] Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3):464–501.

[16] Bertsimas, D., Silberholz, J., and Trikalinos, T. (2018). Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening. *Healthcare Management Science*, 21(1):105–118.

[17] Bonaccorso, G. (2017). *Machine Learning Algorithms*. Packt Publishing Ltd.

[18] Castellani, C., Massie, J., Sontag, M., and Southern, K. W. (2016). Newborn screening for cystic fibrosis. *The Lancet Respiratory Medicine*, 4(8):653–661.

[19] Chatfield, S., Owen, G., Ryley, H., Williams, J., Alfaham, M., Goodchild, M., and Weller, P. (1991). Neonatal screening for cystic fibrosis in wales and the west midlands: clinical assessment after five years of screening. *Archives of Disease in Childhood*, 66(1 Spec No):29–33.

[20] Comeau, A. M., Accurso, F. J., White, T. B., Campbell, P. W., Hoffman, G., Parad, R. B., Wilfond, B. S., Rosenfeld, M., Sontag, M. K., Massie, J., et al. (2007). Guidelines for implementation of cystic fibrosis newborn screening programs: Cystic Fibrosis Foundation workshop report. *Pediatrics*, 119(2):e495–e518.

[21] Comeau, A. M., Parad, R. B., Dorkin, H. L., Dovey, M., Gerstle, R., Haver, K., Lapey, A., O'Sullivan, B. P., Waltz, D. A., Zwerdling, R. G., et al. (2004). Population-based newborn screening for genetic disorders when multiple mutation DNA testing is incorporated: a cystic fibrosis newborn screening model demonstrating increased sensitivity but more carrier detections. *Pediatrics*, 113(6):1573–1581.

[22] Currier, R. J., Sciortino, S., Liu, R., Bishop, T., Koupaei, R. A., and Feuchtbaum, L. (2017). Genomic sequencing in cystic fibrosis newborn screening: what works best, two-tier predefined CFTR mutation panels or second-tier CFTR panel followed by third-tier sequencing? *Genetics in Medicine*, 19(10):1159.

[23] Deneef, P. and Kent, D. L. (1993). Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. *Medical Decision Making*, 13(2):126–132.

[24] Dijk, F. N., McKay, K., Barzi, F., Gaskin, K. J., and Fitzgerald, D. A. (2011). Improved survival in cystic fibrosis patients diagnosed by newborn screening compared to a historical cohort from the same centre. *Archives of Disease in Childhood*, 96(12):1118–1123.

[25] Dodd, R., Notari, E., and Stramer, S. (2002). Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross blood donor population. *Transfusion*, 42(8):975–979.

[26] Doecke, J. D., Laws, S. M., Faux, N. G., Wilson, W., Burnham, S. C., Lam, C., Mondal, A., Bedo, J., Bush, A. I., Brown, B., et al. (2012). Blood-based protein biomarkers for diagnosis of Alzheimer disease. *Archives of Neurology*, 69(10):1318–1325.

[27] Draper, N. R. and Smith, H. (1998). Selecting the "best" regression equation. *Applied Regression Analysis*, pages 327–368.

[28] El-Amine, H., Bish, E. K., and Bish, D. R. (2018). Robust postdonation blood screening under prevalence rate uncertainty. *Operations Research*, 66(1):1–17.

[29] Farrell, P. M., Rosenstein, B. J., White, T. B., Accurso, F. J., Castellani, C., Cutting, G. R., Durie, P. R., LeGrys, V. A., Massie, J., Parad, R. B., et al. (2008). Guidelines for diagnosis of cystic fibrosis in newborns through older adults: Cystic Fibrosis Foundation consensus report. *The Journal of Pediatrics*, 153(2):S4–S14.

[30] Farrell, P. M., Kosorok, M. R., Rock, M. J., Laxova, A., Zeng, L., Lai, H., Hoffman, G., Laessig, R. H., Splaingard, M. L., Wisconsin Cystic Fibrosis Neonatal Screening Study Group and others (2001). Early diagnosis of cystic fibrosis through neonatal screening prevents severe malnutrition and improves long-term growth. *Pediatrics*, 107(1):1–13.

[31] Felder, S. and Mayrhofer, T. (2014). Risk preferences: consequences for test and treatment thresholds and optimal cutoffs. *Medical Decision Making*, 34(1):33–41.

[32] Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal*, 47(4):458–472.

[33] Giusti, R., Badgwell, A., Iglesias, A. D., et al. (2007). New York State cystic fibrosis consortium: the first 2.5 years of experience with cystic fibrosis newborn screening in an ethnically diverse population. *Pediatrics*, 119(2):e460–e467.

[34] Greiner, M., Sohr, D., and Göbel, P. (1995). A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. *Journal of Immunological Methods*, 185(1):123–132.

[35] Hammond, K. B., Abman, S. H., Sokol, R. J., and Accurso, F. J. (1991). Efficacy of statewide neonatal screening for cystic fibrosis by assay of trypsinogen concentrations. *New England Journal of Medicine*, 325(11):769–774.

[36] Hamosh, A., FitzSimmons, S. C., Macek Jr, M., Knowles, M. R., Rosenstein, B. J., and Cutting, G. R. (1998). Comparison of the clinical manifestations of cystic fibrosis in black and white patients. *The Journal of Pediatrics*, 132(2):255–259.

[37] Higgins, T. L., Estafanous, F. G., Loop, F. D., Beck, G. J., Blum, J. M., and Paranandi, L. (1992). Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: a clinical severity score. *JAMA*, 267(17):2344–2348.

[38] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[39] Jiang, W., Freidlin, B., and Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*, 99(13):1036–1043.

[40] Jund, J., Rabilloud, M., Wallon, M., and Ecochard, R. (2005). Methods to estimate

the optimal threshold for normally or log-normally distributed biological tests. *Medical Decision Making*, 25(4):406–415.

[41] Kammesheidt, A., Kharrazi, M., Graham, S., Young, S., Pearl, M., Dunlop, C., and Keiles, S. (2006). Comprehensive genetic analysis of the cystic fibrosis transmembrane conductance regulator from dried blood specimens–implications for newborn screening. *Genetics in Medicine*, 8(9):557.

[42] Kharrazi, M., Yang, J., Bishop, T., Lessing, S., Young, S., Graham, S., Pearl, M., Chow, H., Ho, T., Currier, R., et al. (2015). Newborn screening for cystic fibrosis in California. *Pediatrics*, 136(6):1062–1072.

[43] Kloosterboer, M., Hoffman, G., Rock, M., Gershan, W., Laxova, A., Li, Z., and Farrell, P. M. (2009). Clarification of laboratory and clinical variables that influence cystic fibrosis newborn screening with initial analysis of immunoreactive trypsinogen. *Pediatrics*, 123(2):e338–e346.

[44] Kucirka, L. M., Sarathy, H., Govindan, P., Wolf, J. H., Ellison, T. A., Hart, L. J., Montgomery, R. A., Ros, R. L., and Segev, D. L. (2011). Risk of Window Period HIV Infection in High Infectious Risk Donors: Systematic Review and Meta-Analysis. *American Journal of Transplantation*, 11(6):1176–1187.

[45] Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

[46] Legrys, V. A. and Wood, R. E. (1988). Incidence and implications of false-negative sweat test reports in patients with cystic fibrosis. *Pediatric Pulmonology*, 4(3):169–172.

[47] Massie, J., Curnow, L., Tzanakos, N., Francis, I., and Robertson, C. F. (2006). Markedly elevated neonatal immunoreactive trypsinogen levels in the absence of cystic fibrosis gene mutations is not an indication for further testing. *Archives of Disease in Childhood*, 91(3):222–225.

[48] Mastin, A., Jaillet, P., and Chin, S. (2015). Randomized minmax regret for combinatorial optimization under uncertainty. In *International Symposium on Algorithms and Computation*, pages 491–501. Springer.

[49] Mayeux, R. (2004). Biomarkers: potential uses and limitations. *NeuroRx*, 1(2):182–188.

[50] McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2012). Regression models for group testing data with pool dilution effects. *Biostatistics*, 14(2):284–298.

[51] Paracchini, V., Seia, M., Raimondi, S., Costantino, L., Capasso, P., Porcaro, L., Colombo, C., Coviello, D. A., Mariani, T., Manzoni, E., et al. (2011). Cystic fibrosis newborn screening: distribution of blood immunoreactive trypsinogen concentrations in hypertrypsinemic neonates. In *JIMD Reports-Case and Research Reports, 2012/1*, pages 17–23. Springer.

[52] Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

[53] Pepe, Margaret Sullivan (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

[54] Perakis, G. and Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research*, 56(1):188–203.

[55] Pollitt, R. J. and Matthews, A. J. (2007). Population quantile-quantile plots for monitoring assay performance in newborn screening. *Journal of Inherited Metabolic Disease*, 30(4):607–607.

[56] Price, S., Golden, B., Wasil, E., and Denton, B. T. (2016). Operations research models and methods in the screening, detection, and treatment of prostate cancer: A categorized, annotated review. *Operations Research for Health Care*, 8:9–21.

[57] Rapisuwon, S., Vietsch, E. E., and Wellstein, A. (2016). Circulating biomarkers to monitor cancer progression and treatment. *Computational and Structural Biotechnology Journal*, 14:211–222.

[58] Rohlfs, E. M., Zhou, Z., Heim, R. A., Nagan, N., Rosenblum, L. S., Flynn, K., Scholl, T., Akmaev, V. R., Sirko-Osadsa, D. A., Allitto, B. A., et al. (2011). Cystic fibrosis carrier testing in an ethnically diverse us population. *Clinical Chemistry*, pages clinchem–2010.

[59] Ryckman, K. K., Berberich, S. L., Shchelochkov, O. A., Cook, D. E., and Murray, J. C. (2013). Clinical and environmental influences on metabolic biomarkers collected for newborn screening. *Clinical Biochemistry*, 46(1-2):133–138.

[60] Sato, K. K., Hayashi, T., Harita, N., Yoneda, T., Nakamura, Y., Endo, G., and Kambe, H. (2009). Combined measurement of fasting plasma glucose and hba1c is effective for the prediction of type 2 diabetes: The kansai healthcare study. *Diabetes Care*.

[61] Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67.

[62] Schisterman, E. F., Perkins, N. J., Liu, A., and Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, 16(1):73–81.

[63] Sims, E. J., Clark, A., McCormick, J., Mehta, G., Connett, G., Mehta, A., et al. (2007). Cystic fibrosis diagnosed after 2 months of age leads to worse outcomes and requires more therapy. *Pediatrics*, 119(1):19–28.

[64] Solvang, H. K., Frigessi, A., Kaveh, F., Riis, M. L., Lüders, T., Bukholm, I. R., Kristensen, V. N., and Andreassen, B. K. (2016). Gene expression analysis supports tumor threshold over 2.0 cm for T-category breast cancer. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016(1):6.

[65] Somoza, E. and Mossman, D. (1992). Comparing and optimizing diagnostic tests: an information-theoretical approach. *Medical Decision Making*, 12(3):179–188.

[66] Sontag, M. K., Hammond, K. B., Zielenski, J., Wagener, J. S., and Accurso, F. J. (2005). Two-tiered immunoreactive trypsinogen-based newborn screening for cystic fibrosis in Colorado: screening efficacy and diagnostic outcomes. *The Journal of Pediatrics*, 147(3):S83–S88.

[67] Stephen, J., Murray, G., Cameron, D., Thomas, J., Kunkler, I., Jack, W., Kerr, G., Piper, T., Brookes, C., Rea, D., et al. (2014). Time dependence of biomarkers: non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *British Journal of Cancer*, 111(12):2242.

[68] Subtil, F. and Rabilloud, M. (2010). A Bayesian method to estimate the optimal threshold of a longitudinal biomarker. *Biometrical Journal*, 52(3):333–347.

[69] Subtil, F. and Rabilloud, M. (2014). Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC Medical Informatics and Decision Making*, 14(1):53.

[70] Szefler, S. J., Wenzel, S., Brown, R., Erzurum, S. C., Fahy, J. V., Hamilton, R. G., Hunt, J. F., Kita, H., Liu, A. H., Panettieri, R. A., et al. (2012). Asthma outcomes: biomarkers. *Journal of Allergy and Clinical Immunology*, 129(3):S9–S23.

[71] Tang, W. W., Francis, G. S., Morrow, D. A., Newby, L. K., Cannon, C. P., Jesse, R. L., Storrow, A. B., Christenson, R. H., Apple, F. S., Ravkilde, J., et al. (2007). National Academy of Clinical Biochemistry Laboratory Medicine practice guidelines: clinical utilization of cardiac biomarker testing in heart failure. *Circulation*, 116(5):e99–e109.

[72] Therrell, B. L., Hannon, W. H., Hoffman, G., Ojodu, J., and Farrell, P. M. (2012). Im-

munoreactive trypsinogen (IRT) as a biomarker for cystic fibrosis: challenges in newborn dried blood spot screening. *Molecular Genetics and Metabolism*, 106(1):1–6.

[73] Tluczek, A., Mischler, E. H., Farrell, P. M., Fost, N., Peterson, N. M., Carey, P., Bruns, W. T., and McCarthy, C. (1992). Parents' knowledge of neonatal screening and response to false-positive cystic fibrosis testing. *Journal of Developmental and Behavioral Pediatrics: JDBP*, 13(3):181–186.

[74] Underwood, D. J., Zhang, J., Denton, B. T., Shah, N. D., and Inman, B. A. (2012). Simulation optimization of PSA-threshold based prostate cancer screening policies. *Healthcare Management Science*, 15(4):293–309.

[75] van Giessen, A., de Wit, G. A., Moons, K. G., Dorresteijn, J. A., and Koffijberg, H. (2017). An alternative approach identified optimal risk thresholds for treatment indication: an illustration in coronary heart disease. *Journal of Clinical Epidemiology*.

[76] Vermont, J., Bosson, J., Francois, P., Robert, C., Rueff, A., and Demongeot, J. (1991). Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35(2):141–150.

[77] Wang, D., McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2018). Group testing case identification with biomarker information. *Computational Statistics & Data Analysis*, 122:156–166.

[78] Wein, L. M. and Zenios, S. A. (1996). Pooled testing for HIV screening: capturing the dilution effect. *Operations Research*, 44(4):543–569.

[79] Wells, J., Rosenberg, M., Hoffman, G., Anstead, M., and Farrell, P. M. (2012). A decision-tree approach to cost comparison of newborn screening strategies for cystic fibrosis. *Pediatrics*, pages peds–2011.

[80] Yang, Y., Goldhaber-Fiebert, J. D., and Wein, L. M. (2013). Analyzing screening policies for childhood obesity. *Management Science*, 59(4):782–795.

[81] Ypma, T. J. (1995). Historical development of the Newton–Raphson method. *SIAM Review*, 37(4):531–551.

[82] Yu, W., Liu, T., Valdez, R., Gwinn, M., and Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1):16.

[83] Yue, J., Chen, B., and Wang, M. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, 54(6):1128–1136.

[84] Zenios, S. A. and Wein, L. M. (1998). Pooled testing for HIV prevalence estimation: exploiting the dilution effect. *Statistics in Medicine*, 17(13):1447–1467.

[85] 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.

Appendix A

Appendix for Chapter 2

A.1 Mathematical Proofs

Proof of Theorem 1:

The result simply follows because the **EM** objective function is additively separable in the $\overrightarrow{x}$ vector, and there is no constraint that links the x_j variables. Then, it follows that for each subject $j \in \Omega$, for a given $\overrightarrow{\widehat{p}}$,

$$x_j^{*E} = \begin{cases} 1, & \text{if } c_{FN}\widehat{p}_j \geq c_{FP}(1-\widehat{p}_j) \\ 0, & \text{otherwise} \end{cases}.$$

Equivalently, there exists a $p_{th}^{*E} \equiv \frac{c_{FP}}{c_{FN}+c_{FP}}$ such that:

$$x_j^{*E} = \begin{cases} 1, & \text{if } \widehat{p}_j \geq p_{th}^{*E} \\ 0, & \text{if } \widehat{p}_j < p_{th}^{*E} \end{cases}.$$

Proof of Lemma 1:

Let $x_j^*(p_j)$ denote the optimal solution to the deterministic problem with a given p_j, i.e., it is the optimal solution to **EM** with $\widehat{p}_j = p_j$. Recall that $p_{th}^{*E} = \frac{c_{FP}}{c_{FP}+c_{FN}}$.

Case 1. $\overline{p}_j < p_{th}^{*E} \Rightarrow p_j < p_{th}^{*E}$, $\forall p_j \in [\underline{p}_j, \overline{p}_j]$:

In this case, $x_j^*(p) = 0$, $\forall p \in [\underline{p}_j, \overline{p}_j]$, leading to:

$$Regret(x_j, p_j) = \begin{cases} 0, & \text{if } x_j = 0 \\ c_{FP}(1 - p_j) - c_{FN}p_j = c_{FP} - p_j(c_{FP} + c_{FN}), & \text{if } x_j = 1 \end{cases}.$$

$\Rightarrow max_{p_j \in [\underline{p}_j, \overline{p}_j]} Regret(0, p_j) = 0$

$\Rightarrow max_{p_j \in [\underline{p}_j, \overline{p}_j]} Regret(1, p_j) = Regret(1, \underline{p}_j) = c_{FP} - \underline{p}_j(c_{FP} + c_{FN})$

Case 2. $\underline{p}_j \geq p_{th}^{*E} \Rightarrow p_j \geq p_{th}^{*E}$, $\forall p_j \in [\underline{p}_j, \overline{p}_j]$:

In this case, $x_j^*(p) = 1$, $\forall p \in [\underline{p}_j, \overline{p}_j]$, leading to:

$$Regret(x_j, p_j) = \begin{cases} c_{FN}p_j - c_{FP}(1 - p_j) = p_j(c_{FP} + c_{FN}) - c_{FP}, & \text{if } x_j = 0 \\ 0, & \text{if } x_j = 1 \end{cases}.$$

$\Rightarrow max_{p_j \in [\underline{p}_j, \overline{p}_j]} Regret(0, p_j) = Regret(0, \overline{p}_j) = \overline{p}_j(c_{FN} + c_{FP}) - c_{FP}$

$\Rightarrow max_{p_j \in [\underline{p}_j, \overline{p}_j]} Regret(1, p_j) = 0$

Case 3. $\underline{p}_j < p_{th}^{*E} < \overline{p}_j$.

In this case, $x_j^*(p) = 0$, for $\forall p \in [\underline{p}_j, p_{th}^{*E})$ and $x_j^*(p) = 1$, for $\forall p \in [p_{th}^{*E}, \overline{p}_j]$, leading to the following cases:

For $\forall p_j \in [\underline{p}_j, p_{th}^{*E})$:

$$Regret(x_j, p_j) = \begin{cases} 0, & \text{if } x_j = 0 \\ c_{FP} - p_j(c_{FP} + c_{FN}), & \text{if } x_j = 1 \end{cases}.$$

$\Rightarrow max_{p_j \in [\underline{p}_j, p_{th}^{*E})} Regret(0, p_j) = 0$

$\Rightarrow max_{p_j \in [\underline{p}_j, p_{th}^{*E})} Regret(1, p_j) = c_{FP} - \underline{p}_j(c_{FP} + c_{FN})$

Similarly, for $\forall p_j \in [p_{th}^{*E}, \overline{p}_j]$:

$$Regret(x_j, p_j) = \begin{cases} p_j(c_{FN} + c_{FP}) - c_{FP}, & \text{if } x_j = 0 \\ 0, & \text{if } x_j = 1 \end{cases}.$$

$\Rightarrow max_{p_j \in [p_{th}^{*E}, \overline{p}_j]} Regret(0, p_j) = \overline{p}_j(c_{FP} + c_{FN}) - c_{FP}$

$\Rightarrow max_{p_j \in [p_{th}^{*E}, \overline{p}_j]} Regret(1, p_j) = 0$

Proof of Theorem 2:

The result follows directly from Corollary 2, because the objective function in **RM** is additively separable in the $\overrightarrow{x}$ vector, and there is no constraint that links the x_j variables. Then it follows that for each subject $j \in \Omega$,

$$x_j^{*R} = \begin{cases} 1, & \text{if } \left(c_{FP} - \underline{p}_j(c_{FP} + c_{FN})\right)^+ \leq \left(\overline{p}_j(c_{FN} + c_{FP}) - c_{FP}\right)^+ \\ 0, & \text{otherwise} \end{cases}.$$

There are 4 possible cases:

Case 1: $c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) \geq 0$ and $\overline{p}_j(c_{FN} + c_{FP}) - c_{FP} \geq 0$.

In this case, the optimal solution follows:

$$x_j^{*R} = \begin{cases} 1, & \text{if } c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) \leq \overline{p}_j(c_{FN} + c_{FP}) - c_{FP} \\ 0, & \text{otherwise} \end{cases} \Rightarrow x_j^{*R} = \begin{cases} 1, & \text{if } \frac{\underline{p}_j + \overline{p}_j}{2} \geq p_{th}^{*E} \\ 0, & \text{if } \frac{\underline{p}_j + \overline{p}_j}{2} < p_{th}^{*E} \end{cases}.$$

where $p_{th}^{*E} = \frac{c_{FP}}{c_{FP} + c_{FN}}$.

Case 2: $c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) > 0$ (equivalently $\underline{p}_j < p_{th}^{*E}$) and $\overline{p}_j(c_{FN} + c_{FP}) - c_{FP} < 0$ (equivalently $\overline{p}_j < p_{th}^{*E}$).

In this case, the optimal solution follows:

$$x_j^{*R} = \begin{cases} 1, & \text{if } c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) \leq 0 \\ 0, & \text{otherwise} \end{cases} \Rightarrow x_j^{*R} = 0.$$

This case can be considered as a special case of Case 1, since for all $\underline{p}_j < p_{th}^{*E}$ and $\overline{p}_j < p_{th}^{*E}$, we have $\frac{\underline{p}_j + \overline{p}_j}{2} < p_{th}^{*E} \Rightarrow x_j^{*R} = 0$.

Case 3: $c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) \leq 0$ and $\overline{p}_j(c_{FN} + c_{FP}) - c_{FP} \leq 0$.

In this case, the optimal solution follows:

$$x_j^{*R} = \begin{cases} 1, & \text{if } 0 \leq 0 \qquad \Rightarrow x_j^{*R} = 1. \\ 0, & \text{otherwise} \end{cases}$$

This case is only possible if both inequalities are equal to 0, which leads to $\underline{p}_j = \overline{p}_j = p_{th}^{*E}$. This also can be considered as a special case of Case 1, in which we have $\frac{\underline{p}_j + \overline{p}_j}{2} = p_{th}^{*E} \Rightarrow x_j^{*R} = 1$.

Case 4: $c_{FP} - \underline{p}_j(c_{FP} + c_{FN}) < 0$ and $\overline{p}_j(c_{FN} + c_{FP}) - c_{FP} > 0$.

In this case, the optimal solution follows:

$$x_j^{*R} = \begin{cases} 1, & \text{if } 0 \leq \overline{p}_j(c_{FN} + c_{FP}) - c_{FP} \qquad \Rightarrow x_j^{*R} = 1. \\ 0, & \text{otherwise} \end{cases}$$

This case, as well, can be considered as a special case of Case 1, since for all $\underline{p}_j > p_{th}^{*E}$ and $\overline{p}_j > p_{th}^{*E}$, we have $\frac{\underline{p}_j + \overline{p}_j}{2} > p_{th}^{*E} \Rightarrow x_j^{*R} = 1$.

Proof of Theorem 3:

For a given risk vector $\overrightarrow{p}$, we can write the following expression for $\Pi^R(\overrightarrow{p})$:

Case 1. $p_j \geq p_{th}^{*E}$, $\frac{\underline{p}_j + \overline{p}_j}{2} < p_{th}^{*E}$.

In this case, $x_j^* = 1$ and $x_j^{*R} = 0$, leading to:

$$\Pi^R(p_j) = p_j c_{FN} - c_{FP}(1 - p_j) = p_j(c_{FN} + c_{FP}) - c_{FP}.$$

Case 2. $p_j < p_{th}^{*E}$, $\frac{\underline{p}_j + \overline{p}_j}{2} \geq p_{th}^{*E}$.

In this case, $x_j^* = 0$ and $x_j^{*R} = 1$, leading to:

$$\Pi^R(p_j) = c_{FP}(1 - p_j) - p_j c_{FN} = c_{FP} - p_j(c_{FN} + c_{FP}).$$

Case 3. $p_j < p_{th}^{*E}$, $\frac{\underline{p}_j + \overline{p}_j}{2} < p_{th}^{*E}$.

In this case, $x_j^* = 0$ and $x_j^{*R} = 0$, leading to $\Pi^R(p_j) = 0$.

Case 4. $p_j \geq p_{th}^{*E}$, $\frac{\underline{p}_j + \overline{p}_j}{2} \geq p_{th}^{*E}$.

In this case, $x_j^* = 1$ and $x_j^{*R} = 1$, leading to $\Pi^R(p_j) = 0$.

Similarly, for a given risk vector $\overrightarrow{p}$, for the price of expectation-based, $\Pi^E(\overrightarrow{p})$, we can write:

Case 1. $p_j \geq p_{th}^{*E}$, $\widehat{p}_j < p_{th}^{*E}$.

In this case, $x_j^* = 1$ and $x_j^{*E} = 0$, leading to:

$$\Pi^E(p_j) = p_j c_{FN} - c_{FP}(1 - p_j) = p_j(c_{FN} + c_{FP}) - c_{FP}.$$

Case 2. $p_j < p_{th}^{*E}$, $\widehat{p}_j \geq p_{th}^{*E}$.

In this case, $x_j^* = 0$ and $x_j^{*E} = 1$, leading to:

$$\Pi^E(p_j) = c_{FP}(1 - p_j) - p_j c_{FN} = c_{FP} - p_j(c_{FN} + c_{FP}).$$

Case 3. $p_j < p_{th}^{*E}$, $\widehat{p}_j < p_{th}^{*E}$.

In this case, $x_j^* = 0$ and $x_j^{*E} = 0$, leading to $\Pi^E(p_j) = 0$.

Case 4. $p_j \geq p_{th}^{*E}$, $\widehat{p}_j \geq p_{th}^{*E}$.

In this case, $x_j^* = 1$ and $x_j^{*E} = 1$, leading to $\Pi^E(p_j) = 0$.

A.2 An Equivalent Formulation for Test Efficacy Maximization

We let q denote the disease prevalence rate within the population, and let $\widehat{P}$ and $\widehat{Y}$ respectively denote the estimated disease risk, and the processed biomarker level (i.e., the biomarker level derived via the $h(.)$ function, see Table 2.1) of a random subject.

Recall that x refers to the subject's classification status ($x = 1$ if test positive, and $x = 0$ if test negative), and D refers to the subject's true status ($D = 1$ if true positive, and $D = 0$ if true negative). Then, for any risk threshold $p_{th} \in [0, 1]$, test sensitivity (Se), i.e., the probability of correctly classifying a true positive subject, and test specificity (Sp), i.e., the probability of correctly classifying a true negative subject, follow:

$$Se(p_{th}) = Pr(x = 1|D = 1) = Pr(\widehat{P} \geq p_{th}|D = 1), \;\; Sp(p_{th}) = Pr(x = 0|D = 0) = Pr(\widehat{P} < p_{th}|D = 0).$$

Then, using the law of total probability, the expected cost of misclassification can be written as,

$$\begin{aligned} E\Big[C(p_{th})\Big] &= c_{FN} Pr((1 - x)D = 1) + c_{FP} Pr(x(1 - D) = 1) \\ &= c_{FN}\, Pr(x = 0|D = 1) Pr(D = 1) + c_{FP}\, Pr(x = 1|D = 0) Pr(D = 0). \end{aligned}$$

Corollary 5. *An equivalent formulation for **EM** follows:*

***Problem EM**:*

$$\begin{aligned} &minimize_{p_{th} \in [0,1]} \; E\Big[C(p_{th})\Big] = c_{FN} Pr((1 - x)D = 1) + c_{FP} Pr(x(1 - D) = 1) \\ &\equiv minimize_{p_{th} \in [0,1]} \; q\, c_{FN}\, (1 - Se(p_{th})) + (1 - q)\, c_{FP}\, (1 - Sp(p_{th})) \\ &\equiv maximize_{p_{th} \in [0,1]} \; q\, c_{FN}\, Se(p_{th}) + (1 - q)\, c_{FP}\, Sp(p_{th}) \end{aligned}$$

Thus, the **EM** objective is equivalent to maximizing a weighted sum of test sensitivity and specificity, with respective weights of $q\,c_{FN}$ and $(1-q)\,c_{FP}$. Consequently, the tester may define weights, or target levels, for sensitivity or specificity, rather than specify misclassification costs. To this end, we define $k \equiv \frac{c_{FN}}{c_{FP}}$, and express the optimal **EM** threshold as a function of k, i.e., $p_{th}^{*E}(k) = \frac{c_{FP}}{c_{FN}+c_{FP}} = \frac{1}{k+1}$ (see Theorem 1), leading to $Se(p_{th}^{*E}) = Se(k) = Pr(\widehat{P} \geq \frac{1}{k+1} | D = 1)$ and $Sp(p_{th}^{*E}) = Sp(k) = Pr(\widehat{P} < \frac{1}{k+1} | D = 0)$. Thus, as k increases, the sensitivity increases, while the specificity decreases, and the value of parameter k can be set considering this trade-off.

Remark 6. *Let $\widehat{Y}_+$ and $\widehat{Y}_-$ respectively denote the processed biomarker level of a random true positive and a random true negative subject, with respective cumulative distribution functions, $F_{\widehat{Y}_+}(.)$ and $F_{\widehat{Y}_-}(.)$. To ensure a sensitivity level of at least $\underline{Se}$, $k \geq \left[\left(\frac{1}{g\left(F^{-1}_{\widehat{Y}_+}(1-\underline{Se})\right)}\right) - 1\right]$, and to ensure a specificity level of at least $\underline{Sp}$, $k \leq \left[\left(\frac{1}{g\left(F^{-1}_{\widehat{Y}_-}(\underline{Sp})\right)}\right) - 1\right]$.*

A.3 Case Study

A.3.1 Case Study Results

We do not have reliable data on false negative CF cases in the North Carolina data set. We observe that: (i) the CF prevalence rates for the four racial groups in our data set are lower than those reported in the literature [36, 42, 58] (Table A1); and (ii) the sensitivity level of the **PB** 5% policy (North Carolina's IRT screening policy for the validation data set) reported in the literature is lower than what is calculated from the data set (Table A2). These observations indicate the possible existence of false negative cases in our data set. Consequently, (i) we use Monte Carlo simulation to randomly generate additional CF positive cases for the validation data set, based on the difference between the CF prevalence rate in the data set and the lowest prevalence rate reported in the literature for each race (Table A1), so as to match the sensitivity of the **PB** 5% policy reported in the literature [33];

(ii) randomly assign each generated CF case to a testing day in the validation data set; and (iii) randomly generate the IRT reading for each generated CF case, using a truncated IRT reading distribution for CF positive cases ($\widetilde{Y}_+$), truncated from above by the IRT reading corresponding to the 95^{th} quantile on that particular testing day, obtained from the validation data set. In particular, we fit a distribution to $\widetilde{Y}_+$, the IRT reading of CF positive cases, by using the IRT readings for all CF positive newborns in the entire data set, augmented by the IRT readings of false negative CF cases reported in the literature [42]. The best fit for $\widetilde{Y}_+$ is a Gamma distribution, with a shape parameter of 4.2979 and rate parameter of 0.02557 (based on the Chi-square test). Thus, for each testing day to which a CF positive case is added, the 5% IRT threshold (which varies from 20.1 ng/mL to 66.2 ng/mL in the data set) is calculated and used to truncate the $\widetilde{Y}_+$ distribution from above, to ensure that the generated IRT reading is below the truncation level, to represent that this particular CF positive case is missed by the **PB** 5% policy. This process adds an average of 1.73 ± 0.16 (mean±SD) CF positive cases to the validation data set. Table A2 shows that this process also matches the sensitivity of the **PB** 4% policy with the sensitivity level reported in the literature.

Table A1: CF prevalence rate comparison for the different racial groups

Race	*CF prevalence rate (literature)* [source]	*CF prevalence rate used (literature)*	*CF prevalence rate (original data set)*	*CF prevalence rate (95% CI) (adjusted data set)*[a]
Caucasian	2.40×10^{-4}[42], 3.12×10^{-4}[36], 4.00×10^{-4}[58]	3.12×10^{-4}	2.83×10^{-4}	2.98×10^{-4} $(2.78 \times 10^{-4}, 3.18 \times 10^{-4})$
African American	6.66×10^{-5}[36, 58], 1.10×10^{-4}[42]	6.66×10^{-5}	4.78×10^{-5}	6.76×10^{-5} $(5.25 \times 10^{-5}, 8.26 \times 10^{-5})$
Hispanic	8.81×10^{-5}[58], 1.08×10^{-4}[42], 1.09×10^{-4}[36]	8.81×10^{-5}	8.32×10^{-5}	$9.69 \times 10^{-5}(9.36 \times 10^{-5}, 1.04 \times 10^{-4})$
Asian	1.11×10^{-5}[36], 2.86×10^{-5}[58]	1.11×10^{-5}	0	$5.43 \times 10^{-6}(4.45 \times 10^{-6}, 6.40 \times 10^{-6})$
Overall[b]		2.11×10^{-4}	1.87×10^{-4}	1.92×10^{-4} $(1.88 \times 10^{-4}, 2.19 \times 10^{-4})$

[a] After the addition of CF positive cases via simulation
[b] Derived based on the racial distribution in North Carolina (see Table 4.1)

Table A2: Sensitivity level comparison for **PB** (4% and 5%)

Policy	*Sensitivity (literature [source])*	*Sensitivity (original data set)*	*Sensitivity (adjusted data set) (95% CI)*[a]
PB (4%)	94.60% [33]	97.8%	94.28% (93.94%, 94.62%)
PB (5%)	96.50% [33]	100%	96.37% (96.04%, 96.71%)

[a] After the addition of CF positive cases via simulation

Table A3: Performance of various **PB** policies (Validation data set)

Policy	*False negatives (95% half width)*	*False positives*	*Sensitivity*	*Specificity*
PB (1%)	7.73 (0.16)	2,495	83.80%	98.90%
PB (1.5%)	5.73 (0.16)	3,661	87.99%	98.39%
PB (2%)	4.73 (0.16)	4,778	90.09%	97.90%
PB (2.5%)	3.73 (0.16)	5,927	92.18%	97.40%
PB (3%)	3.73 (0.16)	7,075	92.18%	96.89%
PB (3.5%)	2.73 (0.16)	8,205	94.28%	96.40%
PB (4%)	2.73 (0.16)	9,350	94.28%	95.89%
PB (4.5%)	2.73 (0.16)	10,500	94.28%	95.39%
PB (5%)	1.73 (0.16)	11,636	96.37%	94.89%
PB (5.5%)	1.69 (0.16)	12,780	96.46%	94.39%
PB(6%)	1.65 (0.15)	13,949	96.54%	93.88%

Table A4: Average misclassification cost for various **PB** policies studied in Table A3, in terms of $k = \frac{c_{FN}}{c_{FP}}$ (Validation data set)

	k=1,000	*k=2,000*	*k=4,000*	*k=6,000*	*k=8,000*	*k=10,000*
PB (1%)	10,225	17,955	33,415	48,875	64,335	79,795
PB (1.5%)	9,391	15,121	26,581	38,041	49,501	60,961
PB (2%)	9,508	14,238	23,698	33,158	42,618	52,078
PB (2.5%)	9,657	13,387	20,847	28,307	35,767	43,227
PB (3%)	10,805	14,535	21,995	29,455	36,915	44,375
PB (3.5%)	10,935	13,665	19,125	24,585	30,045	35,505
PB (4%)	12,080	14,810	20,270	25,730	31,190	36,650
PB (4.5%)	13,230	15,960	21,420	26,880	32,340	37,800
PB (5%)	13,366	15,096	18,556	22,016	25,476	28,936
PB (5.5%)	14,470	16,160	19,540	22,920	26,300	29,680
PB (6%)	15,599	17,249	20,549	23,849	27,149	30,449

Table A5: Performance of various **CB** policies (Validation data set)

Policy	*False negatives (95% half width)*	*False positives*	*Sensitivity*	*Specificity*
CB (45)	2.24 (0.15)	14,970	95.31%	93.43%
CB (46)	2.30 (0.15)	13,804	95.18%	93.94%
CB (47)	2.39 (0.15)	12,805	94.99%	94.38%
CB (48)	2.46 (0.15)	11,868	94.85%	94.80%
CB (49)	2.50 (0.15)	10,994	94.76%	95.18%
CB (50)	2.56 (0.14)	10,227	94.64%	95.52%
CB (51)	2.62 (0.14)	9,485	94.51%	95.84%
CB (52)	3.66 (0.14)	8,792	92.33%	96.15%
CB (53)	3.68 (0.14)	8,175	92.29%	96.42%
CB (54)	3.69 (0.14)	7,552	92.27%	96.69%
CB (55)	3.70 (0.13)	6,996	92.25%	96.93%
CB (56)	3.71 (0.13)	6,539	92.22%	97.13%
CB (57)	3.71 (0.13)	6,064	92.22%	97.34%
CB (58)	4.72 (0.13)	5,640	90.11%	97.53%
CB (59)	4.73 (0.13)	5,244	90.09%	97.70%
CB (60)	4.75 (0.13)	4,858	90.05%	97.87%
CB (61)	4.77 (0.12)	4,524	90.00%	98.02%
CB (62)	5.01 (0.12)	4,213	89.50%	98.15%
CB (63)	5.10 (0.12)	3,928	89.31%	98.28%
CB (64)	5.25 (0.12)	3,675	89.00%	98.40%
CB (65)	6.25 (0.12)	3,431	86.90%	98.50%

Table A6: Average misclassification cost for various **CB** policies listed in Table A5, in terms of $k = \frac{c_{FN}}{c_{FP}}$ (Validation data set)

	k=1,000	*k=2,000*	*k=4,000*	*k=6,000*	*k=8,000*	*k=10,000*
CB (45)	17,210	19,450	23,930	28,410	32,890	37,370
CB (46)	16,104	18,404	23,004	27,604	32,204	36,804
CB (47)	15,195	17,585	22,365	27,145	31,925	36,705
CB (48)	14,328	16,788	21,708	26,628	31,548	36,468
CB (49)	13,494	15,994	20,994	25,994	30,994	35,994
CB (50)	12,787	15,347	20,467	25,587	30,707	35,827
CB (51)	12,105	14,725	19,965	25,205	30,445	35,685
CB (52)	12,452	16,112	23,432	30,752	38,072	45,392
CB (53)	11,855	15,535	22,895	30,255	37,615	44,975
CB (54)	11,242	14,932	22,312	29,692	37,072	44,452
CB (55)	10,696	14,396	21,796	29,196	36,596	43,996
CB (56)	10,249	13,959	21,379	28,799	36,219	43,639
CB (57)	9,774	13,484	20,904	28,324	35,744	43,164
CB (58)	10,360	15,080	24,520	33,960	43,400	52,840
CB (59)	9,974	14,704	24,164	33,624	43,084	52,544
CB (60)	9,608	14,358	23,858	33,358	42,858	52,358
CB (61)	9,294	14,064	23,604	33,144	42,684	52,224
CB (62)	9,223	14,233	24,253	34,273	44,293	54,313
CB (63)	9,028	14,128	24,328	34,528	44,728	54,928
CB (64)	8,925	14,175	24,675	35,175	45,675	56,175
CB (65)	9,681	15,931	28,431	40,931	53,431	65,931

A.3.2 Comparison of the proposed two-step regression approach with a single-step logistic regression approach

In this section, we perform a single-step logistic regression, in which the dependent variable is the CF risk, and the independent variables, selected by the backward stepwise variable selection method (based on the Akaike Information Criterion (AIC)) (e.g., [27]), include birth weight, race, and IRT reading (i.e., gender, seasonality, weight-gender correlation, and race-weight correlation variables are eliminated). As explained in Section 2.6.3 for the two-step regression model, we consider the n^{th} root of the difference between the IRT reading

and the average of all IRT readings in the training data set[1], which is 24.65, and find the best n by performing a grid search for various odd values of $n \in \{1, 3, 5, \cdots, 97, 99\}$ [17]; the reason for considering only the odd values of n is because the difference can be negative. We find the best value of n to be 3. Then, we perform a repeated five-fold cross validation with stratified sampling, applied to the training data set, to tune the parameters of the single-step logistic regression, resulting in the following equation for predicting each newborn's CF risk:

$$\widehat{p}_j \quad = \quad g(\widetilde{y}_j, \widetilde{\theta}_j) = \frac{1}{1 + e^{(15.78 - 2.151\,(\widetilde{y}_j - 24.65)^{(\frac{1}{3})} - 0.000663\,W_j + 2.371\,R_j^{AF} + 0.8019\,R_j^{H} + 15.05\,R_j^{A})}}, \quad \text{(A.1)}$$

with a p-value less than 2.2×10^{-16}.

We use the single-step logistic regression in Eq. (A.1), along with **EM** or **RM** for different values of k, to classify each newborn in the validation data set; see Table A7, which reports the average number of false negatives and false positives, the misclassification cost, and the sensitivity and specificity levels for different values of $k = \frac{c_{FN}}{c_{FP}}$, over 400 simulation replications. Further, Fig. A.1 compares the performance of the single-step logistic regression with the two-step regression developed in Section 2.6.3, in terms of sensitivity and specificity (based on Tables 2.4 and A7), and indicates that the proposed two-step regression clearly outperforms the single-step regression, over the entire range of sensitivity/specificity values.

[1] In the two-step regression approach, we use the expected IRT value for each newborn, derived based on external and newborn-specific factors, in this equation. Since the single-step regression approach considers all factors in one step, we use the average of all IRT readings in the training data set in this equation.

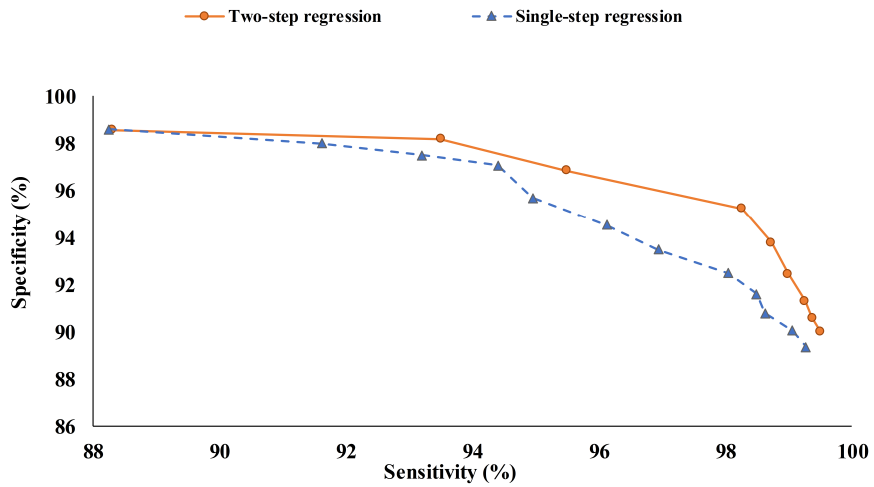

Figure A.1: Sensitivity versus specificity of the single-step and two-step regression models (with variable selection) for different $k = \frac{c_{FN}}{c_{FP}}$ values

Table A7: Single-step logistic regression results with variable selection (the dependent variable is the CF risk, and the independent variables are birth weight, race, and the cube root of the difference between the IRT reading and the IRT reading average for the training data set, i.e., 24.65), and tuned parameters via cross-validation

	Policy	*False negatives (95% half width)*	*False positives*	*Sensitivity*	*Specificity*
k=1,000	**EM**	4.00 (0.16)	3,221	91.62%	98.59%
	RM	1.90 (0.12)	33,166	96.02%	85.44%
k=2,000	**EM**	2.67 (0.11)	6,666	94.4%	97.07%
	RM	0.98 (0.08)	44,433	97.95%	80.49%
k=3,000	**EM**	2.41 (0.11)	9,784	94.95%	95.70%
	RM	0.61 (0.06)	51,293	98.72%	77.48%
k=4,000	**EM**	1.85 (0.09)	12,479	96.12%	94.52%
	RM	0.57 (0.05)	56,053	98.80%	75.39%
k=5,000	**EM**	1.46 (0.09)	14,882	96.94%	93.47%
	RM	0.53 (0.05)	59,677	98.89%	73.80%
k=6,000	**EM**	0.93 (0.08)	17,148	98.05%	92.47%
	RM	0.44 (0.05)	62,472	99.08%	72.57%
k=7,000	**EM**	0.72 (0.08)	19,158	98.49%	91.59%
	RM	0.29 (0.05)	64,814	99.39%	71.55%
k=8,000	**EM**	0.66 (0.06)	21,006	98.62%	90.78%
	RM	0.11 (0.04)	66,715	100%	70.71%
k=9,000	**EM**	0.45 (0.05)	22,711	99.06%	90.03%
	RM	0.05 (0.02)	68,270	99.90%	70.03%
k=10,000	**EM**	0.35 (0.04)	24,262	99.27%	89.35%
	RM	0 (0)	69,719	100%	69.39%