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

May 2018; Revised: October 2018, April 2019; Accepted: May 2019

Saloumeh Sadeghzadeh[1], Ebru K. Bish[2], Douglas R. Bish[2]

[1]School of Management
State University of New York at Binghamton, Vestal

[2]Grado Department of Industrial and Systems Engineering
Virginia Tech, Blacksburg

Abstract

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 or subject-specific factors. These sources of variability can increase the likelihood of subject misclassification based on a biomarker test. We study the minimization of the subject misclassification cost for public health screening of non-infectious diseases, considering regret and expectation-based objectives, and derive various key structural properties of optimal screening policies. Our case study of newborn screening for cystic fibrosis, based on real data from North Carolina, indicates that substantial reductions in classification errors can be achieved through the use of the proposed optimization-based models over current practices.

Keywords: Health care, Public health screening, Biomarker testing, Unobservable risk, Threshold optimization, Newborn screening

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 [38, 65], cardiovascular diseases [64], Alzheimer's disease [24], asthma [63], neurological diseases [43], and various types of cancer [10, 12]. 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*) [38, 65]. 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 [17, 38, 43, 49, 65], level of disease progression, or other medical conditions of the subject [17, 23, 39]. 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 [71, 77], 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 [17, 38, 49, 65]. 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 health care

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 [1]. 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. [6, 7, 9, 8, 57]) and prostate cancer screening (e.g., [12, 15, 50, 67, 78]), as well as for other screening purposes, such as childhood obesity [73]. 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., [44, 55, 70, 71]). 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 [47]. The threshold can correspond to a risk threshold [13, 29, 68], or a biomarker threshold [30, 31, 58, 61, 62, 69], 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., [61, 62]). 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

[61, 62]. 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., [21, 35, 46]). For example, Deneef *et al.* [21] 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.* [46] 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, but 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 presents the notation and the decision problem. Then, Section 3 provides the expectation-based and robust formulations, and Section 4 derives important structural properties of optimal policies. Section 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 6 discusses our case study of cystic fibrosis newborn screening program in North Carolina. Finally, Section 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 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 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 1 for the notation). Note that the width of the uncertainty set on $\overrightarrow{P}$, i.e., the "budget of uncertainty" [14], 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 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)	$\overrightarrow{\widehat{y}} = (\widehat{y}_j)_{j\in\Omega} = (h(\overrightarrow{\theta_j}, \overrightarrow{\widetilde{y}}))_{j\in\Omega}$ $\overrightarrow{\widehat{p}} = (\widehat{p}_j)_{j\in\Omega} = (g(\widehat{y}_j, \widetilde{y}_j))_{j\in\Omega}$	$S(\overrightarrow{P}) = \Big(\big[\underline{p}_j, \overline{p}_j\big]\Big)_{j\in\Omega}$

Remark 1. *Various methods can be employed to derive the support of the risk vector,* $S(\overrightarrow{P}) = \Big(\big[\underline{p}_j, \overline{p}_j\big]\Big)_{j\in\Omega}$*; for instance, by constructing an uncertainty set around* $\overrightarrow{\widehat{Y}}$*, given by* $S(\overrightarrow{\widehat{Y}}) = \Big(\big[\underline{y}_j, \overline{y}_j\big]\Big)_{j\in\Omega}$*, which translates into* $S(\overrightarrow{P}) = \Big(\big[g(\underline{y}_j, \widetilde{y}_j), g(\overline{y}_j, \widetilde{y}_j)\big]\Big)_{j\in\Omega}$*; 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.

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 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:

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

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 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 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:

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

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., [2, 5, 26, 42, 48, 54, 76]), 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 [48]. 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:

$$\begin{aligned} &minimize_{\overrightarrow{x}=(x_j)_{j\in\Omega}}\Big\{max_{\overrightarrow{p}\in S(\overrightarrow{P})}\{Regret(\overrightarrow{x}, \overrightarrow{p})\}\Big\} \\ &subject\ to \qquad x_j\ binary, \forall j \in \Omega. \end{aligned} \tag{2}$$

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.

4 Structural Properties of EM and RM Optimal Solutions

We develop key structural properties of optimal **EM** and **RM** solutions in Section 4.1, and discuss the link between the objectives of minimizing the misclassification cost and maximizing the test efficacy (Appendix B), 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.

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**:*

$$minimize_{p_{th}\in[0,1]} E\left[C(p_{th})|\overrightarrow{\widehat{p}}\right] = c_{FN} \sum_{j\in\Omega:\widehat{p}_j<p_{th}} \widehat{p}_j + c_{FP} \sum_{j\in\Omega:\widehat{p}_j\geq p_{th}} (1-\widehat{p}_j), \tag{3}$$

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 B, 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., [30, 37, 61, 62, 74]). 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.

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**.

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, $\overrightarrow{\hat{p}}$, 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,

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

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:

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

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

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 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.

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 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 [4, 33, 53, 75], 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 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 1 as an example). This follows because*

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

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

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

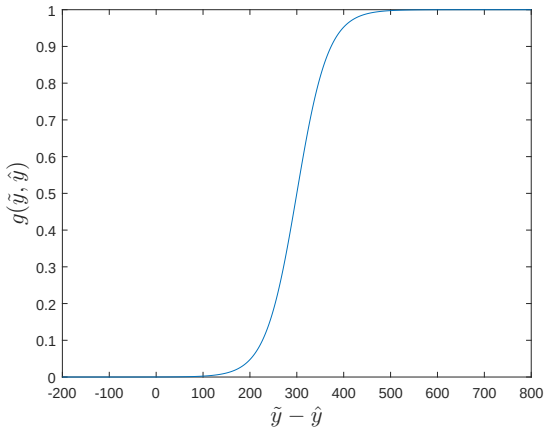

Figure 1: $g(.)$ function that corresponds to the logistic regression model in Remark 2, when $a = -9$ and $b = 0.03$.

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. 1), or very high (e.g., $(\widetilde{y} - \widehat{y}) \in [500, +\infty)$ in Fig. 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, and* $g(.)$ *function given by Remark 2, with* $a = -9$ *and* $b = 0.03$*:*

Table 2: Testing outcomes for Example 1

Subject	$\widetilde{y}$	$\widehat{y}$	$\widetilde{y}-\widehat{y}$	$[\underline{y},\overline{y}]$	$[\underline{p},\overline{p}]=[g(\widetilde{y}-\overline{y}),g(\widetilde{y}-\underline{y})]$	$\frac{\underline{p}+\overline{p}}{2}$	$max_{P\in[\underline{p},\overline{p}]}\left\{\lvert P-\frac{\underline{p}+\overline{p}}{2}\rvert\right\}$
1	700	800	-100	$[750,850]$	$(g(-150),g(-50))=(1.37\times10^{-6},2.75\times10^{-5})$	1.44×10^{-5}	1.31×10^{-5}
2	700	400	300	$[350,450]$	$(g(250),g(350))=(1.824\times10^{-1},8.176\times10^{-1})$	5×10^{-1}	3.2×10^{-1}
3	700	100	600	$[50,150]$	$(g(550),g(650))=(9.994\times10^{-1},1.00)$	9.997×10^{-1}	3×10^{-4}

From Table 2, the maximum deviation of the disease risk used in ***RM****, i.e.,* $\frac{\underline{p}+\overline{p}}{2}$ *(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); 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,* $g(\widehat{y})=P(D=1|\widehat{Y}=\widehat{y})=\frac{qf_{\widehat{Y}_+}(\widehat{y})}{qf_{\widehat{Y}_+}(\widehat{y})+(1-q)f_{\widehat{Y}_-}(\widehat{y})}$.

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., [44]), 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.*

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 [11, 52], CF is one of the most prevalent genetic diseases, and is included in every state's newborn screening panel [1, 45]. Newborn screening for CF allows for early diagnosis, and can substantially improve health outcomes [22, 38]. 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 [28, 56]. 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 [1, 17, 20, 36, 37, 66].

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 [1], 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 [32]. As discussed earlier, newborns with CF tend to have elevated IRT levels [18, 27, 41, 59], 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 [37].

The IRT biomarker fits the modeling assumptions discussed in Section 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., $\overrightarrow{\tilde{y}}$. For example, low temperatures tend to increase the IRT levels of *all* newborns, with or without CF [38, 65]. 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 [17, 38, 49, 65]. 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 6.3.

The remainder of this section is organized as follows. In Section 6.1, we provide an overview of current IRT screening policies used for CF newborn screening. In Section 6.2, we discuss our data sources and calibration. In Section 6.3, we model the relationship between IRT readings and true IRT levels through regression analysis. In Section 6.4, we perform a case study to compare the proposed optimal policies with current IRT screening policies.

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 [37], while Washington uses a threshold of 100 ng/mL [65].

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% [1, 38], while Massachusetts uses a PB policy with a threshold of 5% [19].

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 [37, 38, 49, 65]. Therrell *et al.* [65] state that **PB** outperforms **CB**, especially in regions that experience higher fluctuations in seasonal temperatures, and Kloosterboer *et al.* [38] 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.

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 3 summarizes the demographic characteristics of newborns screened by the NCSLPH during the study period.

Table 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) [1, 65]; and **(2) PB** policy with proportion thresholds of: 4% (Florida, North Carolina, Wisconsin [38, 65]), and 5% (Texas, New York, Massachusetts [65]), as well as other possible **CB** and **PB** policies.

As discussed in Section 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).

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 [38]. 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., [25]) 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) [25]), 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) [25]. 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 $\bar{\bar{y}}_t^R$ 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., [40]). 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:

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

with a p-value less than 2.2×10^{-16}. Thus, Eq. (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 [60]; 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. (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 [40]), 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 3 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., [16]) 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:

$$\widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j) = E(D_j|\widetilde{y}_j - \widehat{y}_j) = \frac{1}{1 + e^{(13.30609 - 2.04352\,(\widetilde{y}_j - \widehat{y}_j)^{\frac{1}{3}} + 2.44\,r_j^{AF} + 0.96496\,r_j^{H} + 14.55\,r_j^{A})}}, \quad j \in \Omega, \ (5)$$

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 (4) and (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. (4)) and $\widehat{p}_j = g(\widehat{y}_j, \widetilde{y}_j)$ (via Eq. (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. (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.

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 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 4-5 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 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 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 B). To study this aspect, Fig. 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 4, A3, and A5. Fig. 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).

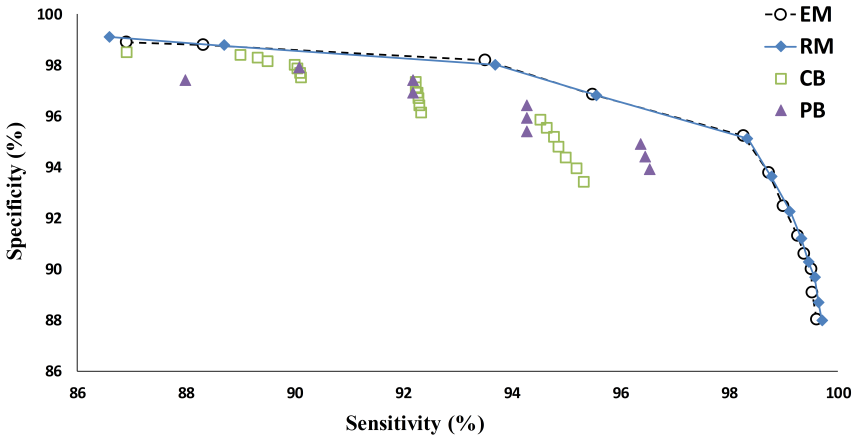

Figure 2: Sensitivity versus specificity of various IRT screening policies

Table 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 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 [72]. 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. 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 5. Moreover, in Table 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 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.

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 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 [3], monitoring the progression of a disease [51], or evaluating the effectiveness of a specific treatment [34]. 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.

Acknowledgments: We are grateful to Professor Ziya, the AE, and three anonymous Reviewers for excellent suggestions that greatly improved the analysis and presentation of the paper. We are thankful to Dr. Scott J. Zimmerman, Director of the North Carolina State Laboratory of Public Health, for offering us valuable insights into public health screening practices for cystic fibrosis. We are also grateful to Dr. Sara Beckloff, Manager of the North Carolina State Laboratory of Public Health's newborn screening, for helping us with the data collection process. This material is based upon work supported in part by the National Science Foundation under Grant No. 1761842. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] (). "Cystic Fibrosis Foundation." https://www.cff.org/, accessed April 2018.

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

[3] 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**, p. 1779.

[4] 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)**, pp. 398–407.

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

[6] Ayer, T., Alagoz, O., and Stout, N.K. (2012). "OR Forum—A POMDP approach to personalize mammography screening decisions." *Operations Research*, **60 (5)**, pp. 1019–1034.

[7] 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)**, pp. 1339–1362.

[8] 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)**, pp. 1333–1358.

[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)**, pp. 600–617.

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

[11] 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)**, p. 231.

[12] 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)**, pp. 815–826.

[13] 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)**, p. 3.

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

[15] 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)**, pp. 105–118.

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

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

[18] 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)**, pp. e495–e518.

[19] 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)**, pp. 1573–1581.

[20] 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)**, p. 1159.

[21] 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)**, pp. 126–132.

[22] 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)**, pp. 1118–1123.

[23] 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)**, pp. 975–979.

[24] 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)**, pp. 1318–1325.

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

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

[27] 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)**, pp. S4–S14.

[28] 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)**, pp. 1–13.

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

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

[31] 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)**, pp. 123–132.

[32] 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)**, pp. 769–774.

[33] 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)**, pp. 2344–2348.

[34] 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)**, pp. 1036–1043.

[35] 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)**, pp. 406–415.

[36] 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)**, p. 557.

[37] 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)**, pp. 1062–1072.

[38] 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)**, pp. e338–e346.

[39] 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)**, pp. 1176–1187.

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

[41] 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)**, pp. 222–225.

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

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

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

[45] 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*. Springer, pp. 17–23.

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

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

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

[49] 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)**, pp. 607–607.

[50] 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**, pp. 9–21.

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

[52] 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*, pp. clinchem–2010.

[53] 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*.

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

[55] 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)**, pp. 73–81.

[56] 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)**, pp. 19–28.

[57] 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)**, p. 6.

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

[59] 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)**, pp. S83–S88.

[60] 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)**, p. 2242.

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

[62] 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)**, p. 53.

[63] 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)**, pp. S9–S23.

[64] 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)**, pp. e99–e109.

[65] Therrell, B.L., Hannon, W.H., Hoffman, G., Ojodu, J., and Farrell, P.M. (2012). "Immunoreactive trypsinogen (IRT) as a biomarker for cystic fibrosis: challenges in newborn dried blood spot screening." *Molecular Genetics and Metabolism*, **106 (1)**, pp. 1–6.

[66] 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)**, pp. 181–186.

[67] 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)**, pp. 293–309.

[68] 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*.

[69] 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)**, pp. 141–150.

[70] 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**, pp. 156–166.

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

[72] 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*, pp. peds–2011.

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

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

[75] 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)**, p. 16.

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

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

[78] 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)**, pp. 529–547.