Authors:
Hussein El Hajj: Ph.D candidate, Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, 24061-01182, United States, email: *hme35@vt.edu*
Douglas Bish: Professor, Department of Information Systems, Statistics, and Management Science, University of Alabama, Tuscaloosa, Alabama, 35487, United States, email: *drbish@cba.ua.edu*
Ebru Bish: Professor, Department of Information Systems, Statistics, and Management Science, University of Alabama, Tuscaloosa, Alabama, 35487, United States, email: *ekbish@cba.ua.edu*

Keywords: Genetic testing, newborn screening, cystic fibrosis, subset-sum problem, robust optimization, price of robustness

Optimal Genetic Screening for Cystic Fibrosis

Hussein El Hajj[1,*], Douglas R. Bish[2], Ebru K. Bish[2]

1. Grado Department of Industrial and Systems Engineering, Virginia Tech
Blacksburg, VA 24061-0118, United States
2. Department of Information Systems, Statistics, and Management Science, University of Alabama
Tuscaloosa, AL 35487, United States

Submitted for publication in Operations Research, the "Policy Modeling and Public Sector OR" area.

October 2018; Revised: December 2019, September 2020, December 2020; Accepted: January 2021

Abstract

Cystic fibrosis (CF) is a life-threatening genetic disorder. Early treatment of CF-positive newborns can extend lifespan, improve quality of life, and reduce healthcare expenditures. As a result, newborns are screened for CF throughout the United States. Genetic testing is costly; therefore, CF screening processes start with a relatively inexpensive, but not highly accurate, biomarker test. Newborns with elevated biomarker levels are further screened via genetic testing for a *panel* of variants (types of mutations), selected among hundreds of CF-causing variants, and newborns with mutations detected are referred for diagnostic testing, which corrects any false-positive screening results. Conversely, a false-negative represents a missed CF diagnosis, and delayed treatment. Therefore, an important decision is which CF-causing variants to include in the genetic testing panel so as to reduce the probability of a false-negative under a testing budget that limits the number of variants in the panel. We develop novel deterministic and robust optimization models, and identify key structural properties of optimal genetic testing panels. These properties lead to efficient, exact algorithms, and key insights. Our case study underscores the value of our optimization-based approaches for CF newborn screening compared to current practices. Our findings have important implications for public policy.

Keywords: Genetic testing, newborn screening, cystic fibrosis, subset-sum problem, robust optimization, price of robustness

[*] Corresponding Author: *hme35@vt.edu*

Abstract

Cystic fibrosis (CF) is a life-threatening genetic disorder. Early treatment of CF-positive newborns can extend lifespan, improve quality of life, and reduce healthcare expenditures. As a result, newborns are screened for CF throughout the United States. Genetic testing is costly; therefore, CF screening processes start with a relatively inexpensive, but not highly accurate, biomarker test. Newborns with elevated biomarker levels are further screened via genetic testing for a *panel* of variants (types of mutations), selected among hundreds of CF-causing variants, and newborns with mutations detected are referred for diagnostic testing, which corrects any false-positive screening results. Conversely, a false-negative represents a missed CF diagnosis, and delayed treatment. Therefore, an important decision is which CF-causing variants to include in the genetic testing panel so as to reduce the probability of a false-negative under a testing budget that limits the number of variants in the panel. We develop novel deterministic and robust optimization models, and identify key structural properties of optimal genetic testing panels. These properties lead to efficient, exact algorithms, and key insights. Our case study underscores the value of our optimization-based approaches for CF newborn screening compared to current practices. Our findings have important implications for public policy.

Keywords: Genetic testing, newborn screening, cystic fibrosis, subset-sum problem, robust optimization, price of robustness

1 Introduction and Motivation

Newborn screening (NBS) tests newborns for genetic disorders that can be life-threatening or that may have a profoundly negative effect on development, if not treated early. NBS is a highly successful public health initiative in the United States (US), and saves thousands of infants and children from disability and death each year through early detection and treatment of genetic disorders (Grosse (2015)). NBS is conducted through in-vitro laboratory tests, performed on dried blood spots routinely collected from newborns (via a heel-prick). In the US, NBS is administered at the state level, and while there is a recommended set of 60 genetic disorders (ACHDNC (2018)), the number of disorders screened for by each state varies from 31-70 (Baby's First Test (2019)). With a prevalence of around 1 in 3,700 newborns, cystic fibrosis (CF) is one of the most prevalent disorders included in NBS (Kammesheidt et al. (2006)). In 2004, the Centers for Disease Control and Prevention recommended CF for NBS (Grosse et al. (2004)), and since 2010 every state's NBS program includes CF screening (Cystic Fibrosis Foundation (2009)). Important testing design decisions arise in NBS, and in this paper we focus on the design of its genetic testing component, using CF as a model disorder.

We first provide a brief background on genetic disorders, and refer the interested reader to in-depth resources, e.g., Jorde et al. (2015). Humans have 23 pairs of chromosomes (22 autosomal pairs and one pair of sex chromosomes). Each chromosome is a sequence (i.e., string) of deoxyribonucleic acid (DNA) molecules; each DNA molecule has one of four bases (adenine (A), thymine (T), guanine (G), and cytosine (C)), and the sequence of these bases encode information. A gene is a segment of chromosome that codes instructions to build a specific protein (e.g., an enzyme). One chromosome of each pair is inherited from each parent, and each pair of autosomal chromosomes has the same set of genes; thus humans have two copies of each autosomal gene, one from each parent. A genetic disorder occurs when a harmful mutation, i.e., an error (extra, missing, or wrong base(s)) occurs in the DNA sequence, which inhibits the correct production of the protein coded by the gene, leading to specific physical symptoms. For instance, CF is caused by mutations to the cystic fibrosis transmembrane conductance regulator (CFTR) gene that inhibit the newborn's ability to make a properly functioning CFTR protein (Kerem et al. (1989); Boyle and Boeck (2013)). Currently there are 312 known CF-causing *variants* (types of harmful mutations), that is, specific, inheritable errors in the CFTR genetic code (CFTR2 (2018)). The CFTR protein plays an important role in regulating the flow of chloride in and out of cells (NIH (2019)). As a result, CF-positive subjects suffer from a chloride imbalance that causes mucus to be more viscous, especially in the lungs. This can be life-threatening: CF's clinical presentation includes chronic lung disease, pancreatic insufficiency, failure to thrive (e.g., 90% of CF-positive infants suffer from fat malabsorption by one year of age), meconium ileus, and infertility (NIH (2019)). Early detection of CF, before symptoms appear, can improve the nutritional status and growth, and extend the lifespan, of the newborn, and reduce hospitalization and complications throughout the newborn's lifetime (Grosse et al. (2004); Accurso et al. (2005); Campbell and White (2005); Farrell et al. (2005); Borowitz et al. (2009)).

Similar to many NBS disorders, CF is an autosomal recessive disorder, which means that to have CF, a newborn must inherit a mutation, of any CF-causing variant, from *each* parent; inheriting only one such mutation, thus having one functional CFTR gene, results in asymptomatic carrier status. (As long as a newborn inherits a CF-causing mutation, the specific variant does not alter the newborn's CF-positivity or carrier status.) Due to a relatively high number of carriers in the US population (1 in 35 individuals, Grosse et al. (2004); Cystic Fibrosis Foundation (2019)) coupled with states' limited screening resources (e.g., testing budget and capacity), and diagnostic testing that is expensive and inconvenient, the main purpose of CF NBS is to identify newborns with CF, and not to determine the carrier status. The limited testing budget/capacity is a major driver for the design of public health screening policies, including NBS (e.g., van den Akker et al. (2006); Mehta (2007); Nshimyumukiza et al. (2014); Grosse (2015); van der Ploeg et al. (2015); CDPH (2019); Schmidt et al. (2019)).

In the US, the following tests, which are conducted on a dried blood spot, are used in CF NBS: **Immunoreactive Trypsinogen (IRT) Test:** CF-positive newborns tend to have higher levels of IRT in their blood than CF-negative newborns (e.g., Cunningham and Taussig (2013)). Measuring IRT levels involves a relatively low-cost biomarker test (around $1.5, Lee et al. (2003)), but unfortunately there is an overlap between the IRT levels of CF-positive and CF-negative newborns (Kloosterboer et al. (2009); Sadeghzadeh et al. (2020)). As a result, the accuracy of the IRT test is not very high, and is dependent on the state's IRT threshold policy. For instance, in the state of New York, which is the subject of our case study, based on NBS data between 2007 and 2012, the IRT test had a positive predictive value (the conditional probability of having CF, given that the IRT test is positive) of only 3.29%, resulting from a sensitivity of 96.54% and a specificity of 99.39% (Kay et al. (2015)).

Mutation Panel (MP) Test: This genetic test searches for a *panel* (i.e., subset) of CF-causing variants, and is more expensive than the IRT test (e.g., a 25-variant panel costs $50.70, Rosenberg and Farrell (2005); Rock et al. (2005)), but it is also highly sensitive and specific in detecting the variants in the panel. A panel can contain *any* subset of variants (e.g., Lim et al. (2016)), but the *panel size* (number of variants in the panel) is restricted, with many current MP technologies having a panel size limit of around 90 variants (Lim et al. (2016)). While the MP test is commonly used by many states, the composition of its panel varies among the states. The American College of Medical Genetics recommends a panel of at least 23 variants (Deignan et al., 2020), and many states use panels of 30 to 50 variants, e.g., New York's panel includes 39 variants, and California's panel includes 40 variants; and both California and New York have customized and changed their panels over the years (Kay et al. (2015); Kharrazi et al. (2015)).

Genetic Sequencing (GS) Test: This genetic test sequences the entire CFTR gene (thus generating the sequence of all the bases) and uses bio-informatics to search for all known CF-causing variants, and is significantly more expensive than both the IRT and MP tests (e.g., $200, Illumina (2019)), but it is also highly sensitive and specific in detecting all known CF-causing variants.

While each state designs its own CF NBS process, all processes begin with the IRT test (due to the low cost), and ends with a classification of the newborn as *screen-negative* or *screen-positive*. Testing stops for newborhs classified as screen-negative, but screen-positives are referred for an expensive diagnostic *Sweat Chloride* (SC) test ($237, Wells et al. (2012)). The SC test requires the newborn to be taken to a specialized testing facility, and can cause parental stress and anxiety as well as out-of-pocket costs (e.g., travel, missed work) (Tluczek et al. (2005); Wells et al. (2012); Cystic Fibrosis Foundation (2017)). Further, while the SC test is highly accurate, it can be inconclusive in around 10% of the cases, requiring a second test on another day (Cystic Fibrosis Foundation (2017); Pagaduan et al. (2018)). Thus, all false screen-positives (*false-positives*) are eventually rectified through the SC test, but at the expense of unnecessary SC testing (Willis (2012)).

With the discovery of the CFTR gene in 1989 (Kerem et al. (1989)), genetic testing for CF became viable. However, due to the high cost of genetic testing, in the US it is used post-IRT, i.e., for newborns with elevated IRT levels, based on the state's IRT threshold policy, so as to mitigate the low positive predictive value of the IRT test. Thus, in between the IRT test and classification, most states use a genetic test or tests, as part of one of the following two processes: **(1)** the most common *one-tier* process, see Fig. 1a (e.g., North Carolina, New York, Virginia, Wisconsin); and **(2)** the *two-tier* process, see Fig. 1b (e.g., California). As Fig. 1 indicates, genetic testing has not replaced the SC test, because the SC test is the only universally accepted test for CF diagnosis. While this may change in the future, this is beyond the scope of this paper. However, our models can be modified to explore the impact of this issue on CF NBS process design.

Due to the importance of genetic testing in CF screening (as well as screening for other genetic disorders) and the complexity of the screening process, we focus on the genetic component of the NBS process, that is, post-IRT testing, and study the one-tier and two-tier genetic testing processes (Fig. 1). For both these processes, the composition of the MP panel, i.e., the specific variants to search for in MP, has significant implications on the classification accuracy and cost of screening, and therefore **optimal panel design**, which we refer to as the *Mutation Variant Selection Problem* (**MSP**), is the focus of this paper. Our **research objectives** are to develop models to determine an optimal MP panel, and to characterize the conditions under which the one-tier or two-tier genetic testing process should be used, so each state can customize their CF NBS process considering their constraints as well as characteristics of the state's population.

In summary, *false-positives* are corrected by the diagnostic SC test (Fig. 1), and hence do not contribute to misclassification, but the unnecessary SC testing is costly, inconvenient, and anxiety-causing for families. Conversely, *false-negatives* (i.e., CF-positive newborns classified as screen-negative) result in delayed CF diagnoses, potentially leading to poor health outcomes and high healthcare expenditures. Each state's public health laboratory is the decision-maker on MP panel composition and NBS process selection, and their primary objective is to reduce the false-negative rate, but their screening effort is naturally constrained by the state's testing budgets (e.g., the high cost of testing is the primary reason all newborns are not screened for CF via the highly sensitive and specific GS test), and they are also sensitive to sending newborns for unnecessary SC testing (even though the latter cost is not borne by the lab). Thus, designing an NBS process necessitates examining the societal perspectives of this decision, to consider both the resources consumed by the state's public health laboratory and the burden of SC testing on families, insurance companies, and the state (e.g., Medicaid); and we study the policy insights of the trade-off between total testing cost (screening and diagnostic) and screening accuracy (false-negative rate) in CF NBS. Towards this end, we represent this trade-off in our formulations of **MSP** through the use of a testing budget, which represents the per person burden for the testing from a societal perspective. We solve **MSP** for a wide range of testing budgets to inform the decision-maker of the trade-offs involved, so that the best decision can be made.

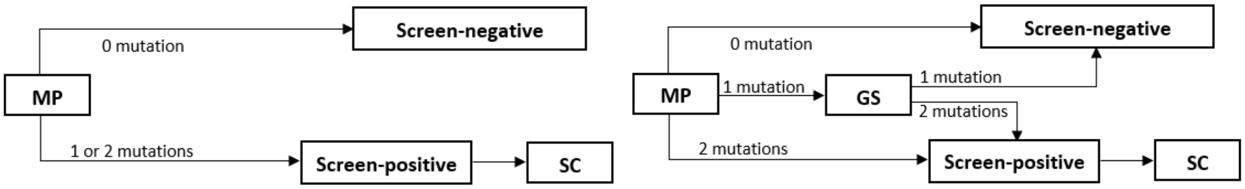
Another important consideration is the uncertainty around the prevalences of CF-causing variants, which stems from many factors, from rare variants to shifting demographics to limited data. Thus, the goal of **MSP** is to determine an MP panel for each process (one-tier and two-tier), so as to minimize the false-negatives, under uncertainty on variant prevalences and considering the trade-off between false-negatives and the testing cost.

Previous literature on NBS processes focuses mostly on descriptive analyses (e.g., Hughes et al. (2015); Kay et al. (2015); Kharrazi et al. (2015); Currier et al. (2017)), with a small number of predictive analyses that use Monte Carlo simulation to estimate the false-negative rate and the testing cost for current testing processes (e.g., Wells et al. (2012)). A few studies use discrete-event simulation to study operational issues in NBS, so as to improve its timeliness (e.g., Cochran et al. (2018)); such operational issues are outside the scope of this paper. While this work falls within the general area of resource allocation in public health screening, the specific application domain, of genetic testing in newborns, coupled with the biological issues surrounding genetic disorders, gives rise to unique decision problems and models, such as **MSP**, which differ substantially from other public health screening domains (e.g., donated blood screening, infectious disease screening), necessitating novel analysis and customized algorithms. For example, considering donated blood screening, Bish et al. (2014) and El Amine et al. (2018) study the problem of allocating a screening budget among a set of infections to minimize the expected risk of a transfusion-transmitted infection, as well as to minimize the maximum regret (which lacks a closed-form expression), formulate these problems as variations of the Knapsack Problem with continuous decision variables, and establish key structural properties of an optimal budget allocation under each objective function. In another application area, infectious disease screening, several papers consider pooled testing in which specimens (e.g., blood) from multiple subjects are combined into testing pools, with each pool tested by only one test, so as to increase the efficiency of testing (e.g., Dorfman (1943); Hwang (1975); McMahan et al. (2012); Aprahamian et al. (2018); Aprahamian et al. (2019)). In this context, the decision problem is to determine pool sizes, and assignment of subjects, with different risk estimates, to testing pools. Compared to this line of research, **MSP** involves individual, not pooled, testing, and the decision is characterized by binary decision variables, representing a subset of variants to include in the panel, among the set of variants that can cause the genetic disorder.

Our contributions in this paper are multi-fold. From a modeling perspective, we introduce **MSP**, a novel decision problem, and develop both deterministic and robust optimization models for **MSP**, under uncertainty on variant prevalences, considering the genetic testing processes used in most states, i.e., the one-tier and two-tier processes. This modeling framework takes a distribution-free approach, which is desirable in practice due to data scarcity. To our knowledge, this paper is the first to provide a mathematical model and prescriptive analysis of this decision problem. From a methodological perspective, we identify key structural properties of optimal panels; and use these properties to develop efficient, exact algorithms. **MSP** is not a trivial problem; in both the deterministic and robust settings, a greedy approach of selecting a number of the highest frequency variants does not always lead to an optimal panel. From an application perspective, we characterize the conditions (e.g., in terms of the cost structure and variant prevalences) under which the two-tier process outperforms the one-tier process; this characterization is important due to the continuously evolving cost structure of genetic testing, newly discovered variants, and changing prevalences. We also develop various insights and quantify the value of robust optimization in our context. Our case study illustrates that, for the given data and cost structure, the two-tier process reduces false-negatives, over its one-tier counterpart, by 18%-46%, depending on the testing budget (for larger budgets the reduction is less, as the panels are larger, and cover more of the higher prevalence variants), while at the same time referring only about 3.6% as many newborns for SC testing compared to the one-tier process. We also find that robust panels, which require only an uncertainty set and the first moment of the prevalence vector of the CF-causing variants, reduce false-negatives by 7.3%, and the worst-case false-negative rate by almost 3.5% on average, compared to the deterministic panels, which are sensitive to estimation errors. (Over the 17 budgets considered, four have the same panel for the robust and deterministic models, and the robust model dominates the deterministic model for all other budget levels.) We also show that our models improve upon current practices without leading to an increase in the testing cost. In particular, both deterministic and robust models lead to a reduction in the false-negative rate, compared to the New York panel, by 5.7% and 17.0%, respectively. These results have important public policy implications. This research is timely, because an early CF diagnosis, and an understanding of the underlying genetic defect, are likely to become more important in the treatment of this disease, especially as new pharmacological and genetic therapies are discovered, placing this type of research on the forefront of personalized, precision medicine.

The remainder of this paper is organized as follows. In Section 2, we present the notation, assumptions, and the decision problem. In Section 3, we provide novel formulations of **MSP**, derive key structural properties of optimal MP panels, develop an efficient, exact algorithm, characterize the conditions under which each of the one-tier and two-tier process dominates, and study the *price of robustness*, i.e., the increase in false-negatives in return for a robust solution. In Section 4, we perform a case study and apply the proposed optimization-based approaches to realistic data. Finally, we conclude, in Section 5, with a summary of our findings and suggestions for future research. To facilitate the presentation, all derivations and proofs, and some results and tables, are relegated to the Appendix.

Figure 1: CF NBS Processes Currently Used in the US



(a) CF NBS with one-tier genetic testing

(b) CF NBS with two-tier genetic testing

2 The Notation, Assumptions, and the Decision Problem

Each individual has a pair of CFTR genes, one from each parent, and each gene can have a CF-causing mutation or not. An individual is CF-negative if they are mutation-free or have exactly one mutation (giving them carrier status), and CF-positive if they have two mutations; in the latter case, both mutations can be of the same variant, or of different variants, and this does not affect the CF-positivity. CF-positive parents are rare, due to the rarity of the disease (there are around 30,000 CF-positive individuals in the US (Cystic Fibrosis Foundation (2018))), and due to CF related health complications. For example, CF reduces fertility, especially in males (Hull and Kass (2000); NIH (2019)); and in 2018 there were only 280 pregnancies in CF-positive women (Cystic Fibrosis Foundation (2018)) out of 3,791,712 births in the US (CDC (2018)). Consequently, in our study, we consider the adult population that excludes CF-positive adults as a proxy for the parent population of interest, as we formally state in Assumption ($\boldsymbol{A_1}$). The scope of our study is the genetic testing component of the CF NBS process, i.e., post-IRT testing.

Throughout, we use upper-case letters to denote random variables, lower-case letters to denote their realization, and bold-face to denote vectors. We denote the sample space or the uncertainty set (the latter is constructed by the decision-maker) of a random variable X by $S(X)$. Let Ω denote the set of variants known to cause CF, which currently contains 312 variants. Let $Q_i, i \in \Omega$, and Q_0 respectively denote the conditional probability that a random parent has one mutation of variant i, and is mutation-free, given that the parent has at most one mutation (i.e., not CF-positive), with $\boldsymbol{Q} = (Q_i)_{i \in \Omega \cup \{0\}}$ representing the prevalence vector in the parent population of newborns undergoing genetic testing, excluding CF-positive parents. Then, $\boldsymbol{Q}$ is a *correlated* random vector, because $\sum_{i \in \Omega \cup \{0\}} Q_i = 1$ with probability one.

There is uncertainty around the prevalences of CF-causing variants. Therefore, we develop a robust optimization model that utilizes an uncertainty set, $S(\boldsymbol{Q})$, and a point estimate of the mean vector, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_i)_{i \in \Omega \cup \{0\}}$, for the random prevalence vector $\boldsymbol{Q}$, that is, the distribution of $\boldsymbol{Q}$ is not needed. Specifically, the robust optimization model utilizes a *correlated uncertainty set*, $S(\boldsymbol{Q}) = \Big\{\boldsymbol{\theta} : \theta_i \in [q_i^{LB}, q_i^{UB}], i \in \Omega \cup \{0\}, \sum_{i \in \Omega \cup \{0\}} \theta_i = 1\Big\}$, where the range of each Q_i, i.e., $q_i^{UB} - q_i^{LB}, i \in \Omega \cup \{0\}$, depends on the available data as well as the level of conservatism the decision-maker is seeking (see Section 4 and Appendix E.1). We assume, without loss of generality, that $q_0^{LB} \geq 1 - \sum_{i \in \Omega} q_i^{UB}$, and $q_0^{UB} \leq 1 - \sum_{i \in \Omega} q_i^{LB}$. To study the price of robustness associated with robust optimization, we also consider a deterministic model that only requires the estimated mean vector, $\hat{\boldsymbol{\mu}} \in S(\boldsymbol{Q})$.

The decision problem, i.e., the *Mutation Variant Selection Problem* (**MSP**), is to determine which variants to search for in MP (i.e., *panel*) for each process represented in Fig.1, so as to minimize the false-negatives, under a technological limit on panel size and a post-IRT testing budget. While the primary objective in NBS is to reduce the false-negatives, a trade-off exists between a false-negative, which leads to a missed diagnosis, and the testing cost (including the cost resulting from false-positives, which require further testing). The testing budget constraint is used to model this trade-off, so as to enable the decision-maker to select the panel based on the trade-offs involved at various budget levels (see Section 1).

We make the following assumptions.

Assumption $(\boldsymbol{A_1})$**.** *Each parent has at most one mutation, that is, each parent is either a carrier (i.e., with one mutation) or mutation-free, independently of the other parent.*

Assumption $(\boldsymbol{A_2})$**.** *All CF-causing mutation variants are known and included in set* Ω.

Assumption $(\boldsymbol{A_3})$**.** *Both genetic tests, MP and GS, and the diagnostic test, SC, are perfectly reliable, i.e., the sensitivity (true positive probability) and specificity (true negative probability) of each genetic test is 1 for each variant searched for in MP, and for the entire set of variants searched for in GS, and SC can identify all CF-positive and CF-negative subjects accurately.*

Assumption $(\boldsymbol{A_1})$ is reasonable, because it is highly unlikely for a newborn to have a CF-positive parent, as we explain at the beginning of this section. Assumption $(\boldsymbol{A_2})$ is reasonable based on current medical knowledge; further, if a CF-causing variant is not yet discovered, then its prevalence is likely very small. Finally, Assumption $(\boldsymbol{A_3})$ is validated by CF genetic testing data (e.g., Kammesheidt et al. (2006); Johnson et al. (2007); Kosheleva et al. (2017); John Hopkins Medicine (2019)), which indicate that both MP and GS have almost perfect sensitivity and specificity. Then, MP will fail to identify a mutation only if the corresponding variant is not included in its panel; and GS will identify all mutations corresponding to the variants in set Ω. Finally, SC is a diagnostic test, which is considered the gold standard for CF diagnosis.

Decision Variable Vector:

$\boldsymbol{x} = (x_i)_{i \in \Omega}$, with $x_i = 1$ if variant i is included in the MP panel, and $x_i = 0$ otherwise

Random Variables:

Q_i, $i \in \Omega \cup \{0\}$: Conditional probability that a random adult has one mutation of variant $i \in \Omega$, or is mutation-free ($i = 0$), given that the adult has at most one mutation

N : Number of mutations, of all variants in set Ω, in a random newborn, with sample space $S(N) = \{0, 1, 2\}$

$\widetilde{N}(\boldsymbol{x})$, $\forall \boldsymbol{x}$: Number of mutations detected by MP in a random newborn, given an MP panel $\boldsymbol{x}$, with sample space $S\left(\widetilde{N}(\boldsymbol{x})\right) = \{0, 1, 2\}$

$Cost^l, l = 1, 2$: Testing cost for genetic screening process type l, with $l = 1, 2$, respectively corresponding to the one-tier and two-tier process

By Assumptions ($\boldsymbol{A_1}$) - ($\boldsymbol{A_3}$), the number of mutations detected by MP is such that $Pr\left(\widetilde{N}(\boldsymbol{x}) \leq N\right) = 1$ for all $\boldsymbol{x}$; and the number of mutations detected by GS, which is independent of $\boldsymbol{x}$, always equals N, i.e., the number of mutations the newborn has.

Parameters:

$m = |\Omega|$: Number of CF-causing variants (i.e., cardinality of set Ω)

$\overline{m}$: Maximum number of variants that can be included in an MP panel due to technological limitations of the testing platform, where $\overline{m} \leq m$

$S(\boldsymbol{Q})$: Uncertainty set around $\boldsymbol{Q}$, constructed by the decision-maker, where $S(\boldsymbol{Q}) = \left\{\boldsymbol{\theta} : \theta_i \in [q_i^{LB}, q_i^{UB}], i \in \Omega \cup \{0\}, \sum_{i \in \Omega \cup \{0\}} \theta_i = 1\right\}$

$\hat{\mu}_i$, $i \in \Omega \cup \{0\}$: Point estimate of $E(Q_i)$, where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_i)_{i \in \Omega \cup \{0\}} \in S(\boldsymbol{Q})$

C_{GS}, C_{SC}, C_{MP} : Cost of each GS test and SC test, and the fixed cost of each MP test, respectively (e.g., cost of testing kit, labor, machine time)

$g\left(\sum_{i \in \Omega} x_i\right)$: Variable cost of MP (e.g., reagent cost associated with each variant), which is strictly increasing in panel size, $\sum_{i \in \Omega} x_i$, with inverse function $g^{-1}(.)$

B $(\geq C_{MP})$: Post-IRT CF testing budget per newborn (i.e., for newborns subject to further testing after IRT)

Since GS searches for all variants in set Ω and SC is not a genetic test (i.e., it does not search for mutations), both GS and SC incur only a fixed cost, of C_{GS} and C_{SC} per newborn, respectively. On the other hand, MP incurs a fixed cost of C_{MP} per newborn and a variable cost, $g\left(\sum_{i \in \Omega} x_i\right)$, that is an increasing function of the number of variants in the panel (and not the specific variants), hence it is invertible. This is realistic for a mutational probe technology (e.g., PCR-based technologies, see Johnson et al. (2007) for details), which is commonly used in CF NBS (see Section 4), where a genetic probe (a sequence of DNA bases) adheres to the strand of DNA with a specific variant if that variant exists, and produces a signal. Thus, for each variant a unique probe is needed, and the variable cost corresponds to the cost of the probe, which is independent of the variant. However, the actual cost structure depends on the pricing practices of manufacturers, therefore we do not make any assumptions on the functional form of $g(.)$.

Then, in both the one-tier and two-tier processes, a false-negative classification, which we denote by event FN, occurs only if the newborn is CF-positive (i.e., $N = 2$) and no mutations are detected in MP (i.e., $\widetilde{N}(\boldsymbol{x}) = 0$); this happens when the newborn's specific variants are not included in the MP panel (see Assumptions ($\boldsymbol{A_2}$) and ($\boldsymbol{A_3}$)). This follows because if one mutation is found in MP, then, depending on the NBS process, the newborn will undergo GS and/or SC testing, both of which will eliminate a false-negative classification; see Fig. 1 (recall that missing a carrier is not considered a false-negative). Then, the probability of a false-negative classification, given a realization $\boldsymbol{\theta}$ of the random vector $\boldsymbol{Q}$, follows (see Appendix A.2 for derivations):

$$Pr(FN(\boldsymbol{x};\boldsymbol{\theta})) = Pr\left(\widetilde{N}(\boldsymbol{x}) = 0, N = 2; \boldsymbol{\theta}\right) = \frac{1}{4}\left[(1-\theta_0) - \left(\sum_{i \in \Omega} x_i \theta_i\right)\right]^2. \tag{1}$$

For both the one-tier and two-tier processes, post-IRT testing starts with MP, and follows stochastically different routes, based on the outcome of MP (i.e., random variable, $\widetilde{N}(\boldsymbol{x})$), see Fig. 1. Using the structure of each testing process depicted in Fig. 1, we derive their expected testing cost (with expectation taken over random variables N and $\widetilde{N}(\boldsymbol{x})$), given a point estimate, $\hat{\boldsymbol{\mu}}$, of the random vector $\boldsymbol{Q}$ (see Appendix A.3 for derivations):

$$\begin{aligned}
E_{N,\widetilde{N}(\boldsymbol{x})}\left[Cost^1(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] =& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left[Pr\left(\widetilde{N}(\boldsymbol{x}) \geq 1\right)\right] C_{SC} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right) C_{SC} - \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2 \left[\frac{C_{SC}}{4}\right] \qquad (2) \\
E_{N,\widetilde{N}(\boldsymbol{x})}\left[Cost^2(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] =& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + Pr\left(\widetilde{N}(\boldsymbol{x}) = 1\right)\left[C_{GS} + Pr\left(N=2|\widetilde{N}(\boldsymbol{x}) = 1\right) C_{SC}\right] + Pr\left(\widetilde{N}(\boldsymbol{x}) = 2\right) C_{SC} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)\left[\frac{2C_{GS} + C_{SC}\left(1-\hat{\mu}_0\right)}{2}\right] - \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2 \left[\frac{2C_{GS} + C_{SC}}{4}\right].
\end{aligned}$$

In our distribution-free approach, we do not take expectation over the random vector $\boldsymbol{Q}$ (which, due to the quadratic terms, would require estimating not only the first two moments of each Q_i, but also the terms, $E[Q_iQ_j]$, for correlated random variables Q_i and Q_j, $\forall i,j \in \Omega, j \neq i$), and simply utilize its point estimate vector, $\hat{\boldsymbol{\mu}}$, in a deterministic manner. Then, the expected testing cost for both genetic screening processes has the same functional form, that is, for $l = 1, 2$:

$$E\left[Cost^l(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] = C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + K_1^l \sum_{i\in\Omega} x_i\hat{\mu}_i - K_2^l \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2, \qquad (3)$$

where (K_1^l, K_2^l) equals $\left(C_{SC}, \frac{C_{SC}}{4}\right)$ for $l = 1$, and $\left(\frac{2C_{GS}+C_{SC}(1-\hat{\mu}_0)}{2}, \frac{2C_{GS}+C_{SC}}{4}\right)$ for $l = 2$.

To simplify the subsequent notation, if a result applies to both genetic screening processes, then we drop the process type index l, and simply refer to the corresponding testing cost as $Cost$, and to the corresponding parameters as K_1 and K_2; we also define $B' \equiv B - C_{MP}$.

3 Optimal Mutation Selection: Formulations and Properties

In **MSP**, the objective is to select a set of variants (panel) for MP so as to minimize the false-negative probability, under a given testing budget and a technological constraint on the maximum number of variants allowable in a panel. In this section, we develop and study two formulations of **MSP**: a *deterministic* formulation that simply uses a point estimate vector, $\hat{\boldsymbol{\mu}}$, and a *robust* formulation that uses an uncertainty set, $S(\boldsymbol{Q})$, for variant prevalences, and the point estimate vector, $\hat{\boldsymbol{\mu}}$. The deterministic formulation allows us to quantify the price of robustness.

We first provide some properties for the false-negative probability and the expected testing cost (Eqs. (1) and (3)). Recall that $\boldsymbol{x}$ is a binary decision variable vector, $\sum_{i\in\Omega\cup\{0\}} \theta_i = 1, \forall\boldsymbol{\theta} \in S(\boldsymbol{Q})$, and $g(.)$ is strictly increasing in panel size.

Property 1.

1. *The objective,* $\left\{\text{minimize}_{\boldsymbol{x}}\, Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\right\}$*, is equivalent to the objective,* $\left\{\text{maximize}_{\boldsymbol{x}} \sum_{i\in\Omega} x_i\theta_i\right\}$*, and the objective,* $\left\{\text{minimize}_{\boldsymbol{x}} \max_{\boldsymbol{\theta}\in S(\boldsymbol{Q})} \left\{Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\right\}\right\}$*, is equivalent to the objective,*

$\left\{\text{maximize}_{\boldsymbol{x}} \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}\right\}$.

2. The expected testing cost, $E[Cost^l(\boldsymbol{x};\hat{\boldsymbol{\mu}})]$*, is strictly increasing in* x_i*, for all* $i \in \Omega$*,* $l = 1, 2$*.*

We are ready to provide the deterministic and robust formulations of **MSP**.

Deterministic MSP (DM):

$$\begin{aligned}
&\underset{\boldsymbol{x}}{\text{minimize}} && Pr\left(FN(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right) && && \underset{\boldsymbol{x}}{\text{maximize}} && \sum_{i\in\Omega} x_i\hat{\mu}_i && (4)\\
&\text{subject to} && E\left[Cost(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] \le B && \Leftrightarrow && && g\left(\sum_{i\in\Omega} x_i\right) + K_1\sum_{i\in\Omega} x_i\hat{\mu}_i - K_2\left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2 \le B' && (5)\\
& && \sum_{i\in\Omega} x_i \le \overline{m} && && && \sum_{i\in\Omega} x_i \le \overline{m} && (6)\\
& && x_i \text{ binary, } i\in\Omega && && && x_i \text{ binary, } i\in\Omega && (7)
\end{aligned}$$

Robust MSP (RM):

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\right\} \quad \Leftrightarrow \quad \underset{\boldsymbol{x}}{\text{maximize}} \quad \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} \qquad (8)$$

$$\text{subject to} \qquad (5), (6), (7).$$

DM objective function (4) minimizes the false-negative probability for a given point estimate vector, $\hat{\boldsymbol{\mu}}$, while **RM** objective function (8) minimizes the worst-case false-negative probability among all possible realizations of the random vector $\boldsymbol{Q}$ (see Property 1). Constraint (5) ensures that the expected testing cost, based on the point estimate $\hat{\boldsymbol{\mu}}$, remains within the testing budget, while Constraint (6) imposes a technological limit on MP panel size. As discussed in Section 1, all false-positive outcomes are resolved by SC (Assumption ($\boldsymbol{A_3}$) and Fig. 1), and hence, are not included in the objective function, but they do contribute to the testing cost in Constraint (5). Let $\boldsymbol{x}^{*k}$ and Z^{*k}, $k \in \{D, R\}$, respectively denote the optimal solution vector and the corresponding objective function value for **DM** ($k = D$) and **RM** ($k = R$).

DM and **RM** lead to novel decision problems and model structures, due to the specific form of the nonlinear cost constraint (Constraint (5)), where the first term is a function of the sum of the binary decision variables (i.e., $g(\sum_{i\in\Omega} x_i)$); and, additionally, due to the correlation structure of the uncertainty set used in **RM**. Consequently, in the following, we establish their structural properties, which allow us to reformulate each problem as a series of sub-problems (Theorem 1 and Remark 2). In particular, each **DM** sub-problem is reformulated as an exact γ-item 0-1 Subset-sum Problem, and each **RM** sub-problem is reformulated as a *novel* variation of an exact γ-item 0-1 maximin Knapsack Problem, with a correlated uncertainty set (for any $\gamma \in Z^+$).

RM can be conservative, because it accounts for the worst possible value of the objective function (see (8)), but **RM** still considers the expectation of the testing cost in the budget constraint (i.e., Constraint (5) is computed based on $\hat{\boldsymbol{\mu}}$). This is to preserve the feasible region of **DM** so that the robust and deterministic solutions can be directly compared and insights can be derived. However, other robust formulations, with different degree of conservatism, are possible, based on how the uncertainty in the testing cost is handled (see Remark 1). In the remainder of the paper, we provide our analysis for **RM**, which forms the basis of the analysis for the other robust models (see Appendix B). We compare all robust models in Section 4.

Remark 1. Observe that for any given panel $\boldsymbol{x}$, the specific prevalence vector realizations that respectively yield the worst-case objective function (maximum false-negative probability) and the worst-case (maximum) testing cost do not necessarily coincide. We suggest two ways to model a robust version of Constraint (5), i.e., without utilizing the point estimate vector, $\hat{\boldsymbol{\mu}}$. These formulations differ in their treatment of the uncertain testing cost (i.e., in the prevalence vector used in Constraint (5)):

Correlated Robust MSP (C-RM) considers the worst-case objective (as in **RM**), that is, for any $\boldsymbol{x}$, it evaluates the false-negative probability at the prevalence vector that maximizes it, i.e., $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \equiv \text{argmax}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\Big\{Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\Big\} \equiv \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\Big\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\Big\}$ (if multiple such vectors exist, then, without loss of optimality, the vector that yields the smallest expected testing cost is selected). The model uses this prevalence vector, $\boldsymbol{\beta}^{FN}(\boldsymbol{x})$, in both objective function (8) and Constraint (5).

Upper Bound Robust MSP (U-RM) considers both a worst-case objective (as in **RM** and **C-RM**), but also a worst-case testing cost in Constraint (5), that is, for any $\boldsymbol{x}$, it evaluates the false-negative probability at prevalence vector $\boldsymbol{\beta}^{FN}(\boldsymbol{x})$, defined above, but evaluates the testing cost at the prevalence vector that maximizes it, i.e., $\boldsymbol{\beta}^{Cost}(\boldsymbol{x}) \equiv \text{argmax}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\Big\{E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]\Big\}$. Thus, the model potentially uses different prevalence vectors in objective function (8) and Constraint (5).

Thus, while **C-RM** is consistent in terms of how the uncertainty is resolved, both **U-RM** and **RM** use potentially different prevalence vectors in their objective function and testing cost constraint. Therefore, the three robust models offer different degrees of conservatism, with **U-RM** always being the most conservative, which is typically followed by **RM** and then **C-RM** (however, the relationship between **RM** and **C-RM** is inconclusive in general).

The trade-off between solution robustness and accuracy has been studied in the OR literature in general, considering various objective functions, such as minimax type or regret-based objectives (e.g., Savage (1951); Perakis and Roels (2008); El Amine et al. (2018)); and various forms of uncertainty sets, including interval type uncertainty sets and others (e.g., ellipsoidal, polyhedral, conic sets), or sets that restrict the number of uncertain parameters in an effort to reduce the conservatism of the robust solution (e.g., Ben-Tal and Nemirovski (2000); Bertsimas and Sim (2004); Bertsimas et al. (2011); Soleimanian and Jajaei (2013)). Our novel robust formulations, **RM, C-RM,** and **U-RM**, all of which rely on a correlated uncertainty set, $S(\boldsymbol{Q}) = \Big\{\boldsymbol{\theta} : \theta_i \in [q_i^{LB}, q_i^{UB}], i \in \Omega\cup\{0\}, \sum_{i\in\Omega\cup\{0\}}\theta_i = 1\Big\}$, the derivation of their structural properties, and the bounding of the price of robustness in our setting, provide a contribution to the literature, while extending this line of research to a new application area, of genetic testing.

Even with a linear objective function implied by Property 1, **DM** remains a difficult problem, because Constraint (5) has not only quadratic terms but also an arbitrary nonlinear function

$g(.)$ that is a function of the sum of the binary decision variables. **RM**, with a nonlinear objective function and a correlated uncertainty set, is at least as difficult as **DM**. While **DM** can be reformulated as a mixed integer linear programming problem (MILP) following the standard linearization techniques (e.g., Adams and Sherali (1986); Smith and Taskin (2008)), the MILP reformulation adds a large number of binary variables and constraints (see Remark C.2 in Appendix C). This is compounded by the fact that the number of binary variables in **DM** and **RM**, which corresponds to the number of known CF-causing variants (m), continues to grow: This number was 130 in April 2012, but increased to 272 by August 2016, and further to 312 by December 2017 (CFTR2 (2018)). Our numerical analysis in Section 3.2 indicates that the problem size, m, coupled with an arbitrary nonlinear cost function, $g(.)$, leads to computational issues for off-the-shelf solvers. Therefore, in the remainder of this section, we establish key structural properties of **DM** and **RM** formulations, and use these properties to develop a customized, exact algorithm, which is able to solve realistic sized **DM** and **RM** problem instances in a highly efficient manner (within a few seconds).

3.1 Equivalent Formulations of DM and RM

We first formulate, and establish various properties of, a specific sub-problem of each of **DM** and **RM**, which we refer to as **DM-S**(γ) and **RM-S**(γ), respectively. This is done by imposing an additional constraint on the original problem that exogenously fixes the number of variants to be used in any MP panel, i.e., $\sum_{i\in\Omega} x_i = \gamma$, for some $\gamma \in Z^+$:$\gamma \leq \overline{m}$. For these specific sub-problems, i.e., when $\sum_{i\in\Omega} x_i = \gamma$ is fixed for some γ, we are able to express Constraint (5) as a linear function of x_i, $i \in \Omega$, reducing these sub-problems to certain special cases of the Knapsack Problem.

Theorem 1. *Consider any $\gamma \in Z^+$: $\gamma \leq \overline{m}$.*

1. ***DM**, under the additional constraint that $\sum_{i\in\Omega} x_i = \gamma$, can be equivalently formulated as:*
   ***Sub-problem DM-S**(γ):*

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & \sum_{i\in\Omega} x_i \hat{\mu}_i \\ \text{subject to} \quad & \sum_{i\in\Omega} x_i \hat{\mu}_i \leq Z_{UB}(\gamma) && (9) \\ & \sum_{i\in\Omega} x_i = \gamma && (10) \\ & x_i \text{ binary}, \ i \in \Omega. && (11) \end{aligned}$$

2. ***RM**, under the additional constraint that $\sum_{i\in\Omega} x_i = \gamma$, can be equivalently formulated as:*
   ***Sub-problem RM-S**(γ):*

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{ \theta_0 + \sum_{i\in\Omega} x_i\theta_i \right\} \Leftrightarrow \quad \underset{\boldsymbol{x},u}{\text{maximize}} \quad u\left(q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}\right) + (1-u)\left(1 - \sum_{i\in\Omega}(1-x_i)\, q_i^{UB}\right) \\ \text{subject to} \quad & (9), (10), (11) \qquad\qquad\qquad\qquad \text{subject to} \quad (9), (10), (11) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u \text{ binary}, && (12) \end{aligned}$$

where $Z_{UB}(\gamma) \equiv \frac{K_1 - \sqrt{(K_1)^2 + 4[g(\gamma) - B']K_2}}{2K_2}$ *is independent of* $\boldsymbol{x}$ *when* $\boldsymbol{x}$ *satisfies that* $\sum_{i \in \Omega} x_i = \gamma$*; and Constraint (9) becomes redundant when* $Z_{UB}(\gamma)$ *does not exist. Further,* $Z_{UB}(\gamma)$ *is strictly decreasing in* γ *in the region that it exists.*

For a given $\gamma \in Z^+ : \gamma \leq \overline{m}$, let $\boldsymbol{x}^{*k}(\gamma), k \in \{D, R\}$, denote the optimal solution vector to Subproblem $\mathbf{DM\text{-}S(\gamma)}$ and $\mathbf{RM\text{-}S(\gamma)}$, respectively; and let $Z^{*D}(\gamma) = \sum_{i \in \Omega} x_i^{*D}(\gamma)\hat{\mu}_i$ and $Z^{*R}(\gamma) = \max\left\{ q_0^{LB} + \sum_{i \in \Omega} x_i^{*R}(\gamma) q_i^{LB}, 1 - \sum_{i \in \Omega}(1 - x_i^{*R}) q_i^{UB} \right\}$, that is, the corresponding objective function value for $\mathbf{DM\text{-}S(\gamma)}$ and $\mathbf{RM\text{-}S(\gamma)}$, respectively, i.e., with the addition of Constraint (10). In addition, for $k = \{D, R\}$, define $\gamma^{*k} \equiv \sum_{i \in \Omega} x_i^{*k}$, where $\boldsymbol{x}^{*k}$ represents the optimal solution vector to the corresponding original problem, i.e., γ^{*D} and γ^{*R} respectively represent the optimal panel size for **DM** and **RM**.

Remark 2.

1. **RM-S**(γ) reduces to a novel variation of the exact γ-item 0-1 maximin Knapsack Problem, with an uncertainty set that has a specific correlation structure (i.e., the random prevalence vector needs to satisfy, $\sum_{i \in \Omega \cup \{0\}} Q_i = 1$), whereas **DM-S**$(\gamma)$ is an exact γ-item 0-1 Subset-sum Problem, which arises as a special case of the exact γ-item 0-1 Knapsack Problem (Capara et al. (2000)). Consequently, while the solution to the standard exact γ-item 0-1 maximin Knapsack Problem (i.e., without a correlated uncertainty set) reduces to the maximization of the objective function when all item values (prevalences in our setting) take on the lower bound of their support (e.g., Chinneck and Ramadan (2000); Nobibon and Leus (2014)), this is no longer the case for **RM-S**(γ), hence the need to formulate its deterministic counterpart in Theorem 1.

2. Various pseudo-polynomial algorithms have been developed in the literature for the exact γ-item 0-1 Knapsack Problem (which also apply to its special case, the exact γ-item 0-1 Subset-sum Problem), including an exact algorithm with complexity $O(mr)$, where $r = \gamma \times Z_{UB}(\gamma)$ (Capara et al. (2000); Kellerer et al. (2004)); and an approximate algorithm with complexity $O(m\gamma/\epsilon)$, where ϵ denotes the maximum absolute difference allowable between the objective function value generated by the algorithm and the optimal objective function value (i.e., accuracy) (Capara et al. (2000)). Various pseudo-polynomial algorithms have been also developed specifically for the exact γ-item 0-1 Subset-sum Problem, including an exact algorithm with complexity $O(mr')$, where $r' = 1 + Z_{UB}(\gamma) - \sum_{i \in \Omega : i = m - \gamma + 1}^{m} \hat{\mu}_i$ (Pisinger (1998)). All three algorithms are pseudo-polynomial due to their dependence on problem parameters, m, γ, r, or r'.

3. The reformulations in Theorem 1 continue to hold in the following cases:

   (a) The term, $K_1 \sum_{i \in \Omega} x_i \hat{\mu}_i - K_2 \left(\sum_{i \in \Omega} x_i \hat{\mu}_i\right)^2$, in Constraint (5) is replaced by any function $f\left(\sum_{i \in \Omega} x_i \hat{\mu}_i\right)$ that is monotone increasing in $\sum_{i \in \Omega} x_i \hat{\mu}_i$ in the interval $[0, 1 - \hat{\mu}_0]$, equivalently, for all panels of size γ, the larger the prevalences of the variants in the panel $(\sum_{i \in \Omega} x_i \hat{\mu}_i)$, the larger the testing cost is. In this case, $Z_{UB}(\gamma)$ in Theorem 1 needs to be replaced by the (unique) root of the function, $-B' + g(\gamma) + f(\sum_{i \in \Omega} x_i \hat{\mu}_i) = 0$, in

$[0, 1-\hat{\mu}_0]$. If $Z_{UB}(\gamma)$ does not exist, then either there exists no feasible solution with panel size γ, or any solution $\boldsymbol{x} : \sum_i x_i = \gamma$ is feasible with respect to Constraint (5).

(b) The function, $g(\sum_{i\in\Omega} x_i)$, is of any functional form (including non-invertible functions), as long as it is a function of panel size, $\gamma = \sum_{i\in\Omega} x_i$. However, in this case, $Z_{UB}(\gamma)$ is not necessarily monotone decreasing in γ.

With the reformulation in Theorem 1, one can utilize any of the aforementioned algorithms developed in the literature (Remark 2) to solve **DM-S(γ)** and **RM-S(γ)** (for the latter, this applies only when the value of $u \in \{0,1\}$ is fixed, see Theorem 1). Further, for the special case where $g\left(\sum_{i\in\Omega} x_i\right) = 0$, i.e., with no MP variable cost, **DM** can be solved via a single 0-1 Subset-sum Problem, and **RM** via two 0-1 Knapsack Problems, one for each value of $u \in \{0,1\}$ (see Remark C.1 in Appendix C).

In the subsequent sections, we utilize the **DM-S**(γ) and **RM-S**(γ) Sub-problems to solve the original problems, **DM** and **RM**, in an efficient manner.

3.2 Structural Properties and an Exact Algorithm

In this section, we analyze some key properties of the $Z_{UB}(\gamma)$ function (see Constraint (9) in Theorem 1) so as to establish bounds on the optimal number of variants to include in the MP panel, i.e., $\gamma^{*k}, k \in \{D, R\}$. These bounds will be useful for developing an efficient algorithm for **DM** and **RM**, and for establishing the termination criteria.

To this end, without loss of generality, we arrange the variants in set Ω such that $\hat{\mu}_i \geq \hat{\mu}_j, \forall i,j \in \Omega, i < j$, for a given $\hat{\boldsymbol{\mu}}$. We also define Ω^{LB} (Ω^{UB}) as an ordered set of q_i^{LB} (q_i^{UB}), $i \in \Omega$, arranged such that $q_i^{LB} \geq q_j^{LB}$ ($q_i^{UB} \geq q_j^{UB}$), $\forall i,j \in \Omega^{LB}(\Omega^{UB}), i < j$. We define $\Delta \equiv \left\lfloor g^{-1}\left(B' - K_1(1-\hat{\mu}_0) + K_2(1-\hat{\mu}_0)^2\right)\right\rfloor$, which provides the number of variants that is feasible to include in an MP panel if every newborn with at least one mutation is sent for further testing beyond MP, that is, considering an upper bound on the number of newborns sent for GS and/or SC testing.

We first characterize optimal **DM-S(γ)** and **RM-S(γ)** solutions for a special case.

Lemma 1. *Consider any $\gamma \in Z^+ : \gamma \leq \overline{m}$. If $\gamma < \Delta$, then Constraint (9) becomes redundant, and the following results hold:*

1. *An optimal **DM-S(γ)** solution consists of γ variants with the highest $\hat{\mu}_i$ values, $i \in \Omega$, breaking ties arbitrarily, with $Z^{*D}(\gamma) = \sum_{i\in\Omega:i\leq\gamma} \hat{\mu}_i$.*

2-i. *If $q_0^{LB} + \sum_{i\in\Omega^{LB}:i\leq\gamma} q_i^{LB} \geq 1 - \sum_{i\in\Omega^{UB}:i\geq\gamma+1} q_i^{UB}$, then an optimal **RM-S(γ)** solution consists of γ variants with the highest q_i^{LB} values, $i \in \Omega^{LB}$, breaking ties arbitrarily, with $Z^{*R}(\gamma) = q_0^{LB} + \sum_{i\in\Omega^{LB}:i\leq\gamma} q_i^{LB}$.*

2-ii. *If $q_0^{LB} + \sum_{i\in\Omega^{LB}:i\leq\gamma} q_i^{LB} \leq 1 - \sum_{i\in\Omega^{UB}:i\geq\gamma+1} q_i^{UB}$, then an optimal **RM-S(γ)** solution consists of γ variants with the highest q_i^{UB} values, $i \in \Omega^{UB}$, breaking ties arbitrarily, with $Z^{*R}(\gamma) = 1 - \sum_{i\in\Omega^{UB}:i\geq\gamma+1} q_i^{UB}$.*

Lemma 1 holds because any solution vector $\boldsymbol{x}$ that satisfies $\sum_{i\in\Omega} x_i \leq \Delta$ is feasible with respect to Constraint (9). In practice, the budget is often constraining. Therefore, while we do not expect the condition in Lemma 1 (i.e., $\gamma < \Delta$) to hold in general, this result establishes a special case when a greedy panel is optimal, with an optimal **DM-S(γ)** panel consisting of the γ highest prevalence variants (part 1), and an optimal **RM-S(γ)** panel consisting of the γ variants, with either the highest lower bounds (part 2-i), or the highest upper bounds (part 2-ii). This characterization of the greedy solution for the robust sub-problem also shows that its solution differs from the robust (maximin) Knapsack Problem, due to its correlated uncertainty set (Remark 2). Further, this lemma assists with the development of bounds on γ^{*D} and γ^{*R}, i.e., the number of variants to include in an optimal **DM** panel and **RM** panel, respectively.

Theorem 2. *Optimal **DM** and **RM** solutions satisfy:*
$\gamma^{*k} \in \left[\min\left\{\gamma_{LB}, \overline{m}\right\}, \min\left\{\gamma_{UB}^{k}, \overline{m}\right\}\right]$, *$k = \{D, R\}$, where $\hat{\mu}_m = min_{i\in\Omega}\{\hat{\mu}_i\}$, and*

$$\gamma_{LB} \equiv \max\left\{0, \Delta\right\},$$

$$\gamma_{UB}^{D} \equiv \max\left\{0, \min\left\{\left\lfloor g^{-1}\left(B' - K_1\left(\sum_{\substack{i\in\Omega:\\ i\leq\gamma_{LB}}} \hat{\mu}_i\right) + K_2\left(\sum_{\substack{i\in\Omega:\\ i\leq\gamma_{LB}}} \hat{\mu}_i\right)^2\right)\right\rfloor, \left\lfloor g^{-1}\left(B' - K_1\hat{\mu}_m + K_2\hat{\mu}_m^2\right)\right\rfloor\right\}\right\}, \text{ and}$$

$$\begin{aligned}\gamma_{UB}^{R} \equiv \max\Bigg\{0, \min\Bigg\{\Bigg\lfloor g^{-1}\Bigg(B' - K_1\Bigg(\max\Bigg\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\leq\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\geq\gamma_{LB}+1}} q_i^{UB}\Bigg\} - \mu_0\Bigg) \\ + K_2\Bigg(\max\Bigg\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\leq\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\geq\gamma_{LB}+1}} q_i^{UB}\Bigg\} - \mu_0\Bigg)^2\Bigg)\Bigg\rfloor, \left\lfloor g^{-1}\left(B' - K_1\hat{\mu}_m + K_2\hat{\mu}_m^2\right)\right\rfloor\Bigg\}\Bigg\},\end{aligned}$$

Observe that the lower bound, γ_{LB}, is derived based on Δ, which corresponds to the minimum feasible panel size, i.e., considering an upper bound on the testing cost beyond MP, that is, when every newborn with at least one mutation were to be identified and sent for GS and/or SC testing. Also observe that if there exists at least one non-empty panel that is feasible to **DM** and **RM**, then the one-variant panel with variant m (i.e., the variant with the smallest prevalence) will also be feasible. Therefore, the upper bound, γ_{UB}^{k}, $k = \{D, R\}$, is based on either this particular one-variant solution (which is feasible if there exists a feasible non-empty panel), or the solution that uses γ_{LB} variants (which is always feasible). On the other hand, if there exists no feasible non-empty panel to the problem, then $\gamma_{LB} = \gamma_{UB}^{k} = 0$, $k = \{D, R\}$, and the optimal solution is given by the empty panel, i.e., $\boldsymbol{x}^* = \boldsymbol{0}$.

Theorem 2 leads to the following form of optimal **DM** and **RM** solutions for some special cases.

Corollary 1.

1. If $\overline{m} \leq \gamma_{LB}$, then Constraint (5) becomes redundant, and $Z^{*D} = Z^{*D}(\gamma = \overline{m}) = \sum_{i\in\Omega: i\leq\overline{m}} \hat{\mu}_i$, and $Z^{*R} = Z^{*R}(\gamma = \overline{m}) = \max\left\{q_0^{LB} + \sum_{i\in\Omega^{LB}: i\leq\overline{m}} q_i^{LB},\ 1 - \sum_{i\in\Omega^{UB}: i\geq\overline{m}+1} q_i^{UB}\right\}$.
2. If $\overline{m} \geq \gamma_{UB}^{D}$ ($\overline{m} \geq \gamma_{UB}^{R}$), then Constraint (6) becomes redundant, and parameter $\overline{m}$ does not affect an optimal **DM** (**RM**) solution.

Thus, in the special case of $\overline{m} \leq \gamma_{LB}$ (i.e., when the budget constraint, Constraint (5), is redundant), optimal **DM** and **RM** solutions can be obtained in a greedy manner (similar to Lemma 1 for the sub-problems). In practice, however, newborn screening *is* constrained by the testing budget, hence a greedy approach does not necessarily lead to an optimal panel for **DM** or **RM**. To see this, consider **DM** as an example; its greedy solution consists of as many of the highest prevalence variants as the testing budget allows. The higher the prevalence of the selected variant, the higher the reduction in the false-negative probability. However, a higher prevalence variant also leads to a higher expected testing cost (Eq. (3)): Notice that while the MP cost, given by $C_{MP} + g\left(\sum_{i\in\Omega} x_i\right)$, is independent of the panel selection, the post-MP testing cost (i.e., the expected cost of subsequent GS and/or SC testing), given by $K_1^l \sum_{i\in\Omega} x_i\hat{\mu}_i - K_2^l \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2$, $l = 1, 2$, is strictly increasing in the panel prevalence, $\sum_{i\in\Omega} x_i\hat{\mu}_i$, for both testing processes (see Property A.1). Thus, it is possible for a larger, non-greedy panel (i.e., one that does not contain the sequential set of the highest prevalent variants) to have a lower false-negative probability than a greedy panel. Indeed, our case study in Section 4 provides multiple problem instances for which a greedy panel is not optimal for **DM** or **RM**.

The equivalent formulations and structural properties lead to the following exact algorithm for **DM** and **RM**.

Theorem 3. *The* ***Mutation Variant Selection Optimization Algorithm (Algorithm MSO)*** *solves* ***DM*** *and* ***RM*** *to optimality, and is pseudo-polynomial, with complexity:*

1. $O(b^D m r')$ *for* ***DM****, where* $r' = 1 + Z_{UB}(\gamma_{LB}) - \sum_{i\in\Omega: i=m-\gamma+1}^{m} \hat{\mu}_i$*; and*
2. $O(2b^R m r)$ *for* ***RM****, where* $r = \gamma_{UB}^R \times Z_{UB}(\gamma_{LB})$*,*

and $b^k = \min\left\{\gamma_{UB}^k, \overline{m}\right\} - \min\left\{\gamma_{LB}, \overline{m}\right\}$*, for* $k \in \{D, R\}$*.*

Algorithm MSO ($k = \{\text{D,R}\}$)

```
Input: $\gamma_{LB}$, $\gamma_{UB}^k$, $x^{*k}\left(\gamma = \min\{\gamma_{LB}, \overline{m}\}\right)$, $Z^{*k}\left(\gamma = \min\{\gamma_{LB}, \overline{m}\}\right)$
for $\gamma \in \left[\min\{\gamma_{LB}, \overline{m}\} + 1, \min\left\{\gamma_{UB}^k, \overline{m}\right\}\right]$ do
    Compute $Z_{UB}(\gamma)$
    if $Z_{UB}(\gamma) \leq Z^{*k}(\gamma - 1)$ then Stop
    else:
        Solve Sub-problem DM-S($\gamma$) (if $k = D$) and RM-S($\gamma$) (if $k = R$), and obtain $\boldsymbol{x}^{*k}(\gamma)$ and $Z^{*k}(\gamma)$
    end if

    if $Z^{*k}(\gamma) < Z^{*k}(\gamma - 1)$ then $Z^{*k}(\gamma) = Z^{*k}(\gamma - 1), \boldsymbol{x}^{*k}(\gamma) = \boldsymbol{x}^{*k}(\gamma - 1)$
    end if
end for
Output($\boldsymbol{x}^{*k} = \boldsymbol{x}^{*k}(\gamma), Z^{*k} = Z^{*k}(\gamma)$)
```

Observe that **Algorithm MSO** always returns an optimal solution. In particular, when $\gamma_{LB} = \gamma_{UB}^k = 0$, the algorithm returns $\boldsymbol{x}^* = \boldsymbol{0}$ (the only feasible solution), and when $\gamma_{UB}^k > 0$, the algorithm proceeds through iterations to determine an optimal solution. Also note that if $\overline{m} \leq \gamma_{LB}$, the algorithm generates the greedy solution, which is optimal in this case (see Corollary 1), in the initialization step, i.e., $Z^{*k} = Z^{*k}(\overline{m})$, and does not go through any iterations; and it accounts for the second part of Corollary 1 through the bounds on γ.

Lemma 2. *If $g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^n$, for all $n \in Z^+ {:} n \geq 2$, then both γ_{LB} and γ_{UB}^k, $k \in \{D,R\}$, are non-increasing in n. Further, γ_{LB} decreases in n at a slower rate than γ_{UB}^k, $k \in \{D,R\}$.*

Lemma 2 indicates that when the $g(.)$ function is polynomial, the higher its degree, the narrower the range, $\gamma_{UB}^k - \gamma_{LB}$, of the number of variants in an optimal panel (Theorem 2), leading to a potential reduction in the number of γ values that need to be considered in **Algorithm MSO** when determining an optimal panel.

Next we perform a numerical study to compare the efficiency of **Algorithm MSO** for **DM** with its MILP reformulation (see Remark C.2) and with an off-the-shelf solver, Gurobi (Gurobi Optimization (2019)). Table 8 (Appendix C) reports the computational times, as well as the number of additional binary variables and constraints required for the MILP reformulation, for various polynomial functions for the MP variable cost, $g(.)$, and various number of variants, m. As Table 8 indicates, **Algorithm MSO** substantially outperforms both the MILP reformulation and the Gurobi solver (without any reformulation). Further, a main advantage of **Algorithm MSO** is that its efficiency does not degrade for higher-degree polynomials for the $g(.)$ function. This is because **Algorithm MSO** relies on solving sub-problems for which the term, $g\left(\sum_{i\in\Omega} x_i\right)$, reduces to a constant for a given γ, i.e., $g(\gamma)$, which is incorporated into parameter $Z_{UB}(\gamma)$ (Theorem 1), and consequently, its run time is not impacted by the functional form of $g(.)$; this is not the case for the other two methods. In particular, when $g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^3$, i.e., in the presence of a cubic constraint, the Gurobi solver was not able to obtain a solution (without any reformulation), while the MILP reformulation required a large number of additional binary variables and constraints, leading to out of memory issues. In fact, in this numerical study **Algorithm MSO** became slightly more efficient for higher-degree polynomials, mainly because of a reduction in the number of iterations needed ($\gamma_{UB}^D - \gamma_{LB}$), see Theorem 2 and Lemma 2. For instance, for $g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^3$, $\gamma_{LB} = \gamma_{UB}^D = 3(= \gamma^{*D})$, with **Algorithm MSO** not requiring any iteration, whereas for $g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^2$, $\gamma_{LB} = 5$, $\gamma_{UB}^D = 7(= \gamma^{*D})$, with **Algorithm MSO** requiring at most 2 iterations, and for $g\left(\sum_{i\in\Omega} x_i\right) = c\sum_{i\in\Omega} x_i$, $\gamma_{LB} = 33$ and $\gamma_{UB}^D = 36$ ($\gamma^{*D} = 35$), with **Algorithm MSO** requiring at most 3 iterations.

In summary, while a higher degree polynomial for the $g(.)$ function typically degrades the run time needed for off-the-shelf solvers and standard linearization techniques, Lemma 2 and our numerical study indicate that this is not necessarily the case for **Algorithm MSO**.

3.3 Comparison of One-tier and Two-tier Genetic Screening Processes

While most states utilize the one-tier genetic screening process for CF NBS, some others use the two-tier process (Fig.1); hence, an important research objective is to characterize the conditions under which the two-tier process outperforms the one-tier process. To this end, let $\boldsymbol{x}^{*k1}$ and $\boldsymbol{x}^{*k2}$, $k \in \{\text{D, R}\}$, respectively denote the optimal **DM** and **RM** solutions for the one-tier and two-tier process. Then, the two-tier process "dominates" the one-tier process in **DM** if $Pr\left(FN(\boldsymbol{x}^{*D2};\hat{\boldsymbol{\mu}})\right) \leq Pr\left(FN(\boldsymbol{x}^{*D1};\hat{\boldsymbol{\mu}})\right)$, and it dominates in **RM** if $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{Pr\left(FN(\boldsymbol{x}^{*R2};\boldsymbol{\theta})\right)\right\} \leq \max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{Pr\left(FN(\boldsymbol{x}^{*R1};\boldsymbol{\theta})\right)\right\}$.

Theorem 4. *For any problem instance:*

1. *If $\frac{C_{SC}}{C_{GS}} \geq \frac{2}{1+\hat{\mu}_0}$, then the two-tier process dominates for both **DM** and **RM**.*
2. *If $\frac{C_{SC}}{C_{GS}} \leq 1$, then the one-tier process dominates for both **DM** and **RM**.*
3. *If $1 < \frac{C_{SC}}{C_{GS}} < \frac{2}{1+\hat{\mu}_0}$, and if there exists at least one feasible solution $\boldsymbol{x}$ to* **DM** *and **RM** for the one-tier process (i.e., Constraints (5)–(7), with $(K_1^1, K_2^1) = \left(C_{SC}, \frac{C_{SC}}{4}\right)$), such that $\sum_{i\in\Omega} x_i\hat{\mu}_i \geq 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0)$, then the two-tier process dominates for both **DM** and **RM**.*

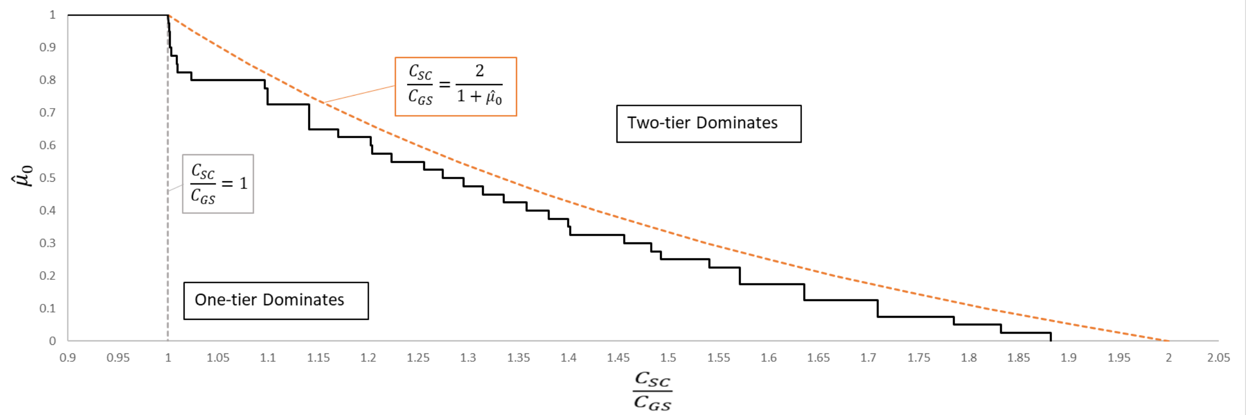
Theorem 4, which delineates the region in which each process dominates, is illustrated in Fig. 2. Observe that the conditions in the first two parts of Theorem 4 (respectively corresponding to the RHS of the orange (dashed) line and the LHS of the gray (dashed and vertical) line in the figure) are functions of only the mutation-free proportion, $\hat{\mu}_0$, and testing costs, C_{GS} and C_{SC}, whereas the condition in the third part (corresponding to the region between the black (solid) line and the orange line, where the black line is numerically computed for each $(\hat{\mu}_0, \frac{C_{SC}}{C_{GS}})$ pair) depends also on budget and cost parameters, B and $g(.)$, as well as the entire prevalence vector, $\hat{\boldsymbol{\mu}}$.

Thus, Theorem 4 completely characterizes the conditions under which each genetic screening process should be utilized. Further, as an immediate consequence of Theorem 4, we can also express the dominance region for the two-tier process as a function of cost parameters and the prevalence vector only.

Corollary 2. If $\frac{C_{SC}}{C_{GS}} \geq \frac{2-\sum_{i\in\Omega:i=1}^{\min\{\gamma_{LB},\overline{m}\}}\hat{\mu}_i}{1+\hat{\mu}_0}$, then the two-tier process dominates for both **DM** and **RM**.

Observe that the dominance region for the two-tier process for **DM** and **RM** expands as: (i) $\frac{C_{SC}}{C_{GS}}$ increases, or (ii) $\hat{\mu}_0$ increases. This can be explained as follows. As $\frac{C_{SC}}{C_{GS}}$ decreases, that is, as SC becomes less costly compared to GS, the one-tier process becomes more desirable, because the SC cost spent on carriers in the one-tier process (i.e., $Pr\left(\widetilde{N}(\boldsymbol{x})=1\right)C_{SC}$) becomes lower than the total cost of GS on carriers, plus cost of GS and SC on CF-positive newborns in the two-tier process (i.e., $Pr\left(\widetilde{N}(\boldsymbol{x})=1\right)C_{GS} + Pr\left(\widetilde{N}(\boldsymbol{x})=1\right)Pr\left(N=2|\widetilde{N}(\boldsymbol{x})=1\right)C_{SC}$); see Eq. (2). For the second part, $Pr\left(\widetilde{N}(\boldsymbol{x})=1\right)$ is strictly increasing in $\hat{\mu}_0$, while $\frac{Pr\left(\widetilde{N}(\boldsymbol{x})=1|N=2\right)}{Pr\left(\widetilde{N}(\boldsymbol{x})=1|N=1\right)}$ is strictly decreasing in $\hat{\mu}_0$ (see Appendix A.1). As a result, higher values of $\hat{\mu}_0$, i.e., the proportion of mutation-free parents, result in more carriers being sent for SC testing in the one-tier process, whereas these carriers are tested via the less expensive GS test in the two-tier process.

Figure 2: Comparison of One-tier and Two-tier Processes for Different Values of $\hat{\mu}_0$ and $\frac{C_{SC}}{C_{GS}}$ for Case Study Parameters ($C_{MP} = \$19.24$, $g(\sum_{i\in\Omega} x_i) = \$1.26 \times (\sum_{i\in\Omega} x_i)$, and with $\hat{\mu}_i, i \in \Omega$, values (from Table 9) Normalized such that $\sum_{i\in\Omega} \hat{\mu}_i = 1 - \hat{\mu}_0$) for a Budget of $B = \$100$



3.4 The Price of Robustness

While the deterministic model **DM** provides a solution that minimizes the false-negative probability for the point estimate of the prevalence vector, the robust model **RM** provides a solution that performs well, in terms of the false-negative probability, for *all* realizations of the random prevalence vector. **RM** does this by minimizing the worst-case false-negative probability (the highest possible false-negative probability for any realization of the prevalence vector, $\boldsymbol{Q}$), but this may come at the expense of an increase in the false-negative probability when the prevalence corresponds to its point estimate, i.e., the input to **DM**. In this section we study this trade-off between solution robustness and the minimization of the false-negative probability on expectation. To isolate the magnitude of this trade-off, we assume that the **DM** input, $\hat{\boldsymbol{\mu}}$, is accurate, i.e., it corresponds to the true mean vector, $\boldsymbol{\mu}$. (Note that the **DM** solution can also lead to a high false-negative probability when $\hat{\boldsymbol{\mu}}$ is a poor estimate of the true mean $\boldsymbol{\mu}$.)

Let $\Pi(\boldsymbol{\mu}) \equiv \frac{Pr\left(FN(\boldsymbol{x}^{*R};\boldsymbol{\mu})\right)}{Pr\left(FN(\boldsymbol{x}^{*D};\boldsymbol{\mu})\right)}$ denote the *price of robustness ratio* for **RM** (e.g., Bertsimas and Sim (2004); Ben-Tal et al. (2009); El Amine et al. (2018)), for a given $\boldsymbol{\mu} \in S(\boldsymbol{Q})$. By this definition, the price of robustness ratio is at least one, and the lower this ratio is, the closer the robust objective function value is to the minimum possible false-negative probability under vector $\boldsymbol{\mu}$.

Theorem 5. *For any given $\boldsymbol{\mu} \in S(\boldsymbol{Q})$, without loss of generality, we represent the uncertainty set of Q_i, $i \in \Omega \cup \{0\}$, $[q_i^{LB}, q_i^{UB}]$, as $[\mu_i\left(1-h_i^{LB}\right), \mu_i\left(1+h_i^{UB}\right)]$, where $h_i^{LB} \equiv \frac{\mu_i - q_i^{LB}}{\mu_i}$ and $h_i^{UB} \equiv \frac{q_i^{UB}-\mu_i}{\mu_i}$. Then, the price of robustness ratio for **RM** can be bounded from above as follows:*

$$\Pi(\boldsymbol{\mu}) \leq \frac{1}{4Pr\left(FN(\boldsymbol{x}^{*D};\boldsymbol{\mu})\right)} \min\left\{1-\mu_0-\min\left\{\left(\frac{\sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i\left(1+h_i^{UB}\right)}{(1+h_{max}^{UB})}\right)^2, \left(\frac{\sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i\left(1-h_i^{LB}\right)}{(1-h_{min}^{LB})}\right)^2\right\},\right.$$

$$\left.\left(1-\mu_0(1-h_0^{LB})-\sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i\left(1-h_i^{LB}\right)\right)^2\right\},$$

where $h_{max}^{UB} \equiv max_{i\in\Omega}\{h_i^{UB}\}$ *and* $h_{min}^{LB} \equiv min_{i\in\Omega}\{h_i^{LB}\}$.

Note that all parameters and optimal solutions in Theorem 5 are dependent on $\boldsymbol{\mu}$ due to the definition of h_i^{LB} and h_i^{UB}, $i \in \Omega$. The bound in Theorem 5 requires knowledge of the optimal **DM** solution; therefore, in Lemma C.1 (Appendix C), we also provide an upper bound that is potentially weaker, but that does not require the optimal **DM** solution. Observe that the upper bound in Theorem 5 is non-decreasing in h_{max}^{UB} and non-increasing in h_{min}^{LB}, because the higher h_{max}^{UB} or the lower h_{min}^{LB} is, the larger the uncertainty set is. Corollary C.1 (Appendix C) provides a special case where the price of robustness ratio is 1. We compare the upper bound in Theorem 5 with the actual price of robustness ratio in the case study.

4 Case Study: Cystic Fibrosis Screening of Newborns

In this section, we present a case study on CF newborn screening, based on CF NBS data from New York (NY) (Kay et al. (2015)). Our focus is on newborns with elevated IRT levels, i.e., those who undergo post-IRT genetic testing (via the dried blood spot). Our objective is to compare the performance of: (i) the one-tier and two-tier genetic screening processes depcted in Fig. 1; (ii) the deterministic and robust models, **DM** and **RM**, and the NY MP panel; and (iii) the three robust model variations, **RM, C-RM**, and **U-RM**.

First we discuss the data sources and model calibration.

4.1 Data Sources and Model Calibration

We first describe the process for generating the variant prevalences for a parent population representative of the parents of newborns genetically tested in NY (i.e., CF-negative parents of post-IRT newborns). We then discuss the construction of the uncertainty set for the robust model, and the testing cost structure.

Both the proportion of the mutation-free population and variant prevalences differ among racial/ethnic groups (Watson et al. (2004); Schrijver et al. (2016)). Therefore, these parameters need to be derived based on NY's diverse population. We use the superscript $r \in R$, where R is the set of groups considered, to denote group-level parameters and random variables, e.g., $\hat{\mu}^r, N^r$. We also define f^r as the proportion of group r in the population, and derive the population-level parameters from their group-level counterparts:

$$\hat{\mu}_i = \sum_{r\in R} f^r \hat{\mu}_i^r, \forall i \in \Omega \cup \{0\}. \tag{13}$$

Due to the low prevalence of CF, coupled with a large number of CF-causing variants, most of which are rare, sparsity in NBS data is an issue that every state must contend with. Thus, we employ a hybrid approach that utilizes data from multiple sources to estimate these parameters, illustrating the types of data a state would have to gather to use our models. These data sources are summarized in Table 1.

Table 1: Data Sources for CF-causing Variants and Population Characteristics

Parameter	Data Source
Set of CF-causing variants (Ω)	CFTR2 data set (CFTR2 (2018))
Proportion of mutation-free parents in group r ($\hat{\mu}_0^r, r \in R$)	NY CF NBS data (Kay et al. (2015))
Variant prevalences in group r parents ($\hat{\mu}_i{}^r$, $i \in \Omega, r \in R$)	Schrijver et al. (2016)
Proportion of group r in the population ($f^r, r \in R$)	NY CF NBS data (Kay et al. (2015))

Population Characteristics: We combine NY's NBS data with other data sets that report the frequency of CF-causing variants in different populations, to estimate $\hat{\mu}$ and to construct the uncertainty set, $S(\boldsymbol{Q})$. Specifically, for NY, Kay et al. (2015) provides six years of CF NBS data on the number of newborns genetically tested and the number of CF-positive newborns identified, for various racial/ethnic groups for the period between 2007 and 2012 (Table 2), but due to the low prevalence of CF, the data on CF-causing variants identified in NY within this six-year period is too sparse to produce accurate estimates of group-level variant frequencies.

On the other hand, the CFTR2 data set (CFTR2 (2018)) reports the frequency of CF-causing variants in more than 88,000 CF-positive subjects from Europe, Canada, and the US, and is the most extensive data set, but does not have racial/ethnic identifiers, and hence is not necessarily representative of the population of NY. Therefore, we use the CFTR2 data set to define the set of CF-causing variants, Ω, and complement this data with Schrijver et al. (2016), which reports the frequencies for the top 50 CF-causing variants for several racial/ethnic groups (each group can have a different set of top 50 variants) derived from more than 35,000 CF-positive subjects in the US. We then incorporate the specific demographics of NY into the estimation.

Specifically, we consider three groups, White (W), Hispanic (H), and Black (B), i.e., $R = \{W, H, B\}$ (these three groups make up 86.3% of the newborns in the NY CF NBS data set in the study period; the data for other specific groups is sparse (Kay et al. (2015)), and normalize the population based on these three groups (Table 2). Parent information is not reported in the NY data set. Therefore, we use the group proportions of newborns in the NY data set as a proxy for the group proportions of their parent population (i.e., parents of post-IRT newborns in NY in the study period, which does not necessarily match the general demographics of NY), and also estimate group-level prevalences assuming that both parents of a newborn in the NY data set belong to the same group as their newborn (this is solely for the estimation procedure; our derivations in Appendix A consider that each parent of a newborn can belong to each group $r, r \in R$, with probability f^r, independently of the other parent).

Based on the data from Kay et al. (2015) (Table 2), we estimate the proportion of mutation-free parents in each group r ($\hat{\mu}_0^r$) via the following relationship (see Eq. (14) in Appendix A.1), and

Table 2: Parameters by Group in NY's CF NBS Data Set for 2007-2012 (from Kay et al. (2015))

Parameter ($r \in R$)	NY (Overall)	W*	H*	B*
Number of CF-positive newborns in group r	237	190	40	7
Number of genetically tested newborns in group r	69,010	32,313	12,780	23,917
Proportion of group r (normalized) in the population (f^r)	1.000	0.468	0.185	0.347

* W: White, H: Hispanic, B: Black

report it in Tables 3 and 9 (Appendix E.2):[1]

$$Pr(N^r = 2) = \frac{1}{4}(1 - \hat{\mu}_0^r)^2 = \frac{\text{number of CF-positive newborns in group } r}{\text{number of genetically tested newborns in group } r}, r \in R.$$

Next, to estimate the prevalence of CF-causing variants in each group (i.e., $\hat{\mu}_i^r, \forall i \in \Omega, r \in R$), we use the data in Schrijver et al. (2016), i.e., each group's 50 most common variants (denoted as set Ω_1^r), from which we calculate $\hat{\mu}_i^r,\ i \in \Omega_1^r$, as follows (see Eq. (15) in Appendix A.1):

$$\hat{\mu}_i^r = \frac{\text{number of mutations of variant } i \text{ in CF-positive newborns in group } r}{\text{number of CF-positive newborns in group } r} \frac{(1 - \hat{\mu}_0^r)}{2}.$$

We do not have group-level frequency data on the remaining set of 262 (=312-50) CF-causing variants (set $\Omega \setminus \Omega_1^r$), i.e., those variants that are not in the top 50 variants for each group, corresponding to those variants with a very low combined frequency (see the last row of Table 3). Therefore, we assume that each of these 262 variants have the same prevalence for a group, and distribute the group's remaining prevalence (remaining proportion of parents with one mutation), i.e., $1 - \sum_{i \in \Omega_1^r \cup \{0\}} \hat{\mu}_i^r$, evenly among these 262 variants:

$$\hat{\mu}_i^r = \frac{1}{262}(1 - \sum_{i \in \Omega_1^r \cup \{0\}} \hat{\mu}_i^r), \forall i \in \Omega \setminus \Omega_1^r,\ r \in R.$$

The variant prevalences in the overall NY parent population of interest is then calculated by Eq. (13). Table 3 reports, for select variants, their prevalence and ranking for the parent population for NY overall and for each group, to demonstrate the differences among the groups. For example, F508del is the most prevalent variant in all groups, and hence in the state, and there is a large gap between the prevalence of F508del and the prevalence of each group's second ranked variant, which differs among the groups. Further, the different groups' rankings do not coincide, for example Variant #9 (3210+1G->A) is the second most prevalent variant in group B, 33rd in group W, and 9th in group H (see Appendix E.2 for a more extensive parameter table). Throughout the remainder of the case study, we refer to specific variants by their NY prevalence ranking, i.e., Variant #1 is F508del (Table 3).

Uncertainty Sets: RM requires an uncertainty set for the variant prevalence vector, $S(\boldsymbol{Q})$, which we construct using the group-level data in Schrijver et al. (2016), which is based on 1,551, 2,578, and

[1] The proportion of mutation-free parents of newborns genetically tested (0.8828) differs from the proportion of mutation-free adults in the general NY population (0.9674, Kay et al. (2015)), because the IRT test removes a large proportion of the CF-negative cases, altering the characteristics of the post-IRT newborn population, and hence the characteristics of their parent population.

Table 3: Prevalences and Rankings for NY, and by Group, for Select Variants (based on Schrijver et al. (2016))

Variant Name	Point Prevalence (Ranking)			
	NY (Overall)	W*	H*	B*
F508del	0.083259 (1)	0.112568 (1)	0.060421 (1)	0.015910 (1)
G542X	0.003435 (2)	0.003527 (3)	0.005818 (2)	0.000547 (5)
G551D	0.002651 (3)	0.003987 (2)	0.000783 (13)	0.000479 (6)
R553X	0.001163 (8)	0.001534 (7)	0.000783 (14)	0.000411 (7)
3210+1G->A	0.001014 (9)	0.000153 (33)	0.001007 (9)	0.00373 (2)
1717+1G->A	0.001002 (10)	0.00138 (8)	0.000671 (17)	0.000171 (19)
3876delA	0.000397 (18)	0.000033 (45)	0.001790 (5)	0.0000242 (44)
R347P	0.000294 (25)	0.000460 (20)	0.000060 (>50)	0.0000242 (47)
W1089X	0.000211 (30)	0.000033 (47)	0.000895 (11)	0.0000242 (>50)
E60X	0.000201 (33)	0.000307 (27)	0.000060 (>50)	0.0000242 (>50)
R75X	0.000173 (42)	0.000033 (>50)	0.000671 (19)	0.000068 (37)
$\hat{\mu}_0$ and $\hat{\mu}_0^r, r \in R$	0.8828	0.8466	0.8881	0.9658
$1 - \sum_{i \in \Omega_1^r \cup \{0\}} \hat{\mu}_i^r, r \in R$		0.008895	0.016900	0.006501

* W: White, H: Hispanic, B: Black

31,286 CF-positive subjects from groups B, H, and W, respectively. In particular, we first derive the lower and upper bounds for each group's prevalences, i.e., $q_i^{r,LB}$ and $q_i^{r,UB}$, for each $Q_i^r, i \in \Omega \cup \{0\}, r \in R$. We do this by constructing a 95% confidence interval for each $Q_i^r, i \in \Omega \cup \{0\}, r \in R$, via the Wilson Score Method with continuity correction (Newcombe (1998)). This method is adopted, because it maintains a value between 0 and 1 for the bounds, and hence, is commonly used for prevalences (e.g., Newcombe (2003)), accounts for continuity correction, utilizes the data available in our setting ($\hat{\mu}_i^r, i \in \Omega, r \in R$, and group sample sizes), and is computationally efficient. Then, the bounds for the overall parent population are derived by, $q_i^{LB} = \sum_{r \in R} f^r q_i^{r,LB}$ and $q_i^{UB} = \sum_{r \in R} f^r q_i^{r,UB}, i \in \Omega \cup \{0\}$, producing ranges that are relatively wider for low frequency (rare) variants. Finally, the overall uncertainty set, $S(\boldsymbol{Q})$, is obtained by superimposing the constraint that $\sum_{i \in \Omega \cup \{0\}} Q_i = 1$, that is, $S(\boldsymbol{Q}) = \left\{\boldsymbol{\theta} : \theta_i \in \left[q_i^{LB}, q_i^{UB}\right], i \in \Omega \cup \{0\}, \sum_{i \in \Omega \cup \{0\}} \theta_i = 1\right\}$; see Appendix E.1 for details.

Testing Parameters: MP cost data are not readily available, list prices are often not public, and the prices paid by the laboratories are often negotiated. Thus, while our analytical results hold for general MP variable cost functions, in the case study we use a linear cost structure, i.e., $g\left(\sum_{i \in \Omega} x_i\right) = c \sum_{i \in \Omega} x_i$, for demonstration purposes. This is a reasonable assumption due to a lack of data, and fits mutational probe technologies, where a genetic probe is added to the MP test for each variant; each probe is similar (i.e., a string of DNA bases that match the variant and surrounding bases, of sufficient length to ensure adhesion with the variant, and not with other DNA strands). Actual pricing may deviate from this linear cost function, but as discussed above, this does not pose a problem for our models.

Using the MP testing cost data in Rosenberg and Farrell (2005), which reports the cost of a one-mutation panel as, $C_{MP}+c = \$20.50$, and the cost of a 25-mutation panel as, $C_{MP}+25c = \$50.70$, we derive $C_{MP} = \$19.24$ and $c = \$1.26$. For genetic sequencing, we set $C_{GS} = \$200$ based on Illumina (2019), and for the SC test, we set $C_{SC} = \$340$ based on Wells et al. (2012).[2] Regarding the panel size limit ($\overline{m}$), many MP technologies have a panel size limit of around 90 variants (Lim et al. (2016)), which we use in the case study; see Table 4 for a summary of testing related parameters and data sources. We note that the testing cost structure used in our case study is mainly for illustrative purposes. Testing costs can differ from state to state, depending on the technology used and whether testing is done by the state laboratory or outsourced. In addition, as genetic testing technology is evolving, so is the cost of genetic testing. Consider that the cost of GS has reduced substantially since its development; for instance, ten years ago, the cost was around ten times higher than today (NIH (2016)). Further, the cost of SC testing is generally not borne by the state laboratory, yet it is an important goal to reduce the number of newborns unnecessarily sent for SC testing, as discussed in Section 1, and we do this through the budget constraint in our models. In particular, we perform our analysis for a set of budgets and study the effect of the budget on the optimal MP panel and the resulting false-negative probability.

Table 4: Testing related Parameters and Data Sources

Parameter	Value	Data Source
Panel size limit ($\overline{m}$)	90	Lim et al. (2016)
MP - fixed cost per test (C_{MP})	\$19.24	Wells et al. (2012)
MP - testing cost per variant (c)	\$1.26	Wells et al. (2012)
GS - cost per test (C_{GS})	\$200	Illumina (2019)
SC - cost per test (C_{SC})	\$340	Rosenberg and Farrell (2005)

4.2 Case Study Results

In this section, we first provide the optimal panels for **DM** and **RM** for the one-tier and two-tier processes under a set of budgets, and discuss panel composition. We then compare the performance of the different panels using a Monte Carlo simulation.

4.2.1 Optimal Panels

We report the optimal panels generated by **DM** and **RM** over a range of budgets, as well as the NY panel and its effective budget for the one-tier process (i.e., NY's process during the study period), given our testing cost parameters (Table 4). The panels are reported in Table 5 in terms of the variant numbers (based on the NY prevalence rank) and, in parentheses, their panel size, leading to the following findings:

- Panel composition: The testing budget, and not the technological limit ($\overline{m} = 90$), restricts the panel sizes in Table 5. In addition, optimal panels do not necessarily contain the most

[2] The cost of the SC test, C_{SC}, includes the testing cost \$237 plus an inconvenience and travel cost of \$103.

Table 5: The Set of Variants in Optimal MP Panels (Panel Size) for **DM** and **RM** for Different Budgets (B), and the NY Panel

Budget ($)	One-tier Process		Two-tier Process	
	DM	**RM**	**DM**	**RM**
65.0	1-10 (10)	1-8, 10, 11 (10)	1-19 (19)	1-17, 19, 25 (19)
67.5	1-11, 15 (12)	1-8, 10-13 (12)	1-21 (21)	1-17, 19, 25-27 (21)
70.0	1-13 (13)	1-13 (13)	1-23 (23)	1-17, 19, 21, 24-27 (23)
72.5	1-15 (15)	1-15 (15)	1-24, 26 (25)	1-19, 21, 22, 24-27 (25)
75.0	1-17 (17)	1-15, 17, 19 (17)	1-27 (27)	1-19, 21-28 (27)
77.5	1-18, 23 (19)	1-17, 19, 25 (19)	1-29 (29)	1-28, 32 (29)
80.0	1-20 (20)	1-17, 19, 25, 27 (20)	1-31 (31)	1-28, 32, 35, 37 (31)
82.5	1-22 (22)	1-17, 19, 21, 25-27 (22)	1-31, 33, 34 (33)	1-28, 32, 34-37 (33)
85.0	1-24 (24)	1-17, 19, 21, 22, 24-27 (24)	1-31, 33-35, 38 (35)	1-28, 32-38 (35)
87.5	1-25, 27 (26)	1-21, 23-27 (26)	1-27, 31, 33-38, 40, 42, 45 (37)	1-19, 21-30, 32-38, 48 (37)
90.0	1-28 (28)	1-28 (28)	1-38 (38)	1-38 (38)
92.5	1-30 (30)	1-28, 32, 35 (30)	1-40 (40)	1-39, 48 (40)
95.0	1-31 (31)	1-19, 21-27, 32-35, 37, 38 (32)	1-42 (42)	1-39, 48, 49, 58 (42)
97.5	1-31, 33, 34 (33)	1-28, 32-35, 38 (33)	1-44 (44)	1-39, 45, 48, 49, 55, 56 (44)
100.0	1-31, 33-35, 38 (35)	1-28, 32-38 (35)	1-46 (46)	1-39, 43-45, 48, 49, 55, 56 (46)
102.5	1-36, 38 (37)	1-30, 32-38 (37)	1-48 (48)	1-39, 43-49, 55, 56 (48)
105.0	1-39 (39)	1-39 (39)	1-49, 52 (50)	1-40, 42-49, 57, 58 (50)
101.5	NY panel (budget applies to one-tier process): 1-20, 22, 25-28, 32-34, 37, 38, 45, 48, 56, 65, 83-85 (37)[a]			

[a] The NY panel actually has 39 variants, but two are not included in the CFTR2 data set as CF-causing variants, because they were later found to be not CF-causing, and thus are excluded from our analysis (as was also done in Hughes et al. (2015), which re-examines the dried blood spots of 439 CF-positive newborns found in NY's CF NBS from 2002 to 2012).

prevalent variants, except for special cases (see Lemma 1). While all variants have the same variable cost (c=$1.26 per variant), variants with higher prevalences reduce the FN probability more and incur higher downstream testing costs (GS and/or SC testing), because of a higher detection rate for both CF-positive cases (which is the goal of NBS) and CF carriers (which is not the goal of NBS, and which unnecessarily consumes testing resources).

- One-tier versus two-tier panels: The two-tier process has larger panels than the one-tier process at each budget, because the use of GS in the two-tier process (after MP, Fig. 1) serves to reduce the referral rate for the high cost SC. This, in turn, allows more variants to be included in the two-tier panel, potentially reducing the FN rate.
- **DM** versus **RM** panels: **DM** and **RM** share the same panel only in four of the 17 budgets considered for one-tier (for $70, $72.5, $90, and $105), and only once for two-tier (for $90). While both **DM** and **RM** panels can deviate from the set of variants with the highest prevalences, **RM** is also sensitive to the prevalence bounds (see Lemma 1). To illustrate, consider Variant #9, which, for budgets of $65 and $67.5, is included in the one-tier **DM** panels, but not in the one-tier **RM** panels. Variant #9 is ranked 33rd in group W, 9th in group H, and 2nd in group B (Table 3), and because of this, the range for Variant #9 is wider than the ranges for Variants #8 and #10 (both of which are included in both **DM** and **RM** panels), with a lower bound that is only 61.1% and 69.8%, respectively, of the other two variants. Thus, **RM** excludes Variant #9 because of its low lower bound; this is in line with Lemma 1, which indicates that **RM** has a tendency to choose variants with higher lower bounds.

- Comparison with the NY panel: NY used a one-tier process and a 37-variant panel during the study period (Kay et al. (2015)). The variants in the NY panel do not perfectly coincide with the 37 highest prevalence variants from our data, e.g., the NY panel includes Variants #83-85 (using our ranking), which were not found in any CF-positive newborn in NY between 2007 and 2012 (Kay et al. (2015)), and hence, are not used in any of our optimal panels. Given our cost structure, the NY panel requires a budget of \$101.5; for a slightly lower \$100 budget, both the one-tier **DM** and **RM** use smaller panels with 35 variants, while the two-tier **DM** and **RM** each contain 46 variants; for both one-tier and two-tier processes, the **DM** and **RM** panels are not identical in composition.

4.2.2 Simulation Analysis

To compare the performance of the panels in Table 5, we perform a Monte Carlo simulation and study the realistic situation in which each group's variant prevalence vector realization may deviate from the point estimate vector, $\hat{\boldsymbol{\mu}}^r, r \in R$, which we set as the expectation of each group's prevalence vector. This allows for uncertainty, but without any errors in the point estimate vector. Thus, the simulation requires prevalence distributions to sample from. Based on the prevalence data in the case study, almost all bounds, $[q_i^{r,LB}, q_i^{r,UB}], i \in \Omega, r \in R$, are asymmetric around their mean ($\hat{\mu}_i^r$). This is because many prevalences have small means and large variances, and thus their lower bounds are very close to zero. Therefore, we use the triangular distribution for each $Q_i^r, i \in \Omega, r \in R$, which can model the asymmetry of the bounds around the mean, and the implied right skewness of the unknown distribution, and is a good modeling choice in the absence of sufficient data to fit a distribution (e.g., Yoe (2019)). We utilize group-level prevalence bounds and first moment estimates (Section 4.1) to calibrate the parameters (*min*, *mode*, and *max*) of each triangular distribution.

Specifically, we analyze the performance of each panel for a common set of 2,000 replications, where each replication corresponds to a population of 100,000 newborns, generated as follows. For each replication, we randomly generate a realization of each group's prevalence vector, $\boldsymbol{\theta}^r, r \in R$, where each $Q_i^r, i \in \Omega, r \in R$, follows a triangular distribution (with parameters, $min_i^r = q_i^{r,LB}, max_i^r = \min\left\{q_i^{r,UB}, 3\hat{\mu}_i^r - 2q_i^{r,LB}\right\}$, and $mode_i^r = max\left\{3\hat{\mu}_i^r - q_i^{r,LB} - q_i^{r,UB}, q_i^{r,LB}\right\}$, i.e., the distribution parameters are calibrated so that its mean matches $\hat{\mu}_i^r$).[3] Due to the correlated structure of the random vector (i.e., $\sum_{i \in \Omega \cup \{0\}} Q_i^r = 1, r \in R$), we use the acceptance-rejection method (Devroye (1986); Martino et al. (2018)) in our sampling. Specifically, for each group $r \in R$, we generate each variant's prevalence $\theta_i^r, i \in \Omega$, from its corresponding triangular distribution, and compute the mutation-free proportion, $\theta_0^r = 1 - \sum_{i \in \Omega} \theta_i^r$: if $\theta_0^r \in [q_0^{r,LB}, q_0^{r,UB}]$, then the sampling for group r is accepted, and otherwise, it is rejected and the process is repeated until an acceptable vector realization for each group is generated. For each newborn, we randomly generate two parents, such that each parent belongs to group $r, r \in R$, with probability f^r, independently

[3] For the case where the calibrated *mode* turns out to be lower than the *min* parameter (i.e., the case where $3\hat{\mu}_i^r - q_i^{r,LB} - q_i^{r,UB} < q_i^{r,LB}$, which happens when the range ($max - min$) is too wide to fit a triangular distribution with the same mean), the mode is adjusted (increased) to equal the lower bound (i.e., $mode_i^r = q_i^{r,LB}$), and the upper bound is adjusted (reduced), i.e., $max_i^r = 3\hat{\mu}_i^r - 2q_i^{r,LB}$, in order to maintain the mean, $\hat{\mu}_i^r$.

of the other parent. For each parent, we randomly generate either zero or one mutation variant according to their group's prevalence vector realization, $\boldsymbol{\theta}^r$. Given the parents' mutational status, we then randomly generate the mutational status of the newborn, i.e., if a parent is mutation-free, then the newborn inherits a mutation-free gene from that parent; and if the parent has a variant, the newborn inherits it with probability 0.5, and inherits a mutation-free gene with probability 0.5. Then, the newborn's number of mutations and types of variants are automatically determined by the pair of genes inherited from the parents. Then, for each panel in Table 5, we determine the corresponding number of FNs and the testing cost per 100,000 newborns in each replication. Based on the simulation results depicted in Figs. 3–9, we provide a comparison of the one-tier and two-tier processes, followed by a comparison of the **DM, RM,** and the NY panels. When discussing the simulation results, the worst-case FN rate refers to the maximum FN rate (over 2,000 replications), and both the FN rate and the worst-case FN rate are per 100,000 newborns.

One-tier versus two-tier processes: For illustrative purposes, we use **DM** for this comparison (the results for **RM** are similar). Based on Theorem 4 and case study parameters, the two-tier process is expected to outperform the one-tier process (i.e., $\frac{C_{SC}}{C_{GS}} = 1.7 \geq \frac{2}{1+\hat{\mu}_0} = 1.07$, satisfying the condition in the first part of Theorem 4), which is indeed the case for the simulation results.

- Due to larger panels, the two-tier process incurs, on average, both a lower FN rate and a lower worst-case FN rate at every budget (Table 5). As budgets increase, the two-tier process remains superior, but the benefits decrease as both panels include more of the higher prevalence variants. For example, at the \$65 budget, the two-tier process uses nine more variants, leading to a 45.6% reduction in the FN rate (from 11.89 to 6.46), and a 36.0% reduction in the worst-case FN rate (from 25 to 16); and at the \$105 budget, the two-tier process uses 11 more variants, leading to a 18.1% reduction in the FN rate (from 2.57 to 2.11), and a 10.0% reduction in the worst-case FN rate (from 10 to 9).

- SC referrals for the two-tier process are, on average, only 3.6% of the referrals for the one-tier process (Fig. 5). This is because the GS test is performed for all newborns with at least one mutation detected in MP (Fig. 1), thus eliminating many carriers from SC testing. This is desirable, and decreases the parental burden associated with SC testing. Therefore, while GS testing is expensive, it increases the accuracy of the genetic screening process for a given budget, and this is one of the reasons the two-tier process performs so well, especially when the cost of GS is sufficiently low compared to the cost of SC testing (Theorem 4).

Figure 3: Average False-negatives per 100,000 Newborns for One-tier and Two-tier **DM**

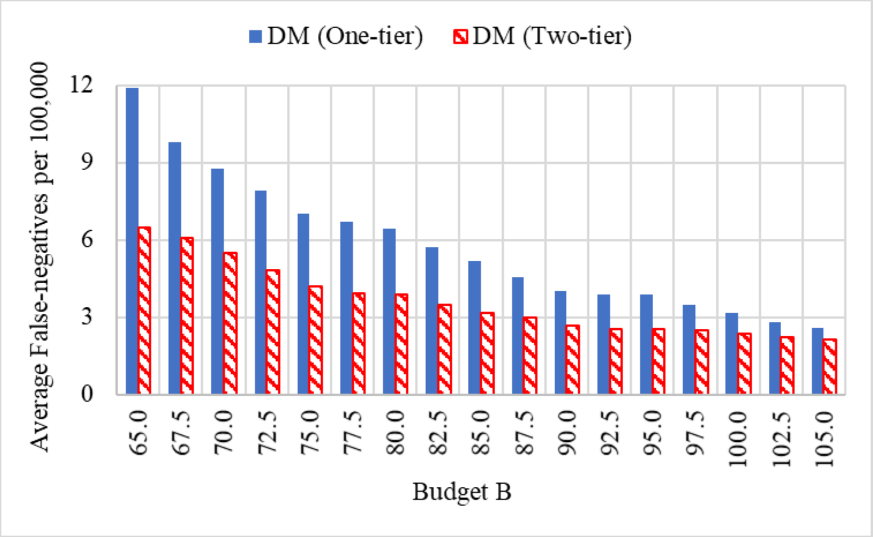
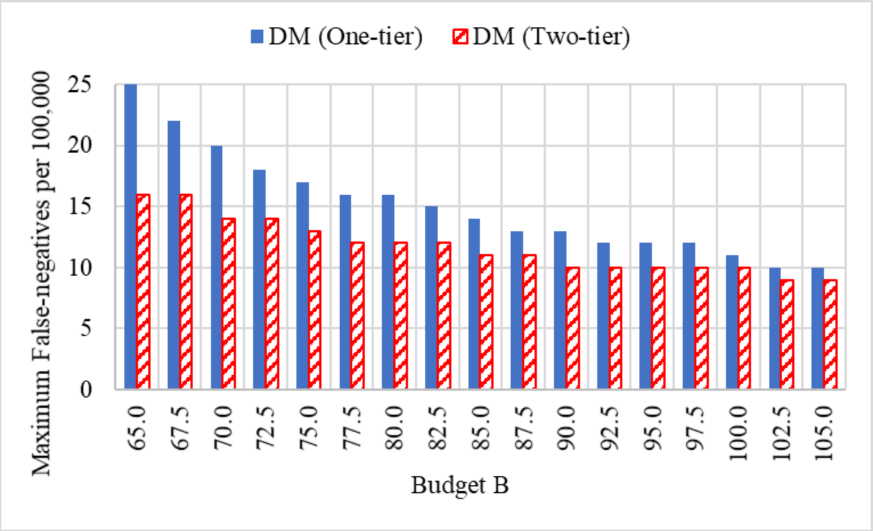


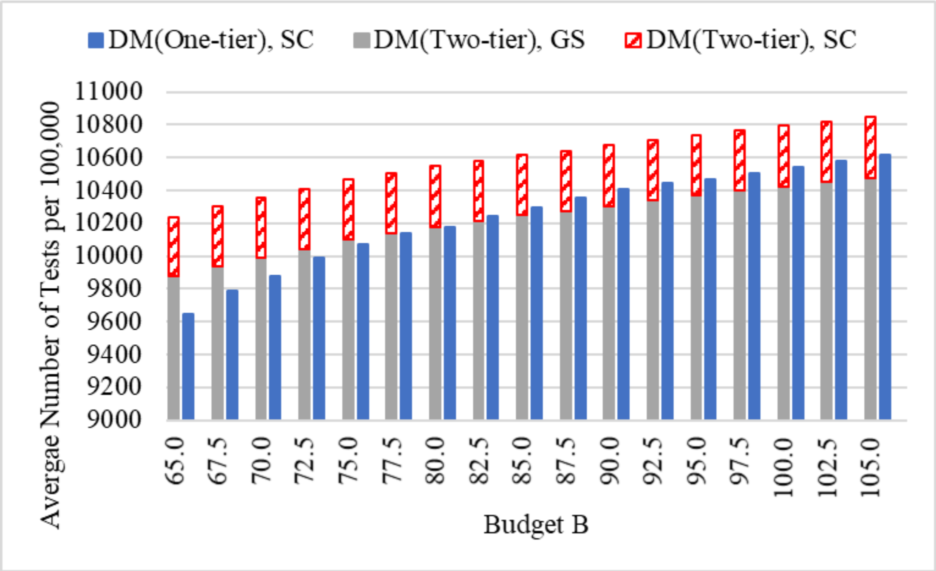
Figure 4: Maximum False-negatives per 100,000 Newborns for One-tier and Two-tier **DM**



DM versus RM: **DM** minimizes the FN probability using a point estimate of the prevalence vector (which equals the true mean prevalence vector, $\hat{\boldsymbol{\mu}}$, used in the simulation), whereas **RM** minimizes the worst-case FN probability for all prevalences that fall in the uncertainty set, $S(\boldsymbol{Q})$. To compare these models, we focus on the one-tier process (the two-tier results are similar).

If the prevalence vector realization corresponds to the point estimate vector used in **DM** (i.e., $\hat{\boldsymbol{\mu}}$), then **RM** incurs, on average, a 0.69% increase in the FN rate over all budgets, compared to **DM**. Table 6 reports the price of robustness ratio and its corresponding upper bound from Theorem 5 for this baseline scenario (which are calculated using the optimal **DM** solution for different budgets, assuming that $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$). While in this setting the price of robustness is fairly small (i.e., the ratio is close to one), in general **RM** can potentially have fewer FNs than **DM**,

Figure 5: Average Number of Sweat Chloride (SC) and Genetic Sequencing (GS) Tests per 100,000 Newborns for One-tier and Two-tier **DM**



that is, when the prevalence realization deviates from the point estimate, which we explore via simulation, leading to the following findings:

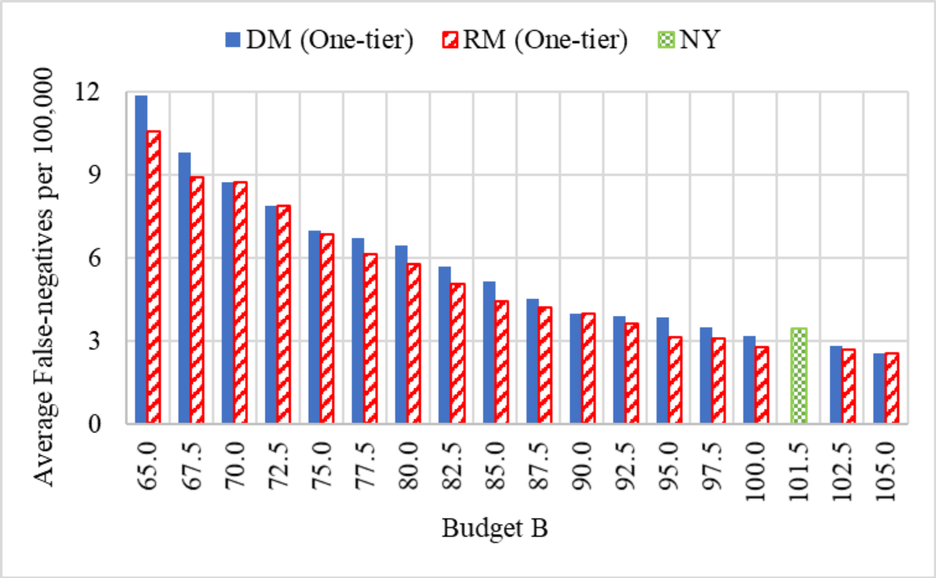
- Over all budgets, **RM** reduces the average FN rate by 7.3% (from 5.75 to 5.33) and the average worst-case FN rate by 3.5% (from 15.06 to 14.53), compared to **DM** (Figs. 6-7). For the FN rate, the maximum reduction is 18.6% at the \$95 budget, and the minimum reduction is 0%, which occurs for the four budgets where **DM** and **RM** panels are identical. When compared to **DM** over all budgets, the FN rate of **RM** is lower 32.5% of the time, equal 60.7% of the time, and higher only 6.8% of the time (Fig. 8). For illustration, consider the \$65 budget (Fig. 9), for which **RM** dominates **DM**, having a lower or equal FN rate 71.1% and 24.9% of the time, respectively, and with an average 11.0% reduction in the FN rate. To test the sensitivity of these results to the triangular distribution assumption, we also conduct a simulation study under the uniform distribution assumption for each $Q_i^r, i \in \Omega, r \in R$, with parameters, $min_i^r = q_i^{r,LB}$ and $max_i^r = q_i^{r,UB}$, and using the acceptance-rejection sampling method, and observe similar results. These results indicate that **RM** is better equipped to reduce FNs in general, and does so without requiring any distribution information on prevalences.

Comparison with the NY panel: The NY panel (with a corresponding cost of \$101.5) is dominated by both one-tier **DM** and **RM** at their lower budget of \$100, despite their smaller panels (35 variants). In particular, both one-tier **DM** and **RM** reduce the average FN rate by 5.7% and 17.0%, respectively, compared to the NY panel (Figs. 6-7). The two-tier **DM** and **RM** (at the same \$100 budget) use larger panels (46 variants), and reduce the average FN rate by 29.5% and 35.2%, respectively, compared to the one-tier NY panel. One-tier and two-tier **DM** and **RM** also

Table 6: Price of Robustness Ratio ($\Pi(\hat{\boldsymbol{\mu}})$) and its Upper Bound from Theorem 5 for Various Budgets

Budget ($)	$\Pi(\hat{\boldsymbol{\mu}})$ (Upper Bound)	
	One-tier	Two-tier
65	1.0019 (2.2710)	1.0117 (3.1460)
70	1.0000 (2.5874)	1.0228 (3.5192)
80	1.0186 (3.2487)	1.0064 (4.2918)
90	1.0000 (3.9859)	1.0000 (4.7868)
100	1.0067 (4.6643)	1.0306 (6.0554)

Figure 6: Average False-negatives per 100,000 Newborns for One-tier **DM**, **RM**, and the NY Panel



reduce the worst-case FN rate (11 for both one-tier **DM** and **RM**, and 10 and 9 for the two-tier **DM** and **RM**, compared to 14 for NY). These results highlight the value of an optimization approach to MP panel design.

To put these results in perspective, consider that during the six-year period between 2007 and 2012, 79,973 newborns underwent post-IRT genetic testing in NY. Based on our prevalence vector ($\hat{\boldsymbol{\mu}}$) and simulation results, the NY panel led to 2.76 FNs for this cohort (equivalently, 3.45 FNs per 100,000) under the one-tier process. Compared to this, **RM**, limited by a budget of $100, not only led to 2.23 FNs per 79,973 (2.78 per 100,000), but also a testing cost reduction of $181,844 under the one-tier process; and to 1.73 FNs per 79,973 (2.16 per 100,000), accompanied by a testing cost reduction of $173,608 under the two-tier process in this six-year period. These results demonstrate the effectiveness of our models, especially when using the two-tier process. NY is adopting the two-tier process (Wadsworth (2017)), which is in line with our findings.

Figure 7: Maximum False-negatives per 100,000 Newborns for One-tier **DM**, **RM**, and the NY Panel

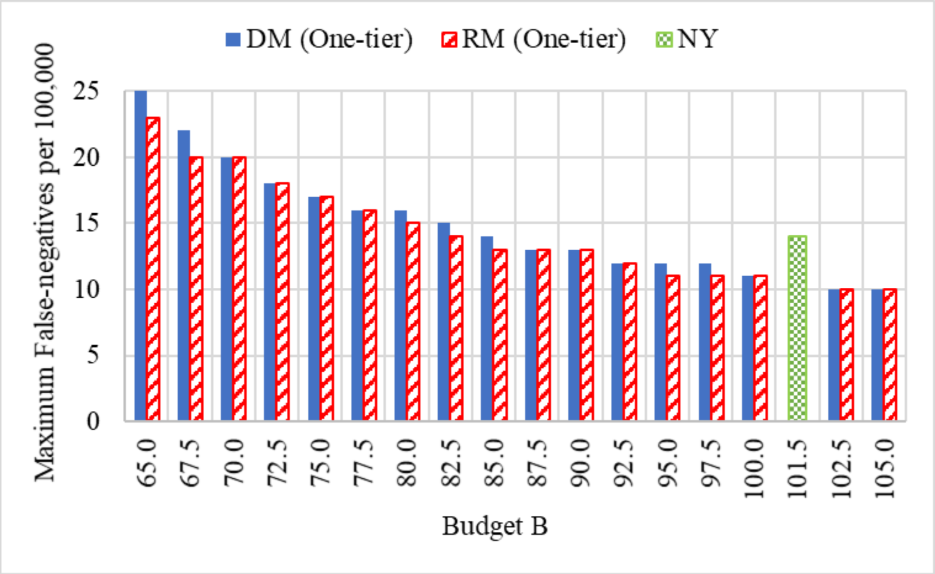
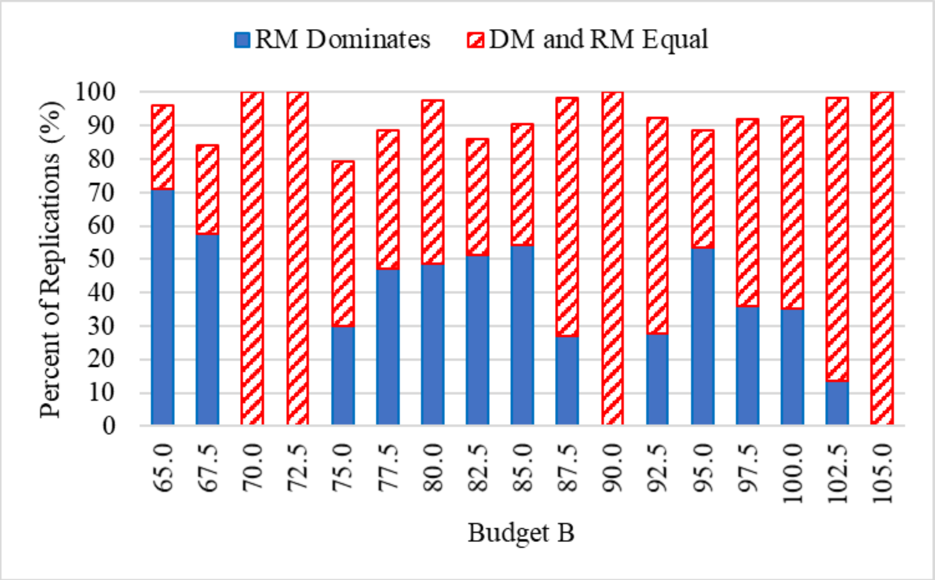


Figure 8: Percent of Replications in which the One-tier **RM** Dominates, or is Equal to, the One-tier **DM**, in Terms of False-negatives per 100,000 Newborns

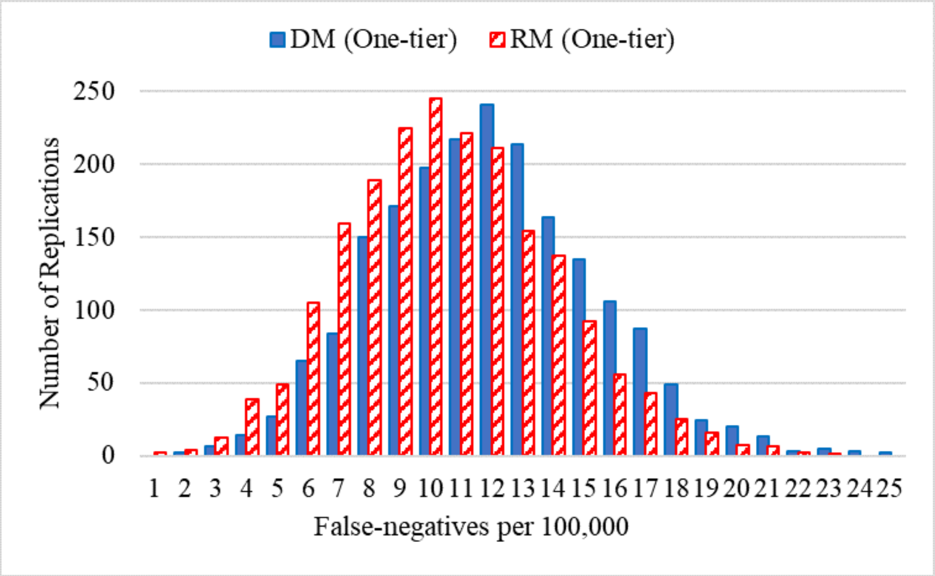


4.2.3 Robust Model Variations

We use Monte Carlo simulation to illustrate the performance of **RM**, **C-RM** and **U-RM** (Remark 1 and Appendix B), as well as the deterministic model, **DM**, and provide some results for a budget of \$65. For each replication, we compute the average testing cost per newborn and report the percent of replications in which this cost exceeds the per newborn budget of \$65 (Table 7).

Our results indicate that the correlated **C-RM** yields the lowest FN rate (8.88) but this comes

Figure 9: Histogram of the False-negatives per 100,000 Newborns for One-tier **DM** and **RM** for a Budget of $B = \$65$



at the expense of a higher per newborn testing cost (an average of \$68.03, over all replications), and exceeds the per newborn budget in every replication. On the other hand, **U-RM** is the most conservative in terms of the testing cost (due to its worst-case cost constraint), as a result, its testing cost (\$61.45 on average) is always below the per newborn budget, but this model yields the highest FN rate (13.55). Finally, **RM** incurs a testing cost (\$64.61 on average) that exceeds the budget in only 0.1% of the replications, and incurs an FN rate of 10.58. We also note that **RM**, which shares the feasible region of **DM**, leads to both a lower FN rate and a lower testing cost than **DM**, underscoring the value of robust optimization in our setting. These results provide insight on the trade-off between the FN probability and the testing cost, and should assist the decision-maker in model selection.

Table 7: Average and Maximum False-negatives per 100,000 Newborns, Average and Maximum Testing Cost per Newborn, and Percent of Replications in which the Average Testing Cost is Higher than Budget, for a Budget of $B = \$65$ (One-tier).

Model	Average FN Rate (per 100,000)	Maximum FN Rate (per 100,000)	Average Testing Cost (per Newborn)	Maximum Testing Cost (per Newborn)	Over-budget Percent of Replications
DM	11.89	25	64.62	65.02	0.05%
RM	10.58	23	64.61	65.09	0.10%
C-RM	8.88	20	67.55	68.03	100%
U-RM	13.55	28	61.45	61.89	0%

5 Conclusions and Future Work

In this paper, we study the optimal design of the genetic portion of CF NBS processes (see Fig. 1). Specifically, we develop and study two distribution-free models, a robust optimization model and a deterministic optimization model, which can be used to design optimal MP panels for the one-tier and two-tier processes, considering the trade-off between the false-negative and false-positive classification rates, and the state's unique characteristics (e.g., population demographics, variant prevalences) and constraints (e.g., testing resources). Further, these models can also be used to redesign the screening process as genetic testing costs change with evolving technology. We provide design insight, models, and algorithms to help practitioners develop optimal screening processes. To our knowledge, this research is the first to use optimization models and methodologies to study the design of CF screening processes. Most of the literature that examines screening processes is in the medical realm, and mainly uses historical data sets and performs descriptive analysis on existing screening processes to determine which process performs best on the given data set. These methods have many shortcomings, the limited data do not provide high levels of confidence in the results, and there is no guarantee that the current screening processes are even near-optimal. Therefore, we hope that our work will motivate further research in this important area, especially as more data on CF-causing variants is collected, which is an important input to our models.

Cystic fibrosis is one of many genetic disorders for which newborns are screened. Newborn screening is performed at the state level, and states choose both the disorders to screen for, and the screening process to use. As such, the disorders screened for, and the processes used, can, and do, differ by state. Our ultimate research goal is to produce decision support models that can help states design optimal and robust screening processes, given each state's unique characteristics and constraints. Having efficient processes that are cost-effective is essential for screening, as this is one of the deciding factors for whether to screen for a disease or not.

Designing optimal screening panels and processes for genetic disorders is timely and important, because genetic testing technologies are improving and becoming less costly, and treatment options for genetic disorders is expanding, thus increasing the number of disorders for which screening is appropriate. Further, for CF and other disorders, identifying the specific genetic variant(s) causing the disorder is becoming more relevant. Currently there are generic CF treatments that are highly beneficial, but customized treatment options that target the underlying defect are becoming available. For example, CF-causing mutation variants have been classified into six types, based on the mechanism through which they compromise the CFTR protein (resulting in CF). Thus there is a strong genotypic/phenotypic correlation in this gene. For example, the use of Ivacaftor in patients having some class III mutation variants is now FDA-licensed (FDA (2018)), and studies show substantial clinical benefit of these types of customized therapies. Further, there is evidence that some therapies can reduce the loss of pancreatic function, which can occur just months after birth (Bell et al. (2015)). Thus early diagnosis, and an understanding of the underlying genetic defect, are likely to become more important in the treatment of this disease, especially as new pharmacological and genetic therapies are discovered, placing this type of research on the forefront of the new wave of personalized, precision medicine.

Future work in this area includes expanding our models to study different methods for constructing the uncertainty set and sampling under the correlated structure of the uncertainty set; and to consider genetic tests that may have less than perfect accuracy (sensitivity and/or specificity) for other genetic disorders, CF variants that may have varying, or uncertain, consequences, and some measures of equity in the selection of variants, as the prevalence of certain variants can vary among different ethnic/racial groups. It is also an important research direction to construct mathematical models to examine, and optimize, the entire screening process, which includes the IRT test (the IRT decision rules can influence the testing budget available for genetic testing, the overall misclassification rates, and the post-IRT prevalences). As another exciting research direction, one can integrate these optimization models with data analytics methodologies to consider subject risk factors for genetic disorders, so that risk-based newborn screening schemes can be developed.

Acknowledgments: We are grateful to Professor Özlem Ergün, the Associate Editor, and three anonymous Reviewers for excellent suggestions that greatly improved the analysis and presentation of the paper. 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.

Biographies:

Hussein El Hajj is a Ph.D candidate in the Grado Department of Industrial and Systems Engineering at Virginia Tech. His research interests lie in the application of operations research and data analytics to solve problems related to healthcare systems and public policy. He worked with Ebru Bish and Douglas Bish, and this paper is part of his Ph.D. dissertation. He is a finalist for the 2020 INFORMS Bonder Scholarship for Applied Operations Research in Health Services, and was a runner-up for the 2019 INFORMS Health Application Society Student Paper Competition.

Douglas Bish is a professor in the Department of Information Systems, Statistics and Management Science in the Culverhouse College of Business at the University of Alabama. His research interests are in applying operations research methodologies and analytics to problems in healthcare, emergency management, and supply chains. This paper is part of a project funded by the National Science Foundation (NSF) that studies the design of pooled screening strategies in public health applications. He is a prior recipient of the NSF CAREER award and the INFORMS Pierskalla Award for the best paper in healthcare.

Ebru Bish is a professor of operations management in the Culverhouse College of Business at the University of Alabama. Her research focus is on decision-making under uncertainty using stochastic modeling, robust and data-driven optimization, and data analytics, with application to public health policy and healthcare system design. This paper is part of a broader body of work funded by the National Science Foundation that focuses on the design of pooled screening schemes for disease biomarkers. She is a recipient of the INFORMS Pierskalla Award for the best paper in healthcare, and she has served as the President of the INFORMS Health Applications Society.

References

Accurso, F.J., Sontag, M.K., and Wagener, J.S. (2005). "Complications associated with symptomatic diagnosis in infants with cystic fibrosis." *The Journal of Pediatrics*, **(3 Suppl)**, pp. S37–S41.

ACHDNC (2018). "Advisory Committee on Heritable Disorders in Newborns and Children - Recommended Uniform Screening Panel Core Conditions." *https://www.hrsa.gov/sites/default/files/hrsa/advisory-committees/heritable-disorders/rusp/rusp-uniform-screening-panel.pdf*, Accessed in November 2019.

Adams, W.P. and Sherali, H.D. (1986). "A tight linearization and an algorithm for zero-one quadratic programming problems." *Management Science*, **32 (10)**, pp. 1274–1290.

Aprahamian, H., Bish, D.R., and Bish, E.K. (2019). "Optimal risk-based group testing." *Management Science*, **65 (9)**, pp. 4365–4384.

Aprahamian, H., Bish, E.K., and Bish, D.R. (2018). "Adaptive risk-based pooling in public health screening." *IISE Transactions*, **50 (9)**, pp. 753–766.

Baby's First Test (2019). *http://www.babysfirsttest.org/newborn-screening/states*, Accessed in November 2019.

Bell, S.C., Boeck, K.D., and Amaral, M.D. (2015). "New pharmacological approaches for cystic fibrosis: Promises, progress, pitfalls." *Pharmacology & Therapeutics*, **145**, pp. 19 – 34.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press. Princeton, New Jersey, USA.

Ben-Tal, A. and Nemirovski, A. (2000). "Robust solutions of linear programming problems contaminated with uncertain data." *Mathematical Programming*, **88 (3)**, pp. 411–424.

Bertsimas, D., Brown, D.B., and Caramanis, C. (2011). "Theory and applications of robust optimization." *Society for Industrial and Applied Mathematics*, **5 (1)**, pp. 15–20.

Bertsimas, D. and Sim, M. (2004). "The price of robustness." *Operations Research*, **52 (1)**, pp. 35–53.

Bish, D.R., Bish, E.K., Xie, R.S., and Stramer, S.L. (2014). "Going beyond "same-for-all" testing of infectious agents in donated blood." *IIE Transactions*, **46 (11)**, pp. 1147–1168.

Borowitz, D., Parad, R.B., Sharp, J.K., Sabadosa, K.A., Robinson, K.A., Rock, M.J., Farrell, P.M., Sontag, M.K., Rosenfeld, M., Davis, S.D., et al. (2009). "Cystic Fibrosis Foundation practice guidelines for the management of infants with cystic fibrosis transmembrane conductance regulator-related metabolic syndrome during the first two years of life and beyond." *The Journal of Pediatrics*, **155 (6)**, pp. S106–S116.

Boyle, M.P. and Boeck, K.D. (2013). "A new era in the treatment of cystic fibrosis: correction of the underlying CFTR defect." *The Lancet Respiratory Medicine*, **1 (2)**, pp. 158 – 163.

Campbell, P.W. and White, T.B. (2005). "Newborn screening for cystic fibrosis: an opportunity to improve care and outcomes." *The Journal of Pediatrics*, **(3 Suppl)**, pp. S2–S5.

Capara, A., Kellerer, H., Pferschy, U., and Pisinger, D. (2000). "Approximation algorithms for knapsack problems with cardinality constraints." *European Journal of Operational Research*, **123 (2)**, pp. 333–345.

CDPH (2019). “California Department of Public Health: Genetic disease screening program.” *https://www.cdph.ca.gov/Programs/CFH/DGDS/CDPH%20Document%20Library/FY%202019-20%20GDSP%20November%20Estimate%20%201.9.19%20DOF%20Final.pdf*, Accessed in November 2019.

Center for Disease Control and Prevention (CDC) (2018). “Birth Data.” *https://www.cdc.gov/nchs/nvss/births.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fnchs%2Fbirths.htm*, Accessed in September 2020.

CFTR2 (2018). *https://www.cftr2.org/*, Accessed in May 2018.

Chinneck, J. and Ramadan, K. (2000). “Linear programming with interval coefficients.” *Journal of the Operational Research Society*, **51 (2)**, pp. 209–220.

Cochran, A.L., Tarini, B.A., Kleyn, M., and Zayas-Cabán, G. (2018). “Newborn screening collection and delivery processes in Michigan Birthing Hospitals: strategies to improve timeliness.” *Maternal and Child Health Journal*, **22 (10)**, pp. 1436–1443.

Cunningham, J. and Taussig, L. (2013). *An Introduction to Cystic Fibrosis for Patients and Their Families.* Cystic Fibrosis Foundation, 6th edition.

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)**, pp. 1159–1163.

Cystic Fibrosis Foundation (2009). “All fifty states to screen newborns for cystic fibrosis by 2010.” *https://www.cff.org/About-Us/Media-Center/Press-Releases/All-Fifty-States-to-Screen-Newborns-for-Cystic-Fibrosis-by-2010/*, Accessed in May 2018.

Cystic Fibrosis Foundation (2017). “Sweat Test.” *https://www.cff.org/What-is-CF/Testing/Sweat-Test/*, Accessed in November 2019.

Cystic Fibrosis Foundation (2018). “2018 patient registry annual data report.” *https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2018-Patient-Registry-Annual-Data-Report.pdf*, Accessed in September 2020.

Cystic Fibrosis Foundation (2019). “Carrier testing for cystic fibrosis.” *https://www.cff.org/What-is-CF/Testing/Carrier-Testing-for-Cystic-Fibrosis/*, Accessed in November 2019.

Deignan, J.L., Astbury, C., and Cutting, G.R.e.a. (2020). “CFTR variant testing: a technical standard of the American College of Medical Genetics and Genomics (ACMG).” *Genetics in Medicine*, **22 (5)**, p. 1288–1295.

Department of Health, Wadsworth Center (2017). “Changes to the New York newborn screening program’s cystic fibrosis screening protocol.” *https://www.wadsworth.org/sites/default/files/WebDoc/26704185/CF%20Algorithm%20Change%20Notification%20Effect 1-17.pdf*, Accessed in December 2019.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* Springer-Verlag, Montreal, Canada.

Dorfman, R. (1943). “The detection of defective members of large populations.” *The Annals of Mathematical Statistics*, **14 (4)**, pp. 436–440.

El Amine, H., Bish, E.K., and Bish, D.R. (2018). “Robust post-donation blood screening under prevalence rate uncertainty.” *Operations Research*, **66 (1)**, pp. 1–17.

Farrell, P.M., Lai, H.J., Li, Z., Kosorok, M.R., Laxova, A., and Green, C.G. (2005). "Evidence on improved outcomes with early diagnosis of cystic fibrosis through neonatal screening: enough is enough!" *Pediatrics*, **147 (3, Suppl)**, pp. S30–S36.

FDA (2018). "Approved drug products." *https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/UCM071436.p* Accessed in December 2019.

Grosse, S.D. (2015). "Showing value in newborn screening: Challenges in quantifying the effectiveness and cost-effectiveness of early detection of phenylketonuria and cystic fibrosis." *Healthcare*, **3 (4)**, pp. 1133–1157.

Grosse, S.D., Boyle, C.A., Botkin, J.R., Comeau, A.M., Kharrazi, M., Rosenfeld, M., and Wilfond, B.S. (2004). "Newborn screening for cystic fibrosis: evaluation of benefits and risks and recommendations for state newborn screening programs." *MMWR: Morbidity & Mortality Weekly Report*, **53 (40)**, pp. 1–36.

Gurobi Optimization (2019). "Gurobi optimizer reference manual." *https://www.gurobi.com/documentation/9.0/refman/index.html*, Accessed in December 2019.

Hughes, E.E., Stevens, C.F., Saavedra-Matiz, C.A., Tavakoli, N.P., Krein, L.M., Parker, A., Zhang, Z., Maloney, B., Vogel, B., DeCelie-Germana, J., et al. (2015). "Clinical sensitivity of cystic fibrosis mutation panels in a diverse population." *Human Mutation*, **37 (2)**, pp. 201–208.

Hull, S.C. and Kass, N.E. (2000). "Adults with cystic fibrosis and (in) fertility: How has the health care system responded?" *Journal of Andrology*, **21 (6)**, p. 809.

Hwang, F.K. (1975). "A generalized binomial group testing problem." *Journal of the American Statistical Association*, **70 (352)**, pp. 923–926.

Illumina (2019). "MiSeqDx cystic fibrosis clinical sequencing assay." *https://www.illumina.com/products/by-type/ivd-products/miseqdx-cystic-fibrosis-clinical-sequencing-assay.html*, Accessed in November 2019.

John Hopkins Medicine (2019). "Cystic Fibrosis and CF-Related Disorders NGS Panel." *https://www.hopkinsmedicine.org/dnadiagnostic/tests/tests/cystic-fibrosis-and-cf-related-disorders-ngs-panel*, Accessed in June 2019.

Johnson, M.A., Yoshitomi, M.J., and Richards, C.S. (2007). "A comparative study of five technologically diverse CFTR testing platforms." *Journal of Molecular Diagnostics*, **9 (3)**, pp. 401–407.

Jorde, L.B., Carey, J.C., and Bamshad, M.J. (2015). *Medical Genetics e-Book.* Elsevier Health Sciences, 5th edition. Philadelphia, Pennsylvania, USA.

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)**, pp. 557–562.

Kay, D.M., Maloney, B., Hamel, R., Pearce, M., DeMartino, L., McMahon, R., McGrath, E., Krein, L., Vogel, B., Saavedra-Matiz, C.A., Caggana, M., and Tavakoli, N.P. (2015). "Screening for cystic fibrosis in New York state: considerations for algorithm improvements." *European Journal of Pediatrics*, **175 (2)**, pp. 181–193.

Kellerer, H., Pferschy, U., and Pisinger, D. (2004). *Knapsack Problems.* Springer, Berlin. Verlag Berlin Heidelberg, New York, USA.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L. (1989). "Identification of the cystic fibrosis gene: genetic analysis." *Science*, **245 (4922)**, pp. 1073–1080.

Kharrazi, M., Yang, J., Bishop, T., Lessing, S., Young, S., Graham, S., Pearl, M., Chow, H., Ho, T., Currier, R., Gaffneya, L., and Feuchtbaum, L. (2015). "Newborn screening for cystic fibrosis in California." *Pediatrics*, **136 (6)**, pp. 1062–1072.

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.

Kosheleva, K., Faulkner, N., Robinson, K., and Umbarger, M. (2017). "Validation of a novel copy number variant detection algorithm for CFTR from targeted next generation sequencing data." *Fertility and Sterility*, **108 (3)**, p. e269.

Lee, D.S., Rosenberg, M.A., Peterson, A., Makholm, L., Hoffman, G., Laessig, R.H., and Farrell, P.M. (2003). "Analysis of the costs of diagnosing cystic fibrosis with a newborn screening program." *The Journal of Pediatrics*, **142 (6)**, pp. 617–623.

Lim, R.M., Silver, A.J., Silver, M.J., Borroto, C., Spurrier, B., Petrossian, T.C., Larson, J.L., and Silver, L.M. (2016). "Targeted mutation screening panels expose systematic population bias in detection of cystic fibrosis risk." *Genetics in Medicine*, **18 (2)**, p. 174.

Martino, L., Luengo, D., and Míguez, J. (2018). "Accept–Reject Methods." In *Independent Random Sampling Methods*. Springer International Publishing, Switzerland, pp. 65–113.

McMahan, C.S., Tebbs, J.M., and Bilder, C.R. (2012). "Informative Dorfman screening." *Biometrics*, **68 (1)**, pp. 287–296.

Mehta, A. (2007). "A global perspective on newborn screening for cystic fibrosis." *Current Opinion in Pulmonary Medicine*, **13 (6)**, pp. 510–514.

National Institutes of Health (NIH) (2016). "The cost of sequencing a human genome." *https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost*, Accessed in December 2019.

National Institutes of Health (NIH) (2019). "CFTR gene." *https://ghr.nlm.nih.gov/gene/CFTR#*, Accessed in November 2019.

Newcombe, R.G. (1998). "Two-sided confidence intervals for the single proportion: comparison of seven methods." *Statistics in Medicine*, **17 (8)**, pp. 857–872.

Newcombe, R.G. (2003). "Confidence intervals for the mean of a variable taking the values 0, 1 and 2." *Statistics in Medicine*, **22 (17)**, pp. 2737–2750.

Nobibon, F.T. and Leus, R. (2014). "Complexity results and exact algorithms for robust knapsack problems." *Journal of Optimization Theory and Applications*, **161 (2)**, pp. 533–552.

Nshimyumukiza, L., Bois, A., Daigneault, P., Lands, L., Laberge, A.M., Fournier, D., Duplantie, J., Giguere, Y., Gekas, J., Gagné, C., et al. (2014). "Cost effectiveness of newborn screening for cystic fibrosis: a simulation study." *Journal of Cystic Fibrosis*, **13 (3)**, pp. 267–274.

Pagaduan, J.V., Ali, M., Dowlin, M., Suo, L., Ward, T., Ruiz, F., and Devaraj, S. (2018). "Revisiting sweat chloride test results based on recent guidelines for diagnosis of cystic fibrosis." *Practical Laboratory Medicine*, **10**, pp. 34–37.

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

Pisinger, D. (1998). "A fast algorithm for strongly correlated knapsack problems." *Discrete Applied Mathematics*, **89 (1)**, pp. 197–212.

Rock, M.J., Hoffman, G., Laessig, R.H., Kopish, G.J., Litsheim, T.J., and Farrell, P.M. (2005). "Newborn screening for cystic fibrosis in Wisconsin: nine-year experience with routine trypsinogen/DNA testing." *The Journal of Pediatrics*, **147 (3)**, pp. S73–S77.

Rosenberg, M.A. and Farrell, P.M. (2005). "Assessing the cost of cystic fibrosis diagnosis and treatment." *The Journal of Pediatrics*, **147 (3)**, pp. S101–S105.

Sadeghzadeh, S., Bish, E.K., and Bish, D.R. (2020). "Optimal data-driven policies for disease screening under noisy biomarker measurement." *IISE Transactions*, **52 (2)**, pp. 166–180.

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

Schmidt, M., Werbrouck, A., Verhaeghe, N., De Wachter, E., Simoens, S., Annemans, L., and Putman, K. (2019). "A model-based economic evaluation of four newborn screening strategies for cystic fibrosis in Flanders, Belgium." *Acta Clinica Belgica*, pp. 1–9.

Schrijver, I., Pique, L., Graham, S., Pearl, M., Cherry, A., and Kharrazi, M. (2016). "The spectrum of CFTR variants in nonwhite cystic fibrosis patients: implications for molecular diagnostic testing." *Journal of Molecular Diagnostics*, **18 (1)**, pp. 39–50.

Smith, J.C. and Taskin, Z.C. (2008). "A tutorial guide to mixed-integer programming models and solution techniques." *Optimization in Medicine and Biology*, pp. 521–548.

Soleimanian, A. and Jajaei, G.S. (2013). "Robust nonlinear optimization with conic representable uncertainty set." *European Journal of Operational Research*, **228 (2)**, pp. 337 – 344.

Tluczek, A., Koscik, R.L., Farrell, P.M., and Rock, M.J. (2005). "Psychosocial risk associated with newborn screening for cystic fibrosis: Parents' experience while awaiting the sweat-test appointment." *Pediatrics*, **115 (6)**, pp. 1692–1703.

van den Akker, M.E., Dankert, H.M., Verkerk, P.H., Dankert-Roelse, J.E., et al. (2006). "Cost-effectiveness of 4 neonatal screening strategies for cystic fibrosis." *Pediatrics*, **118 (3)**, pp. 896–905.

van der Ploeg, C., van den Akker, M.E., Vernooij-van Langen, A., Elvers, L., Gille, J., Verkerk, P., Dankert-Roelse, J., study group, C., et al. (2015). "Cost-effectiveness of newborn screening for cystic fibrosis determined with real-life data." *Journal of Cystic Fibrosis*, **14 (2)**, pp. 194–202.

Watson, M.S., Cutting, G.R., Desnick, R.J., Driscoll, D.A., Klinger, K., Mennuti, M., Palomaki, G.E., Popovich, B.W., Pratt, V.M., Rohlfs, E.M., Strom, C.M., Richards, C.S., Witt, D.R., and Grody, W.W. (2004). "Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel." *Genetics in Medicine*, **6 (5)**, pp. 387–391.

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*, **129 (2)**, pp. e339–e347.

Willis, M.S. (2012). "Multiple positive sweat chloride tests in an infant asymptomatic for cystic fibrosis." *Laboratory Medicine*, **43 (2)**, pp. 1–5.

Yoe, C. (2019). *Principles of Risk Analysis: Decision Making Under Uncertainty.* CRC press. Boca Raton, Florida, USA.

Optimal Genetic Screening for Cystic Fibrosis

Hussein El Hajj[1*], Douglas R. Bish[2], Ebru K. Bish[2]

1. Grado Department of Industrial and Systems Engineering, Virginia Tech
Blacksburg, VA 24061-0118, United States
2. Department of Information Systems, Statistics, and Management Science, University of Alabama
Tuscaloosa, AL 35487, United States

Submitted for publication in Operations Research, the "Policy Modeling and Public Sector OR" area.

[*] Corresponding Author: *hme35@vt.edu*

APPENDIX

A Derivations and Supporting Results

We first provide some derivations and supporting results that will be used in the remainder of the Appendix.

A.1 Probability Mass Functions

We first derive the probability mass functions for random variables N, $\widetilde{N}(\boldsymbol{x})|N$, and $\widetilde{N}(\boldsymbol{x})$. We use the notation | or ; interchangeably to denote probabilistic conditioning.

Recall that each parent provides one CFTR gene, randomly selected among the parent's two CFTR genes, to their newborn; and each parent can have at most one CF-causing mutation (Assumption ($\boldsymbol{A_1}$)). Consequently, if both parents are carriers, then there is a 0.25 probability that the newborn will have CF (i.e., with two mutations), a 0.5 probability that the newborn will be a carrier (i.e., with one mutation), and a 0.25 probability that the newborn will be mutation-free. On the other hand, if one parent is a carrier while the other parent is mutation-free, then there is a 0.5 probability that the newborn will be a carrier, and a 0.5 probability that the newborn will be mutation-free. Let $A_{i,j}$, $i, j \in \Omega \cup \{0\}$, denote the event that the mother of a random newborn has one mutation of variant i, and the father has one mutation of variant j, with $i = 0$ or $j = 0$ if the corresponding parent is mutation-free.

We derive the following expressions for any prevalence vector realization $\boldsymbol{\theta} \in S(\boldsymbol{Q})$; thus, all the following probabilities are conditioned on $\boldsymbol{\theta}$, but we omit the conditioning notation in most parts of this section to simplify the notation. The expressions for any specific prevalence vector, e.g., $\hat{\boldsymbol{\mu}} \in S(\boldsymbol{Q})$, can be simply obtained by replacing $\boldsymbol{\theta}$ with the prevalence vector of interest:

Probability mass function of N:

By the law of total probability, we can write, for n=0,1,2:

$$\begin{aligned} Pr\left(N=n\right) = &\sum_{i\in\Omega\cup\{0\}}\sum_{j\in\Omega\cup\{0\}} Pr\left(N=n|A_{i,j}\right)Pr\left(A_{i,j}\right) \\ =&Pr\left(N=n|A_{0,0}\right)Pr\left(A_{0,0}\right) + \sum_{i\in\Omega}\left[Pr\left(N=n|A_{i,0}\right)Pr\left(A_{i,0}\right) + Pr\left(N=n|A_{0,i}\right)Pr\left(A_{0,i}\right)\right] \\ &+ \sum_{i\in\Omega}\sum_{j\in\Omega} Pr\left(N=n|A_{i,j}\right)Pr\left(A_{i,j}\right), \text{ where} \end{aligned}$$

$$Pr\left(A_{0,0}\right) = {\theta_0}^2,\ Pr\left(A_{0,i}\right) = Pr\left(A_{i,0}\right) = \theta_i\theta_0, \text{ and } Pr\left(A_{i,j}\right) = Pr\left(A_{j,i}\right) = \theta_i\theta_j, \forall i,j\in\Omega \text{ (by Assumption } (\boldsymbol{A_1})).$$

Then, using the equality that $\theta_0 = 1 - \sum_{i\in\Omega}\theta_i$, we derive:

$$\begin{aligned} Pr\left(N=0\right) =&(1){\theta_0}^2 + 2\sum_{i\in\Omega}\left(\frac{1}{2}\right)\theta_i\theta_0 + \sum_{i\in\Omega}\sum_{j\in\Omega}\left(\frac{1}{4}\right)\theta_i\theta_j = {\theta_0}^2 + \theta_0\left(1-\theta_0\right) + \frac{1}{4}\left(1-\theta_0\right)^2 = \frac{1}{4}\left(1+\theta_0\right)^2 \\ Pr\left(N=1\right) =&(0) + 2\sum_{i\in\Omega}\left(\frac{1}{2}\right)\theta_i\theta_0 + \sum_{i\in\Omega}\sum_{j\in\Omega}\left(\frac{1}{2}\right)\theta_i\theta_j = \theta_0\left(1-\theta_0\right) + \frac{1}{2}\left(1-\theta_0\right)^2 = \frac{1}{2}\left(1-{\theta_0}^2\right) \qquad (14) \\ Pr\left(N=2\right) =&(0) + 2(0) + \sum_{i\in\Omega}\sum_{j\in\Omega}\left(\frac{1}{4}\right)\theta_i\theta_j = \frac{1}{4}\left(1-\theta_0\right)^2. \end{aligned}$$

To derive the conditional probability mass function of $\tilde{N}(\boldsymbol{x})|N$, we define Y_i, $i \in \Omega$, as the number of mutations of variant i in a random newborn, with $S\left(Y_i\right) = \{0, 1, 2\}$. First note the following equivalent events:

$\{Y_i = n, N = n\} \equiv \{Y_i = n, Y_k = 0, k \in \Omega\backslash\{i\}\}$, for $i \in \Omega$, $n = 0, 1, 2$, and
$\{Y_i = 1, Y_j = 1, N = 2\} \equiv \{Y_i = 1, Y_j = 1, Y_k = 0, k \in \Omega\backslash\{i, j\}\}$, for $i, j \in \Omega$, $i \neq j$.

Hence, we can write, for $n = 0, 1, 2$:

$$Pr\left(Y_i = n | N = n\right) = \frac{Pr\left(Y_i = n, Y_k = 0, k \in \Omega\backslash\{i\}\right)}{Pr\left(N = n\right)} = \frac{\sum_{s\in\Omega\cup\{0\}}\sum_{j\in\Omega\cup\{0\}} Pr\left(Y_i = n, Y_k = 0, k \in \Omega\backslash\{i\}|A_{s,j}\right) Pr\left(A_{s,j}\right)}{Pr\left(N = n\right)}.$$

Then:

$$\begin{aligned}
Pr\left(Y_i = 0|N = 0\right) =& 1, \forall i \in \Omega \\
Pr\left(Y_i = 1|N = 1\right) =& \frac{2\left(\frac{1}{2}\right)\theta_i\theta_0 + \left(\frac{1}{2}\right)\theta_i^2 + 2\sum_{j\in\Omega\backslash\{i\}}\frac{1}{4}\theta_i\theta_j}{Pr\left(N = 1\right)} = \frac{\theta_i\theta_0 + \frac{1}{2}\theta_i^2 + \frac{1}{2}\theta_i\left(1 - \theta_0 - \theta_i\right)}{Pr\left(N = 1\right)} \\
=& \frac{\frac{1}{2}\theta_i\left(1 + \theta_0\right)}{\frac{1}{2}\left(1 - {\theta_0}^2\right)} = \frac{\theta_i}{\left(1 - \theta_0\right)}, \qquad \forall i \in \Omega \\
Pr\left(Y_i = 2|N = 2\right) =& \frac{\left(\frac{1}{4}\right)\theta_i^2}{\frac{1}{4}\left(1 - \theta_0\right)^2} = \frac{\theta_i^2}{\left(1 - \theta_0\right)^2}, \qquad \forall i \in \Omega \\
Pr\left(Y_i = 1, Y_j = 1|N = 2\right) =& \frac{2Pr\left(Y_i = 1, Y_j = 1, Y_k = 0, k \in \Omega\backslash\{i,j\}|A_{i,j}\right)Pr\left(A_{i,j}\right)}{Pr\left(N = 2\right)} = \frac{2\left(\frac{1}{4}\right)\theta_i\theta_j}{\frac{1}{4}\left(1 - \theta_0\right)^2} \\
=& \frac{2\theta_i\theta_j}{\left(1 - \theta_0\right)^2}, \qquad \forall i, j \in \Omega, i \neq j,
\end{aligned}$$

$$\begin{align}
E[Y_i|N = 2] =& 2Pr\left(Y_i = 2|N = 2\right) + \sum_{j\in\Omega\backslash\{i\}} Pr\left(Y_i = 1, Y_j = 1|N = 2\right) = \frac{2\theta_i^2}{\left(1 - \theta_0\right)^2} + \frac{2\sum_{j\in\Omega\backslash\{i\}}\theta_i\theta_j}{\left(1 - \theta_0\right)^2} \notag \\
=& \frac{2\theta_i^2 + 2\theta_i(1 - \theta_0 - \theta_i)}{\left(1 - \theta_0\right)^2} = \frac{2\theta_i(1 - \theta_0)}{\left(1 - \theta_0\right)^2} = \frac{2\theta_i}{\left(1 - \theta_0\right)}, \qquad \forall i \in \Omega. \tag{15}
\end{align}$$

In the case study of Section 4, the term $E[Y_i|N = 2, \hat{\boldsymbol{\mu}}]$ is estimated as $\frac{\text{number of mutations of variant } i \text{ in CF-positive newborns}}{\text{number of CF-positive newborns}}$, based on the data reported in Schrijver et al. (2016) on variant frequencies in 35,415 CF-positive newborns (i.e., among the newborn population with $\{N = 2\}$).

Conditional probability mass function of $\widetilde{N}(\boldsymbol{x})|N$:

By Assumptions $(\boldsymbol{A_2})$ and $(\boldsymbol{A_3})$, we have that $Pr\left(\widetilde{N}(\boldsymbol{x}) > n|N = n\right) = 0$, $n = 0, 1, 2$, $\forall \boldsymbol{x}$. Then:

$$\begin{align}
Pr\left(\widetilde{N}(\boldsymbol{x}) = 0|N = 0\right) =& 1 \notag \\
Pr\left(\widetilde{N}(\boldsymbol{x}) = 0|N = 1\right) =& \sum_{i\in\Omega}\left(1 - x_i\right)Pr\left(Y_i = 1|N = 1\right) = \frac{\sum_{i\in\Omega}\left(1 - x_i\right)\theta_i}{\left(1 - \theta_0\right)} \notag \\
Pr\left(\widetilde{N}(\boldsymbol{x}) = 1|N = 1\right) =& \sum_{i\in\Omega} x_i Pr\left(Y_i = 1|N = 1\right) = \frac{\sum_{i\in\Omega} x_i\theta_i}{\left(1 - \theta_0\right)} \notag \\
Pr\left(\widetilde{N}(\boldsymbol{x}) = 0|N = 2\right) =& \sum_{i\in\Omega}\left(1 - x_i\right)Pr\left(Y_i = 2|N = 2\right) + \sum_{i\in\Omega}\sum_{j\in\Omega:j\geq i+1}\left(1 - x_i\right)\left(1 - x_j\right)Pr\left(Y_i = 1, Y_j = 1|N = 2\right) \notag \\
=& \frac{\sum_{i\in\Omega}\left(1 - x_i\right)\theta_i^2}{\left(1 - \theta_0\right)^2} + \frac{2\sum_{i\in\Omega}\sum_{j\in\Omega:j\geq i+1}\left(1 - x_i\right)\left(1 - x_j\right)\theta_i\theta_j}{\left(1 - \theta_0\right)^2} \tag{16} \\
=& \frac{\sum_{i\in\Omega}\sum_{j\in\Omega}\left(1 - x_i\right)\left(1 - x_j\right)\theta_i\theta_j}{\left(1 - \theta_0\right)^2} \notag \\
Pr\left(\widetilde{N}(\boldsymbol{x}) = 1|N = 2\right) =& \sum_{i\in\Omega}\sum_{j\in\Omega} x_i\left(1 - x_j\right)Pr\left(Y_i = 1, Y_j = 1|N = 2\right) = \frac{2\sum_{i\in\Omega}\sum_{j\in\Omega} x_i\left(1 - x_j\right)\theta_i\theta_j}{\left(1 - \theta_0\right)^2} \notag
\end{align}$$

$$
\begin{aligned}
Pr\left(\widetilde{N}(\boldsymbol{x})=2|N=2\right)=&\sum_{i\in\Omega}x_iPr\left(Y_i=2|N=2\right)+\sum_{i\in\Omega}\sum_{j\in\Omega:j\geq i+1}x_ix_jPr\left(Y_i=1,Y_j=1|N=2\right)\\
=&\frac{\sum_{i\in\Omega}x_i\theta_i^2}{\left(1-\theta_0\right)^2}+\frac{2\sum_{i\in\Omega}\sum_{j\in\Omega:j\geq i+1}x_ix_j\theta_i\theta_j}{\left(1-\theta_0\right)^2}=\frac{\sum_{i\in\Omega}\sum_{j\in\Omega}x_ix_j\theta_i\theta_j}{\left(1-\theta_0\right)^2}.
\end{aligned}
$$

Probability mass function of $\widetilde{N}(\boldsymbol{x})$:
By the law of total probability, we can write, for $\tilde{n}=0,1,2$, $\forall\boldsymbol{x}$:

$$
Pr\left(\widetilde{N}(\boldsymbol{x})=\tilde{n}\right)=\sum_{n=\tilde{n}}^{2}Pr\left(\widetilde{N}(\boldsymbol{x})=\tilde{n}|N=n\right)Pr\left(N=n\right).
$$

Then:

$$
\begin{aligned}
Pr\left(\widetilde{N}(\boldsymbol{x})=0\right)=&(1)\left(\frac{1}{4}\left(1+\theta_0\right)^2\right)+\left(\frac{\sum_{i\in\Omega}\left(1-x_i\right)\theta_i}{\left(1-\theta_0\right)}\right)\left(\frac{1}{2}\left(1-{\theta_0}^2\right)\right)\\
&+\left(\frac{\sum_{i\in\Omega}\sum_{j\in\Omega}\left(1-x_i\right)\left(1-x_j\right)\theta_i\theta_j}{\left(1-\theta_0\right)^2}\right)\left(\frac{1}{4}\left(1-\theta_0\right)^2\right)\\
=&\left(\frac{1}{4}\left(1+\theta_0\right)^2\right)+\frac{1}{2}\left(1+\theta_0\right)\left(\sum_{i\in\Omega}\left(1-x_i\right)\theta_i\right)+\frac{1}{4}\left(\sum_{i\in\Omega}\sum_{j\in\Omega}\left(1-x_i\right)\left(1-x_j\right)\theta_i\theta_j\right)\\
Pr\left(\widetilde{N}(\boldsymbol{x})=1\right)=&\left(\frac{\sum_{i\in\Omega}x_i\theta_i}{\left(1-\theta_0\right)}\right)\left(\frac{1}{2}\left(1-{\theta_0}^2\right)\right)+\left(\frac{2\sum_{i\in\Omega}\sum_{j\in\Omega}x_i\left(1-x_j\right)\theta_i\theta_j}{\left(1-\theta_0\right)^2}\right)\left(\frac{1}{4}\left(1-\theta_0\right)^2\right) \qquad (17)\\
=&\frac{1}{2}\left(\left(\sum_{i\in\Omega}x_i\theta_i\right)\left(1+\theta_0\right)+\sum_{i\in\Omega}\sum_{j\in\Omega}x_i\left(1-x_j\right)\theta_i\theta_j\right)\\
Pr\left(\widetilde{N}(\boldsymbol{x})=2\right)=&\left(\frac{\sum_{i\in\Omega}\sum_{j\in\Omega}x_ix_j\theta_i\theta_j}{\left(1-\theta_0\right)^2}\right)\left(\frac{1}{4}\left(1-\theta_0\right)^2\right)=\frac{1}{4}\left(\sum_{i\in\Omega}\sum_{j\in\Omega}x_ix_j\theta_i\theta_j\right).
\end{aligned}
$$

The probability mass functions for the corresponding random variables for each group $r\in R$ follow similarly, with group index r added as a superscript, e.g., N^r.

A.2 False-negative probability

For any $\boldsymbol{\theta}\in S(\boldsymbol{Q})$, we have:

$$
\begin{aligned}
Pr(FN(\boldsymbol{x};\boldsymbol{\theta}))=&Pr\left(\widetilde{N}(\boldsymbol{x})=0,N=2;\boldsymbol{\theta}\right)=Pr\left(\widetilde{N}(\boldsymbol{x})=0|N=2,\boldsymbol{\theta}\right)Pr\left(N=2;\boldsymbol{\theta}\right)\\
=&\frac{1}{4}\sum_{i\in\Omega}\sum_{j\in\Omega}\left(1-x_i\right)\left(1-x_j\right)\theta_i\theta_j \qquad \text{(by Eqs. (14) and (16))}\\
=&\frac{1}{4}\sum_{i\in\Omega}\left(1-x_i\right)\theta_i\sum_{j\in\Omega}\left(1-x_j\right)\theta_j\\
=&\frac{1}{4}\left[\left(1-\theta_0\right)^2-2\left(1-\theta_0\right)\sum_{i\in\Omega}x_i\theta_i+\left(\sum_{i\in\Omega}x_i\theta_i\right)^2\right]\\
=&\frac{1}{4}\left[\left(1-\theta_0\right)-\left(\sum_{i\in\Omega}x_i\theta_i\right)\right]^2.
\end{aligned}
$$

A.3 Expected Testing Cost

One-tier genetic screening process:

$$\begin{aligned}
E_{N,\tilde{N}(\boldsymbol{x})}\left[Cost^1(\boldsymbol{x};\boldsymbol{\theta})\right] =& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left[Pr\left(\widetilde{N}(\boldsymbol{x}) \geq 1; \boldsymbol{\theta}\right)\right] C_{SC} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \frac{C_{SC}}{2}\left[(1+\theta_0)\sum_{i\in\Omega} x_i\theta_i + \sum_{i\in\Omega}\sum_{j\in\Omega} x_i\left(1-x_j\right)\theta_i\theta_j + \frac{1}{2}\sum_{i\in\Omega}\sum_{j\in\Omega} x_i x_j\theta_i\theta_j\right] \text{ (by Eq. (17))} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \frac{C_{SC}}{2}\left[(1+\theta_0)\sum_{i\in\Omega} x_i\theta_i + (1-\theta_0)\sum_{i\in\Omega} x_i\theta_i - \frac{1}{2}\sum_{i\in\Omega}\sum_{j\in\Omega} x_i x_j\theta_i\theta_j\right] \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left(\sum_{i\in\Omega} x_i\theta_i\right) C_{SC} - \left(\sum_{i\in\Omega} x_i\theta_i\right)^2\left[\frac{C_{SC}}{4}\right].
\end{aligned}$$

Two-tier genetic screening process:

$$\begin{aligned}
E_{N,\tilde{N}(\boldsymbol{x})}\left[Cost^2(\boldsymbol{x};\boldsymbol{\theta})\right] =& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + Pr\left(\widetilde{N}(\boldsymbol{x}) = 1; \boldsymbol{\theta}\right)\left[C_{GS} + Pr\left(N = 2|\widetilde{N}(\boldsymbol{x}) = 1, \boldsymbol{\theta}\right) C_{SC}\right] + Pr\left(\widetilde{N}(\boldsymbol{x}) = 2; \boldsymbol{\theta}\right) C_{SC} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left[\frac{C_{GS}\left(1+\theta_0\right)\sum_{i\in\Omega} x_i\theta_i}{2} + \frac{\left(C_{GS}+C_{SC}\right)\sum_{i\in\Omega}\sum_{j\in\Omega} x_i\left(1-x_j\right)\theta_i\theta_j}{2}\right] \\
&+ \left[\frac{C_{SC}\sum_{i\in\Omega}\sum_{j\in\Omega} x_i x_j\theta_i\theta_j}{4}\right] \qquad \text{(by Eqs. (16) and (17))} \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left(\frac{\sum_{i\in\Omega} x_i\theta_i}{2}\right)\left[2C_{GS} + C_{SC}\left(1-\theta_0\right) - \left(C_{GS}+C_{SC}\right)\sum_{i\in\Omega} x_i\theta_i + \frac{C_{SC}\left(\sum_{i\in\Omega} x_i\theta_i\right)}{2}\right] \\
=& C_{MP} + g\left(\sum_{i\in\Omega} x_i\right) + \left(\sum_{i\in\Omega} x_i\theta_i\right)\left[\frac{2C_{GS} + C_{SC}\left(1-\theta_0\right)}{2}\right] - \left(\sum_{i\in\Omega} x_i\theta_i\right)^2\left[\frac{2C_{GS}+C_{SC}}{4}\right].
\end{aligned}$$

A.4 Supporting Results

In this section, we provide a set of results that will be used in the subsequent proofs. We use the notation $K_1(\theta_0)$ to express parameter K_1 as a function of θ_0; and drop the process type index, $l = 1, 2$, when the result or proof holds for both process types.

Property A.1. *The following result applies to both process types $l = 1, 2$, hence we drop the process type index l. For any $\boldsymbol{\theta} \in S(\boldsymbol{Q})$, we have that:*

1. *$K_1(\theta_0) > 2(1-\theta_0)K_2$.*
2. *$E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]$ is strictly increasing in $\sum_{i\in\Omega} x_i\theta_i$, and non-increasing in θ_0.*

Proof. Consider any $\boldsymbol{\theta} \in S(\boldsymbol{Q})$.

1. From Eq. (3), for process type 1, $K_1^1(\theta_0) = C_{SC} > \frac{C_{SC}}{2} = 2K_2^1 \geq 2(1-\theta_0)K_2^1$, and for process type 2, $K_1^2(\theta_0) = \left[\frac{2C_{GS}+C_{SC}(1-\theta_0)}{2}\right] > (1-\theta_0)\left[\frac{2C_{GS}+C_{SC}}{2}\right] = 2(1-\theta_0)K_2^2$, and the result follows.

2. From Part 1 of this property, we have that for both process types $l = 1, 2$, $K_1^l(\theta_0) > 2(1-\theta_0)K_2^l \geq 2K_2^l\sum_{i\in\Omega} x_i\theta_i \geq 0$. Then, $\frac{\partial E\left[Cost^l(\boldsymbol{x};\boldsymbol{\theta})\right]}{\partial\left(\sum_{i\in\Omega} x_i\theta_i\right)} = K_1^l(\theta_0) - 2K_2^l\sum_{i\in\Omega} x_i\theta_i > 0$, $l = 1, 2$. We also have that $\frac{\partial E\left[Cost^l(\boldsymbol{x};\boldsymbol{\theta})\right]}{\partial\left(K_1^l(\theta_0)\right)} = \sum_{i\in\Omega} x_i\theta_i \geq 0$, $l = 1, 2$, $\frac{\partial\left(K_1^1(\theta_0)\right)}{\partial\theta_0} = 0$, and $\frac{\partial\left(K_1^2(\theta_0)\right)}{\partial\theta_0} = \frac{-C_{SC}\sum_{i\in\Omega} x_i\theta_i}{2} \leq 0$. Therefore, we have that, for both process types $l = 1, 2$, $E\left[Cost^l(\boldsymbol{x};\boldsymbol{\theta})\right]$ is strictly increasing in $\sum_{i\in\Omega} x_i\theta_i$ and non-decreasing in $K_1^l(\theta_0)$, and $K_1^l(\theta_0)$ is non-increasing in θ_0. Then, it follows that $E\left[Cost^l(\boldsymbol{x};\boldsymbol{\theta})\right]$ is non-increasing in θ_0, $l = 1, 2$. □

Property A.2.

1. $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega} x_i\theta_i\right\} = \min\left\{\sum_{i\in\Omega} x_i q_i^{UB}, 1-q_0^{LB}-\sum_{i\in\Omega}\left(1-x_i\right)q_i^{LB}\right\}.$

2. $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0+\sum_{i\in\Omega} x_i\theta_i\right\} = \max\left\{q_0^{LB}+\sum_{i\in\Omega} x_i q_i^{LB}, 1-\sum_{i\in\Omega}\left(1-x_i\right)q_i^{UB}\right\}.$

Proof. By definition, $\sum_{i\in\Omega}\theta_i = 1-\theta_0$, $\forall\boldsymbol{\theta}\in S(\boldsymbol{Q})$, and hence, $\theta_0+\sum_{i\in\Omega}x_i\theta_i = 1-\sum_{i\in\Omega}(1-x_i)\theta_i$.

1. We have that, $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\} \leq \sum_{i\in\Omega}x_iq_i^{UB}$, and $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\} \leq 1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}$, and at least one of these inequalities must be binding in the solution to $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\}$.
2. We have that, $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0+\sum_{i\in\Omega}x_i\theta_i\right\} \geq q_0^{LB}+\sum_{i\in\Omega}x_iq_i^{LB}$, and $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0+\sum_{i\in\Omega}x_i\theta_i\right\} \geq 1-\sum_{i\in\Omega}(1-x_i)q_i^{UB}$, and at least one of these inequalities must be binding in the solution to $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0+\sum_{i\in\Omega}x_i\theta_i\right\}$. □

Property A.3. *The following result applies to both process types $l=1,2$, hence we drop the process type index l:*

$$\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]\right\} = C_{MP}+g\left(\sum_{i\in\Omega}x_i\right)+\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{K_1(\theta_0)\sum_{i\in\Omega}x_i\theta_i - K_2\left(\sum_{i\in\Omega}x_i\theta_i\right)^2\right\}$$
$$= C_{MP}+g\left(\sum_{i\in\Omega}x_i\right)+\min\left\{C_{UB}^1(\boldsymbol{x}),C_{UB}^2(\boldsymbol{x})\right\},$$

where $C_{UB}^1(\boldsymbol{x}) = K_1\left(q_0^{LB}\right)\sum_{i\in\Omega}x_iq_i^{UB} - K_2\left(\sum_{i\in\Omega}x_iq_i^{UB}\right)^2$, *and* $C_{UB}^2(\boldsymbol{x}) = K_1\left(q_0^{LB}\right)\left(1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}\right) - K_2\left(1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}\right)^2$.

Proof. For any process type l=1,2, and for any panel $\boldsymbol{x}$, we have that $E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]$ is strictly increasing in $\sum_{i\in\Omega}x_i\theta_i$, and non-increasing in θ_0 (Property A.1). Further, because $\boldsymbol{\theta}\in S(\boldsymbol{Q})$, by definition of $S(\boldsymbol{Q})$ we have that, $1-q_0^{LB}\leq\sum_{i\in\Omega}q_i^{UB} = \sum_{i\in\Omega}x_iq_i^{UB}+\sum_{i\in\Omega}(1-x_i)q_i^{UB}$. By Property A.2, the term, $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\}$, attains two possible values, both of which are considered below:

<u>Case 1</u>: $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\} = \sum_{i\in\Omega}x_iq_i^{UB} \leq 1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}$. Hence, we have that $1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB} \leq \sum_{i\in\Omega}x_iq_i^{UB} \leq 1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}$. Then, we have that $q_0^{LB}+\sum_{i\in\Omega}x_iq_i^{UB}+\sum_{i\in\Omega}(1-x_i)q_i^{LB}\leq 1$ and $q_0^{LB}+\sum_{i\in\Omega}x_iq_i^{UB}+\sum_{i\in\Omega}(1-x_i)q_i^{UB}\geq 1$. Hence, there exists a realization $\boldsymbol{\theta}$ such that $\theta_0=q_0^{LB}$, $\sum_{i\in\Omega}x_i\theta_i=\sum_{i\in\Omega}x_iq_i^{UB}$, and $\sum_{i\in\Omega}(1-x_i)\theta_i = 1-q_0^{LB}-\sum_{i\in\Omega}x_iq_i^{UB}$. Therefore, we must have that $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]\right\} = C_{MP}+C_{UB}^1(\boldsymbol{x}) \leq C_{MP}+C_{UB}^2(\boldsymbol{x})$.

<u>Case 2</u>: $\sum_{i\in\Omega}x_iq_i^{LB}\leq\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\sum_{i\in\Omega}x_i\theta_i\right\} = 1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}\leq\sum_{i\in\Omega}x_iq_i^{UB}$. Hence, there exists a realization $\boldsymbol{\theta}$ such that $\theta_0=q_0^{LB}$, $\sum_{i\in\Omega}x_i\theta_i = 1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{LB}$, and $\sum_{i\in\Omega}(1-x_i)\theta_i = \sum_{i\in\Omega}(1-x_i)q_i^{LB}$. Therefore, we must have that $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]\right\} = C_{MP}+C_{UB}^2(\boldsymbol{x})\leq C_{MP}+C_{UB}^1(\boldsymbol{x})$, completing the proof. □

Remark A.1. Consider the second-degree polynomial:

$$-K_2(Z(\gamma))^2+K_1Z(\gamma)+g\left(\gamma\right)-B'=0. \tag{18}$$

Let $Z_1(\gamma)$ and $Z_2(\gamma)$ respectively denote the smallest and largest roots of this polynomial. We have that:

$$Z_{1,2}(\gamma)=\frac{-K_1\pm\sqrt{\delta(\gamma)}}{-2K_2}=\frac{K_1\pm\sqrt{\delta(\gamma)}}{2K_2},$$

where $\delta(\gamma) \equiv (K_1)^2 + 4[g(\gamma) - B']K_2$.

B Robust Model Variations: Formulations and Properties

We provide the formulations and properties of the two additional robust models introduced in Remark 1, namely **Correlated Robust MSP (C-RM)** and **Upper Bound Robust MSP (U-RM)**. Recall that $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \equiv \text{argmax}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\right\} \equiv \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}$ (if multiple such vectors exist, then, without loss of optimality, the vector that yields the smallest expected testing cost, i.e., the LHS of Constraint (5), is selected); and $\boldsymbol{\beta}^{Cost}(\boldsymbol{x}) \equiv \text{argmax}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]\right\}$.

B.1 Correlated Robust MSP (C-RM)

C-RM:

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad \beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) \tag{19}$$

$$\text{subject to} \quad g\left(\sum_{i\in\Omega} x_i\right) + K_1\left(\beta_0^{FN}(\boldsymbol{x})\right)\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) - K_2\left(\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})\right)^2 \leq B' \tag{20}$$

$$\boldsymbol{\beta}^{FN}(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}, \quad \forall \boldsymbol{x} \tag{21}$$

$$\sum_{i\in\Omega} x_i \leq \overline{m}$$

$$x_i \text{ binary, } i \in \Omega.$$

Next, we provide a reformulation of **C-RM** under the additional constraint, $\sum_{i\in\Omega} x_i = \gamma$, for any $\gamma \in \mathbb{Z}^+ : \gamma \leq \overline{m}$, which we refer to as Sub-problem **C-RM-S(γ)**. To this end, we first provide a property that follows in light of Property A.2.

Lemma B.1. *For any given $\boldsymbol{x}$, $\beta_0^{FN}(\boldsymbol{x})$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$ take one of the following pairs of values:*

1. *If $q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB} \geq 1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$, then $\beta_0^{FN}(\boldsymbol{x}) = q_0^{LB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$.*

2. *If $1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB} \geq q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}$, then there may be multiple solutions for $\beta_0^{FN}(\boldsymbol{x})$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$, and the solution that yields the smallest expected cost is given by either one of the following solutions:*
   - *(a) $\beta_0^{FN}(\boldsymbol{x}) = q_0^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - q_0^{UB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$, or*
   - *(b) $\beta_0^{FN}(\boldsymbol{x}) = 1 - \sum_{i\in\Omega} x_i q_i^{LB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$.*

Proof. Consider any $\boldsymbol{x}$. We have that,

$$\begin{aligned}\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) &= \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} \\ &= \max\left\{q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}, 1 - \sum_{i\in\Omega}(1-x_i)\, q_i^{UB}\right\},\end{aligned} \tag{22}$$

where the first equality follows by definition (Eq. (21)) and the second equality follows by Property A.2. Hence, if we can identify a prevalence vector $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \in S(\boldsymbol{Q})$ that satisfies Eq. (22), then this must be the vector that maximizes the false-negative probability for the given $\boldsymbol{x}$ vector. In order for $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \in S(\boldsymbol{Q})$, it must satisfy the following conditions:

$$\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} \beta_i^{FN}(\boldsymbol{x}) = 1, \tag{23a}$$

$$q_0^{LB} \le \beta_0^{FN}(\boldsymbol{x}) \le q_0^{UB}, \text{ and} \tag{23b}$$

$$q_i^{LB} \le \beta_i^{FN}(\boldsymbol{x}) \le q_i^{UB}, \ i \in \Omega. \tag{23c}$$

In what follows, we consider all possible values of the term, $\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$:

<u>Case 1:</u> $q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB} \ge 1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$. Then, Eq. (22) reduces to:

$$\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}.$$

Based on Eqs. (23b) and (23c), the unique solution in this case is given by $\beta_0^{FN}(\boldsymbol{x}) = q_0^{LB}$, $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$, and $\sum_{i\in\Omega}(1-x_i)\beta_i^{FN} = 1 - \beta_0^{FN}(\boldsymbol{x}) - \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - q_0^{LB} - \sum_{i\in\Omega} x_i q_i^{LB}$; hence all the conditions in Eq. (23) are satisfied, and $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \in S(\boldsymbol{Q})$.

<u>Case 2:</u> $1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB} \ge q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}$. Then, Eq. (22) reduces to:

$$\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB}. \tag{24}$$

In this case, there may be multiple solutions to Eqs. (23)-(24), and without loss of optimality, we choose the one that yields the smallest expected testing cost, $E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]$.

First, note the following set of constraints on β_0^{FN} and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$, implied by Eqs. (23)-(24):

$$\begin{aligned}
\beta_0^{FN}(\boldsymbol{x}) &\le q_0^{UB} && \text{(Eq. (23b))},\\
\beta_0^{FN}(\boldsymbol{x}) = 1 - \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) - \sum_{i\in\Omega}(1-x_i)q_i^{UB} &\le 1 - \sum_{i\in\Omega} x_i q_i^{LB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB} && \text{(From Eqs. (23c) and (24))},\\
\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) &\ge \sum_{i\in\Omega} x_i q_i^{LB} && \text{(From Eq. (23c)), and}\\
\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - \beta_0^{FN}(\boldsymbol{x}) - \sum_{i\in\Omega}(1-x_i)q_i^{UB} &\ge 1 - q_0^{UB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB} && \text{(From Eqs. (23b) and (24))}.
\end{aligned}$$

Also recall that $E\left[Cost(\boldsymbol{x};\boldsymbol{\theta})\right]$ is strictly increasing in $\sum_{i\in\Omega} x_i\theta_i$, and non-increasing in θ_0 (Property A.1). Therefore, $\beta_0^{FN}(\boldsymbol{x})$ must take its highest feasible value, and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$ must take its lowest feasible value.

Therefore, if we can show that there exists a vector $\boldsymbol{\beta}^{FN}(\boldsymbol{x})$ such that:

$$\begin{aligned}
\beta_0^{FN}(\boldsymbol{x}) &= \min\left\{q_0^{UB}, 1 - \sum_{i\in\Omega} x_i q_i^{LB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB}\right\}, \quad \text{and}\\
\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) &= \max\left\{\sum_{i\in\Omega} x_i q_i^{LB}, 1 - q_0^{UB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB}\right\},
\end{aligned} \tag{25}$$

then this must be the vector that satisfies the definition of $\boldsymbol{\beta}^{FN}(\boldsymbol{x})$, and yields the smallest expected cost. From Eq. (24) we have that $\sum_{i\in\Omega}(1-x_i)\beta_i^{FN} = 1 - \beta_0^{FN}(\boldsymbol{x}) - \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega}(1-x_i)q_i^{UB}$. In the following, we analyze all possible solutions to Eq. (25):

Sub-case 2(a): If $1-q_0^{UB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB} \ge \sum_{i\in\Omega} x_i q_i^{LB} \Leftrightarrow q_0^{UB} \le 1-\sum_{i\in\Omega} x_i q_i^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}$. Then, by Eq. (25), it follows that, $\beta_0^{FN}(\boldsymbol{x}) = q_0^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1-q_0^{UB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}$; hence all the conditions in Eq. (23) are satisfied, and $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \in S(\boldsymbol{Q})$.

Sub-case 2(b): If $\sum_{i\in\Omega} x_i q_i^{LB} \ge 1-q_0^{UB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB} \Leftrightarrow q_0^{UB} \ge 1-\sum_{i\in\Omega} x_i q_i^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}$. Then, by Eq. (25), it follows that, $\beta_0^{FN}(\boldsymbol{x}) = 1-\sum_{i\in\Omega} x_i q_i^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$; hence all the conditions in Eq. (23) are satisfied, and $\boldsymbol{\beta}^{FN}(\boldsymbol{x}) \in S(\boldsymbol{Q})$. This completes the proof. □

Recall that $Z_{UB}(\gamma,\theta_0) \equiv \frac{K_1(\theta_0)-\sqrt{(K_1(\theta_0))^2+4[g(\gamma)-B']K_2}}{2K_2}$ (Theorem 1), which we express as a function of both γ and θ_0.

Lemma B.2. *For any $\gamma \in \mathbb{Z}^+ : \gamma \le \overline{m}$, **C-RM**, under the additional constraint that $\sum_{i\in\Omega} x_i = \gamma$, can be equivalently formulated as follows:*
***Sub-problem C-RM-S**(γ):*

$$
\begin{aligned}
\underset{\boldsymbol{x}}{\text{maximize}} \quad & \beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) \\
\text{subject to} \quad & \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) \le Z_{UB}(\gamma, \beta_0^{FN}(\boldsymbol{x})) \\
& \boldsymbol{\beta}^{FN}(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{ \theta_0 + \sum_{i\in\Omega} x_i\theta_i \right\}, \quad \forall \boldsymbol{x} \\
& \sum_{i\in\Omega} x_i = \gamma \\
& x_i \text{ binary}, \ i\in\Omega,
\end{aligned}
\tag{26}
$$

with the following deterministic counterpart:
***Deterministic Counterpart of Sub-problem C-RM-S**(γ):*

$$
\begin{aligned}
\underset{\boldsymbol{x},u,v,w}{\text{maximize}} \quad & u\left(q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}\right) + (1-u)\left(1-\sum_{i\in\Omega}(1-x_i)q_i^{UB}\right) \\
\text{subject to} \quad & (1-w)\sum_{i\in\Omega} x_i q_i^{LB} \le u \times Z_{UB}(\gamma, q_0^{LB}) + v \times Z_{UB}\left(\gamma, 1-\sum_{i\in\Omega} x_i q_i^{LB} - \sum_{i\in\Omega}(1-x_i)q_i^{UB}\right) \\
& 1-\sum_{i\in\Omega}(1-x_i)q_i^{UB} \le u\left(q_0^{LB}+\sum_{i\in\Omega} x_i q_i^{LB}\right) + v\left(q_0^{UB}+\sum_{i\in\Omega} x_i q_i^{LB}\right) + w\left(q_0^{UB} + Z_{UB}(\gamma, q_0^{UB})\right) \\
& (1-u)\sum_{i\in\Omega} x_i q_i^{LB} \le v\left(1-q_0^{LB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}\right) + w\left(1-q_0^{UB}-\sum_{i\in\Omega}(1-x_i)q_i^{UB}\right) \\
& \sum_{i\in\Omega} x_i = \gamma \\
& x_i \text{ binary}, \ i\in\Omega \\
& u+v+w = 1 \\
& u, v, w \text{ binary.}
\end{aligned}
$$

Proof. The formulation of **C-RM-S(γ)** follows directly from Theorem 1. The formulation for its deterministic counterpart follows directly from Lemma B.1, because the case where $u = 1$ (hence $v = 0, w = 0$) corresponds to the case where $\beta_0^{FN}(\boldsymbol{x}) = q_0^{LB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$ (Case 1 in Lemma B.1); the case where $w = 1$ (hence $u = 0, v = 0$) corresponds to the case where $\beta_0^{FN}(\boldsymbol{x}) = q_0^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - q_0^{UB} - \sum_{i\in\Omega}(1 - x_i)q_i^{UB}$ (Case 2(a) in Lemma B.1); and the case where $v = 1$ (hence $u = 0, w = 0$) corresponds to the case where $\beta_0^{FN}(\boldsymbol{x}) = 1 - \sum_{i\in\Omega} x_i q_i^{LB} - \sum_{i\in\Omega}(1 - x_i)q_i^{UB}$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{LB}$ (Case 2(b) of Lemma B.1). Thus, the deterministic counterpart formulation accounts for all possible values for $\beta_0^{FN}(\boldsymbol{x})$ and $\sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$, and chooses a vector $(u(\boldsymbol{x}), v(\boldsymbol{x}), w(\boldsymbol{x}))$ for each possible $\boldsymbol{x}$, that maximizes the objective function. □

Remark B.1. The solutions with $u = 1, w = 1$, and $v = 1$ in the deterministic counterpart of Sub-problem **C-RM-S**(γ) (Lemma B.2) respectively represent Parts 1, 2(a), and 2(b) of Lemma B.1. In each case, **C-RM-S**(γ) reduces to a 0-1 multiconstraint Knapsack Problem. Thus, in order to solve **C-RM-S**(γ), it is sufficient to solve three 0-1 multiconstraint Knapsack Problems, one for each case, i.e., $u = 1, w = 1$, and $v = 1$.

B.2 Upper Bound Robust MSP (U-RM)

U-RM:

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{maximize}} \quad & \beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) \\
\text{subject to} \quad & g\left(\sum_{i\in\Omega} x_i\right) + K_1\left(\beta_0^{Cost}(\boldsymbol{x})\right)\sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x}) - K_2\left(\sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x})\right)^2 \le B' \qquad (27) \\
& \boldsymbol{\beta}^{FN}(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}, \quad \forall \boldsymbol{x} \\
& \boldsymbol{\beta}^{Cost}(\boldsymbol{x}) = \text{argmax}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{K_1(\theta_0)\sum_{i\in\Omega} x_i\theta_i - K_2\left(\sum_{i\in\Omega} x_i\theta_i\right)^2\right\}, \quad \forall \boldsymbol{x} \\
& \sum_{i\in\Omega} x_i \le \overline{m} \\
& x_i \text{ binary}, \ i \in \Omega.
\end{aligned}$$

Next, we provide a reformulation of **U-RM** under the additional constraint, $\sum_{i\in\Omega} x_i = \gamma$, for any $\gamma \in \mathbb{Z}^+ : \gamma \le \overline{m}$, which we refer to as Sub-problem **U-RM-S(γ)**, along with its deterministic counterpart.

Lemma B.3. *For any $\gamma \in \mathbb{Z}^+ : \gamma \le \overline{m}$, **U-RM**, under the additional constraint that $\sum_{i\in\Omega} x_i = \gamma$, can be equivalently formulated as follows:*
***Sub-problem U-RM-S**(γ):*

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{maximize}} \quad & \beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) \\
\text{subject to} \quad & \sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x}) \le Z_{UB}(\gamma, \beta_0^{Cost}(\boldsymbol{x})) \\
& \boldsymbol{\beta}^{FN}(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}, \quad \forall \boldsymbol{x}
\end{aligned}$$

$$\boldsymbol{\beta}^{Cost}(\boldsymbol{x}) = argmax_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{ K_1(\theta_0)\sum_{i\in\Omega} x_i\theta_i - K_2\left(\sum_{i\in\Omega} x_i\theta_i\right)^2 \right\}, \quad \forall \boldsymbol{x}$$

$$\sum_{i\in\Omega} x_i = \gamma$$

$$x_i \text{ binary, } i \in \Omega,$$

with the following deterministic counterpart:

Deterministic Counterpart of Sub-problem U-RM-S(γ):

$$\begin{aligned}
\underset{\boldsymbol{x},u,v}{\text{maximize}} \quad & u\left(q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}\right) + (1-u)\left(1 - \sum_{i\in\Omega}(1-x_i)\, q_i^{UB}\right) \\
\text{subject to} \quad & v\sum_{i\in\Omega} x_i q_i^{UB} + (1-v)\left(1 - q_0^{LB} - \sum_{i\in\Omega}(1-x_i)\, q_i^{LB}\right) \le Z_{UB}(\gamma, q_0^{LB}) \qquad (28) \\
& \sum_{i\in\Omega} x_i = \gamma \\
& x_i \text{ binary, } i \in \Omega \\
& u, v \text{ binary.}
\end{aligned}$$

Proof. Consider any $\gamma \in Z^+ : \gamma \le \overline{m}$, and recall that $Z(\gamma;\boldsymbol{\theta}) \equiv \sum_{i\in\Omega} x_i\theta_i$: $\sum_{i\in\Omega} x_i = \gamma$ and $Z_{UB}(\gamma, K_1(q_0^{LB})) = \frac{K_1(q_0^{LB}) - \sqrt{(K_1(q_0^{LB}))^2 + 4[g(\gamma) - B']K_2}}{2K_2}$. Because the $Z(\gamma;\boldsymbol{\theta})$ function is not one-to-one, we define its set of solutions for a given γ as, $S\left(Z(\gamma;\boldsymbol{\theta})\right) = \left\{\sum_{i\in\Omega} x_i\theta_i : \sum_{i\in\Omega} x_i = \gamma,\ x_i \text{ binary}, i \in \Omega\right\}$. By definition of γ, Constraint (6) is automatically satisfied. Then, the feasible region of **U-RM**, i.e., the region where Constraint (5) is also satisfied, can be expressed as follows:

$$\begin{aligned}
\underset{Z(\gamma;\boldsymbol{\theta})}{\text{maximize}} \quad & \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + Z(\gamma;\boldsymbol{\theta})\right\} \\
\text{subject to} \quad & K_1\left(q_0^{LB}\right) \max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{Z(\gamma;\boldsymbol{\theta})\right\} - K_2\left(\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{Z(\gamma;\boldsymbol{\theta})\right\}\right)^2 \le B' - g(\gamma) \qquad (29) \\
& Z(\gamma;\boldsymbol{\theta}) \in S\left(Z(\gamma;\boldsymbol{\theta})\right). \qquad (30)
\end{aligned}$$

Then the result follows directly by Theorem 1. The deterministic counterpart of Sub-problem **U-RM-S(γ)** follows directly from Properties A.2 and A.3, because for the objective function, the case where $u = 1$ corresponds to the case where $\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}$; and the case where $u = 0$ corresponds to the case where $\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x}) = 1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$; and for the budget constraint, i.e., Constraint (28), the case where $v = 1$ corresponds to the case where $\sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i^{UB}$; and the case where $v = 0$ corresponds to the case where $\sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x}) = 1 - q_0^{LB} - \sum_{i\in\Omega} x_i q_i^{LB}$. Thus, the deterministic counterpart formulation accounts for all possible values for $\beta_0^{FN}(\boldsymbol{x}) + \sum_{i\in\Omega} x_i\beta_i^{FN}(\boldsymbol{x})$, and $\sum_{i\in\Omega} x_i\beta_i^{Cost}(\boldsymbol{x})$, and chooses a vector $(u(\boldsymbol{x}), v(\boldsymbol{x}))$ for each possible $\boldsymbol{x}$, that maximizes the objective function. □

Remark B.2. For each value of $u \in \{0,1\}$ and $v \in \{0,1\}$, **U-RM-S(γ)** reduces to an exact γ-item 0-1 Knapsack Problem. Thus, in order to solve **U-RM-S(γ)**, it is sufficient to solve four exact γ-item 0-1 Knapsack Problems, one for each pair of values, $(u,v) \in \{(0,0),(0,1),(1,0),(1,1)\}$.

C Additional Properties and Reformulations

A Special Case of DM and RM:

Remark C.1. For the special case where $g\left(\sum_{i\in\Omega} x_i\right) = 0$, i.e., with no MP variable cost, $Z_{UB}(\gamma)$ reduces to a constant independent of γ, with $Z_{UB}\left(\gamma; g\left(\sum_{i\in\Omega} x_i\right) = 0\right) = \frac{K_1-\sqrt{(K_1)^2-4B'K_2}}{2K_2}$ (see Theorem 1), and Constraint (10) can be eliminated without loss of optimality. In this special case, to solve **DM** to optimality, it is sufficient to solve a single 0-1 Subset-sum Problem, and to solve **RM** to optimality, it is sufficient to solve two 0-1 Knapsack Problems, one for each value of $u \in \{0,1\}$.

A Mixed Integer Linear Programming Reformulation for DM:

Remark C.2.

1. When the MP variable cost, $g\left(\sum_{i\in\Omega} x_i\right)$, is a polynomial of degree $n \in Z^+ : n \geq 2$, **DM** can be reformulated as a mixed integer linear programming problem (MILP), following the standard linearization techniques (e.g., Adams and Sherali (1986); Smith and Taskin (2008)), that is, by using the fact that $(x_i)^k = x_i$, $\forall k \in \{2, \cdots, n\}$, $i \in \Omega$, and introducing additional binary variables, $y_{i_1,i_2,...,i_w} \equiv x_{i_1} \times x_{i_2} \times ... \times x_{i_w}$, along with constraints, $\left\{ \sum_{q=i_1}^{i_w} x_q - w + 1 \leq y_{i_1,i_2,...,i_w},\ y_{i_1,i_2,...,i_w} \leq x_q,\ q \in \{i_1, i_2, ..., i_w\} \right\}$, for all distinct w-tuples $\{i_1, ..., i_w\} \in \Omega : i_1 \neq i_2 \cdots \neq i_w$, $\forall w \in \{2, \cdots, n\}$, leading to $\sum_{w=2}^{n} \binom{m}{w}$ additional binary variables, and $\sum_{w=2}^{n} (w+1) \times \binom{m}{w}$ additional constraints.

   For the special case where $g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^n$, for some constant $c > 0$, constraints $\left\{ y_{i_1,i_2,...,i_w} \leq x_q, q \in \{i_1, i_2, ..., i_w\} \right\}$ become redundant, $\forall w \geq 3$ (because $E[Cost(\boldsymbol{x}, \boldsymbol{y}; \hat{\boldsymbol{\mu}})]$ is strictly increasing in $y_{i_1,i_2,...,i_w}$), reducing the number of additional constraints to $\sum_{w=2}^{n} \binom{m}{w} + 2 \times \binom{m}{2}$.

2. For the linear variable cost case where $g\left(\sum_{i\in\Omega} x_i\right) = c\sum_{i\in\Omega} x_i$, for some constant $c > 0$, constraints $\left\{ x_{i_1} + x_{i_2} - 1 \leq y_{i_1,i_2},\ i_1, i_2 \in \Omega, i_1 \neq i_2 \right\}$ become redundant (because $E[Cost(\boldsymbol{x}, \boldsymbol{y}; \hat{\boldsymbol{\mu}})]$ is strictly decreasing in y_{i_1,i_2}), and the MILP reformulation requires $\binom{m}{2}$ additional binary variables and $2 \times \binom{m}{2}$ additional constraints.

Table 8 provides the computational times, as well as the number of additional binary variables and constraints required for the MILP reformulation for **DM**, provided in Remark C.2, for various polynomial functions for the MP variable cost, $g(.)$, and various number of variants, m.

Table 8: Comparison of **Algorithm MSO**, MILP Reformulation, and Gurobi Solver for Various **DM** Problem Instances*

m	MILP Reformulation		Run Time (seconds)		
	Number of Additional Binary Variables	Number of Additional Constraints	MILP Reformulation	Gurobi	**MSO**
$g\left(\sum_{i\in\Omega} x_i\right) = c\sum_{i\in\Omega} x_i$					
312	48,516	97,032	86	7	1
346	59,685	119,370	91	7	1
400	79,800	159,600	99	8	1
450	101,025	202,050	225	9	1
500	124,750	249,500	582	10	2
550	150,975	301,950	out of memory	10	2
600	179,700	359,400	out of memory	10	2
1,000	499,500	999,000	out of memory	28	3
2,000	1,999,000	3,998,000	out of memory	138	3
$g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^2$					
312	48,516	145,548	93	7	1
346	59,685	179,055	122	8	1
400	79,800	239,400	313	10	1
450	101,025	303,075	out of memory	10	1
500	124,750	374,250	out of memory	12	1
550	150,975	452,925	out of memory	13	1
600	179,700	539,100	out of memory	15	1
1,000	499,500	1,498,500	out of memory	out of memory	1
2,000	1,999,000	5,997,000	out of memory	out of memory	2
$g\left(\sum_{i\in\Omega} x_i\right) = c\left(\sum_{i\in\Omega} x_i\right)^3$					
312	5,061,836	5,158,868	out of memory	could not solve	1
346	6,903,565	7,022,935	out of memory	could not solve	1
400	10,666,600	10,826,200	out of memory	could not solve	1
450	15,187,425	15,389,475	out of memory	could not solve	1
500	20,833,250	21,082,750	out of memory	could not solve	1
550	27,729,075	28,031,025	out of memory	could not solve	1
600	35,999,900	36,359,300	out of memory	could not solve	1
1,000	166,666,500	167,665,500	out of memory	could not solve	1
2,000	1,333,333,000	1,337,331,000	out of memory	could not solve	1

*Performed on a computer with an *i*-7 processor @2.90GHz, 16Gb RAM, and 64-*bit* operating system.

Price of Robustness Ratio:

The bounds on the price of robustness ratio provided in Theorem 5 require the optimal **DM** solution. In the following, we also provide bounds that are potentially weaker, but that do not require the optimal **DM** solution.

Lemma C.1. *For any given $\boldsymbol{\mu} \in S(\boldsymbol{Q})$:*

$$\Pi(\boldsymbol{\mu}) \leq \frac{1}{\left(1-\mu_0 - Z_{UB}(\gamma_{LB})\right)^2} \min\left\{ 1-\mu_0 - \min\left\{ \left(\frac{\sum_{i\in\Omega: i\leq\gamma_{LB}} \mu_i \left(1+h_i^{UB}\right)}{(1+h_{max}^{UB})} \right)^2, \left(\frac{\sum_{i\in\Omega: i\leq\gamma_{LB}} \mu_i \left(1-h_i^{LB}\right)}{(1-h_{min}^{LB})} \right)^2 \right\}, \right.$$
$$\left. \left(1-\mu_0(1-h_0^{LB}) - \sum_{i\in\Omega: i\leq\gamma_{LB}} \mu_i \left(1-h_i^{LB}\right) \right)^2 \right\},$$

where γ_{LB} is as defined in Theorem 2.

Proof. The result follows directly from Theorem 5 and from the fact that $Z^{*D}(\gamma_{LB}) \leq Z^{*D} \leq Z_{UB}(\gamma^{*D}) \leq Z_{UB}(\gamma_{LB})$, where the first and second inequalities follow by Theorem 3, and the third inequality follows because $Z_{UB}(\gamma)$ is strictly decreasing in γ (see Theorem 1). □

Corollary C.1. For any given $\boldsymbol{\mu} \in S(\boldsymbol{Q})$, if $h_i^{LB} = h^{LB}$ and $h_i^{UB} = h^{UB}$, for all $i \in \Omega$, then $\boldsymbol{x}^{*D}(\boldsymbol{\mu}) = \boldsymbol{x}^{*R}(\boldsymbol{\mu}) \Leftrightarrow \Pi(\boldsymbol{\mu}) = 1$.

Proof. When $h_i^{LB} = h^{LB}$ for all $i \in \Omega$, $h_{min}^{LB} = h_{max}^{LB} = h^{LB}$. Similarly, when $h_i^{UB} = h^{UB}$ for all $i \in \Omega$, $h_{min}^{UB} = h_{max}^{UB} = h^{UB}$ (Theorem 5). Then, the result follows from Theorem 1, because **DM** and **RM** have the same feasible region, and because the objective, $\left\{\text{maximize}_{\boldsymbol{x}} \sum_{i\in\Omega} x_i\mu_i\right\} = \left\{\text{maximize}_{\boldsymbol{x}}\, 1-\mu_0 - \sum_{i\in\Omega}(1-x_i)\mu_i\right\}$, is equivalent to both objectives, $\left\{\text{maximize}_{\boldsymbol{x}}\, q_0^{LB} + (1-h^{LB})\sum_{i\in\Omega} x_i\mu_i\right\}$ and $\left\{\text{maximize}_{\boldsymbol{x}}\, 1-(1+h^{UB})\sum_{i\in\Omega}(1-x_i)\mu_i\right\}$. Thus, the optimal **DM** and **RM** solutions coincide. □

D Proofs of the Results in Section 3

Proof of Property 1.

1. We have that $Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right) = \frac{1}{4}\left(1-\theta_0 - \sum_{i\in\Omega} x_i\theta_i\right)^2$, $\forall\boldsymbol{\theta} \in S(\boldsymbol{Q})$ (Appendix A and Eq. (1)). Then, the proof follows because $\theta_0 + \sum_{i\Omega} x_i\theta_i \leq 1, \forall\boldsymbol{x}$ (by definition of $\boldsymbol{\theta}$), and the term, $\sum_{i\in\Omega} x_i\theta_i$, is strictly increasing in x_i, $i \in \Omega$. Therefore, we have that $\left\{\text{ minimize}_{\boldsymbol{x}}\, Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right) \right\} \equiv \left\{\text{ maximize}_{\boldsymbol{x}} \sum_{i\in\Omega} x_i\theta_i\right\}$, and $\left\{\text{minimize}_{\boldsymbol{x}} \max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{Pr\left(FN(\boldsymbol{x};\boldsymbol{\theta})\right)\right\}\right\} \equiv \left\{\text{maximize}_{\boldsymbol{x}} \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}\right\}$.
2. Since the result holds for $E\left[Cost^l(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right]$, $l = 1, 2$, we drop the process type index, l, and derive, $\forall i \in \Omega$:

$$\frac{\partial E\left[Cost(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right]}{\partial x_i} = \frac{\partial g\left(\sum_{i\in\Omega} x_i\right)}{\partial x_i} + \hat{\mu}_i K_1 - 2\hat{\mu}_i K_2 \sum_{i\in\Omega} x_i\hat{\mu}_i \geq \hat{\mu}_i K_1 - 2\hat{\mu}_i(1-\hat{\mu}_0)K_2 > 0,$$

which follows because $g\left(\sum_{i\in\Omega} x_i\right)$ is strictly increasing in x_i, $\sum_{i\in\Omega} x_i\hat{\mu}_i \leq 1-\hat{\mu}_0$ (by definition of $\hat{\boldsymbol{\mu}}$), and $K_1 > 2(1-\hat{\mu}_0)K_2$ for both process types (by Property A.1). This completes the proof. □

Proof of Theorem 1. Consider any $\gamma \in Z^+ \colon \gamma \leq \overline{m}$, and recall that $Z(\gamma) \equiv \sum_{i\in\Omega} x_i\hat{\mu}_i$: $\sum_{i\in\Omega} x_i = \gamma$ and $Z_{UB}(\gamma) = \frac{K_1 - \sqrt{(K_1)^2 + 4[g(\gamma)-B']K_2}}{2K_2}$. Because the $Z(\gamma)$ function is not one-to-one, we define its set of solutions for a given γ as, $S\left(Z(\gamma)\right) = \left\{\sum_{i\in\Omega} x_i\hat{\mu}_i : \sum_{i\in\Omega} x_i = \gamma,\ x_i \text{ binary}, i \in \Omega\right\}$. By definition of γ, Constraint (6) is automatically satisfied. Then, the feasible region of **DM** and **RM**, i.e., the region where Constraint (5) is also satisfied, can be expressed as follows:

$$K_1 Z(\gamma) - K_2 \left(Z(\gamma)\right)^2 \leq B' - g(\gamma) \tag{31}$$

$$Z(\gamma) \in S\left(Z(\gamma)\right). \tag{32}$$

Then, the formulations of Sub-problems $\mathbf{DM\text{-}S(}\gamma\mathbf{)}$ and $\mathbf{RM\text{-}S(}\gamma\mathbf{)}$ given in the theorem follow if we can show the equivalence between Constraints (31)-(32) and Constraint (9). To prove this result, it is sufficient to show that any feasible solution, $Z(\gamma) \in S\left(Z(\gamma)\right)$, to Constraint (31) belongs to the region $(0, Z_1(\gamma)]$, because $Z_1(\gamma) = Z_{UB}(\gamma)$ by definition.

To this end, first observe, by Remark A.1, that Constraint (31) is satisfied if and only if $Z(\gamma) \in (-\infty, Z_1(\gamma)] \cup [Z_2(\gamma), +\infty)$. Observe also that, by definition, $Z(\gamma) = \sum_{i\in\Omega} x_i \hat{\mu}_i \in [0, 1-\hat{\mu}_0]$, for all $\gamma \geq 0$ and $\hat{\boldsymbol{\mu}} \in S(\boldsymbol{Q})$. Assume, to the contrary, that there exists a feasible solution, $Z(\gamma) \in S\left(Z(\gamma)\right)$, to Constraint (31) such that $Z(\gamma) \geq Z_2(\gamma)$, which must then be in region $[Z_2(\gamma), 1-\hat{\mu}_0]$, that is, we must have $Z_2(\gamma) \leq (1-\hat{\mu}_0)$. However, the condition that $Z_2(\gamma) \leq (1-\hat{\mu}_0) \Leftrightarrow \sqrt{\delta(\gamma)} \leq -K_1 + 2(1-\hat{\mu}_0)K_2 < 0$ for both process types (by Property A.1). Therefore, we reach a contradiction, and it must be true that $Z_2(\gamma) > (1-\hat{\mu}_0)$, and hence, there exists no feasible solution, $Z(\gamma) \in S\left(Z(\gamma)\right)$, to Constraint (31) that belongs to region $[Z_2(\gamma), 1-\hat{\mu}_0]$. Thus, we must have $Z(\gamma) \in (0, Z_1(\gamma)]$, where $Z_{UB}(\gamma) = Z_1(\gamma)$. Then, Constraint (31) can be equivalently written as: $\{Z(\gamma) \leq Z_{UB}(\gamma)\}$, showing the equivalence between Constraints (31)-(32) and Constraint (9). Further, observe that when $Z_{UB}(\gamma)$ does not exist, we have that $\delta(\gamma) < 0$, and Constraint (31) is satisfied for any $\boldsymbol{x}$. It remains to prove that $Z_{UB}(\gamma)$ is strictly decreasing in γ, which directly follows because we have that, $\frac{\partial Z_{UB}(\gamma)}{\partial \gamma} = -\frac{\partial g(\gamma)}{\partial \gamma \sqrt{\delta(\gamma)}} < 0$. The deterministic counterpart of Sub-problem $\mathbf{RM\text{-}S(}\gamma\mathbf{)}$ follows directly by Property A.2, because the case where $u = 1$ corresponds to the case where $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} = q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}$; and the case where $u = 0$ corresponds to the case where $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} = 1 - \sum_{i\in\Omega}(1-x_i)q_i^{UB}$. Thus, the deterministic counterpart formulation accounts for all possible values for $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\}$, and chooses a value $u(\boldsymbol{x})$ for each possible $\boldsymbol{x}$, that maximizes the objective function. This completes the proof. □

Proof of Lemma 1. Consider the polynomial in Remark A.1 (Eq. 18)). In what follows, we consider all possible values of $\delta(\gamma) = (K_1)^2 + 4[g(\gamma) - B']K_2$, and discuss the optimal solutions to $\mathbf{DM\text{-}S(}\gamma\mathbf{)}$ and $\mathbf{RM\text{-}S(}\gamma\mathbf{)}$ in each case. Specifically, Case 1 corresponds to the case where $\delta(\gamma) \leq 0$, and hence $Z_{UB}(\gamma)$ does not exist, and Case 2 corresponds to the case where $\delta(\gamma) > 0$, and hence $Z_{UB}(\gamma)$ exists, where either $Z_{UB}(\gamma) \geq 1-\hat{\mu}_0$ (Sub-case 2(a)), or $Z_{UB}(\gamma) < 1-\hat{\mu}_0$ (Sub-case 2(b)). Recall that $\Delta \equiv \lfloor g^{-1}\left(B' - K_1(1-\hat{\mu}_0) + K_2(1-\hat{\mu}_0)^2\right)\rfloor$, we define $\Delta' \equiv \left\lfloor g^{-1}\left(\frac{-(K_1)^2}{4K_2} + B'\right)\right\rfloor$.

Case 1: $\delta(\gamma) \leq 0$:

$$\delta(\gamma) \leq 0 \Leftrightarrow \gamma \leq \Delta' = \left\lfloor g^{-1}\left(\frac{-\left(K_1\right)^2}{4K_2} + B'\right)\right\rfloor.$$

In this case, the polynomial in Eq. (18) has either no real root, i.e., when $\gamma < g^{-1}\left(\frac{-(K_1)^2}{4K_2} + B'\right)$, or only one real root, i.e., when $\gamma = g^{-1}\left(\frac{-(K_1)^2}{4K_2} + B'\right)$, and as a result, Constraint (9) is satisfied for any $\boldsymbol{x}$ and becomes redundant. By Property A.2 we also have that, $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} = \max\left\{q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}, 1 - \sum_{i\in\Omega}\left(1-x_i\right) q_i^{UB}\right\}$. Further, to ensure that $\gamma \geq 1$, we must also have that $g^{-1}\left(\frac{-(K_1)^2}{4K_2} + B'\right) \geq 1$.

1. Recall that the objective function of $\mathbf{DM\text{-}S(}\gamma\mathbf{)}$ is given by $\left\{\text{maximize}_{\boldsymbol{x}} \sum_{i\in\Omega} x_i\hat{\mu}_i\right\}$. Therefore, when Constraint (9) is redundant, an optimal $\mathbf{DM\text{-}S(}\gamma\mathbf{)}$ solution is given by $x_i^{*D}(\gamma) = 1$, for $i \in \Omega, i = 1, 2, ..., \gamma$, and $x_j^{*D}(\gamma) = 0$, for $j \in \Omega, j = \gamma+1, ..., m$.

2. Recall that the objective function of $\mathbf{RM\text{-}S(\gamma)}$ is given by $\left\{\text{maximize}_{\boldsymbol{x}}\ \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0+\sum_{i\in\Omega}x_i\theta_i\right\}\right\}$. From Property A.2, if $q_0^{LB}+\sum_{i\in\Omega^{LB}:i\leq\gamma}q_i^{LB}\geq 1-\sum_{i\in\Omega^{UB}:i\geq\gamma+1}q_i^{UB}$, then the optimal $\mathbf{RM\text{-}S(\gamma)}$ solution is given by $x_i^{*R}(\gamma)=1$, for $i\in\Omega^{LB}, i=1,2,...,\gamma$, and $x_j^{*R}(\gamma)=0$, for $j\in\Omega^{LB}, j=\gamma+1,...,m$; and if $q_0^{LB}+\sum_{i\in\Omega^{LB}:i\leq\gamma}q_i^{LB}\leq 1-\sum_{i\in\Omega^{UB}:i\geq\gamma+1}q_i^{UB}$, then the optimal $\mathbf{RM\text{-}S(\gamma)}$ solution is given by $x_i^{*R}(\gamma)=1$, for $i\in\Omega^{UB}, i=1,2,...,\gamma$, and $x_j^{*R}(\gamma)=0$, for $j\in\Omega^{UB}, j=\gamma+1,...,m$.

Case 2: $\delta(\gamma)>0$:

$$\delta(\gamma)>0\Leftrightarrow\gamma>\Delta'=\left\lfloor g^{-1}\left(\frac{-\left(K_1\right)^2}{4K_2}+B'\right)\right\rfloor.$$

In this case, the polynomial in Eq. (18) has two real roots. Furthermore, by feasibility of an optimal $\mathbf{DM\text{-}S(\gamma)}$ and $\mathbf{RM\text{-}S(\gamma)}$ solution, we know, by Theorem 1, that $Z^{*k}(\gamma)\in(0,Z_{UB}(\gamma)], k\in\{D,R\}$. Since $\mathbf{DM\text{-}S(\gamma)}$ and $\mathbf{RM\text{-}S(\gamma)}$ are infeasible if $Z_{UB}(\gamma)<0$, we must have that:

$$0\leq Z_{UB}(\gamma)\Leftrightarrow 0\leq\frac{K_1-\sqrt{\delta(\gamma)}}{2K_2}\Leftrightarrow -K_1\leq-\sqrt{\delta(\gamma)}\Leftrightarrow\gamma\leq g^{-1}\left(B'\right).$$

We analyze two sub-cases.
Sub-case 2(a): $Z_{UB}(\gamma)\geq 1-\hat{\mu}_0$:
$Z_{UB}(\gamma)\geq 1-\hat{\mu}_0\Leftrightarrow\Delta'\leq\gamma\leq\Delta$. By definition of $\hat{\boldsymbol{\mu}}$, we have that $\sum_{i\in\Omega}\hat{\mu}_i=1-\hat{\mu}_0$, and hence $Z^{*D}(\gamma)=\sum_{i\in\Omega}x_i^{*D}\hat{\mu}_i\leq 1-\hat{\mu}_0$, and Constraint (9) becomes redundant. Then,

1. An optimal $\mathbf{DM\text{-}S(\gamma)}$ solution is given by $x_i^{*D}(\gamma)=1$, for $i\in\Omega, i=1,2,...,\gamma$, and $x_j^{*D}(\gamma)=0$, for $j\in\Omega, j=\gamma+1,...,m$.
2. From Property A.2, if $q_0^{LB}+\sum_{i\in\Omega^{LB}:i\leq\gamma}q_i^{LB}\geq 1-\sum_{i\in\Omega^{UB}:i\geq\gamma+1}q_i^{UB}$, then the optimal $\mathbf{RM\text{-}S(\gamma)}$ solution is given by $x_i^{*R}(\gamma)=1$, for $i\in\Omega^{LB}, i=1,2,...,\gamma$, and $x_j^{*R}(\gamma)=0$, for $j\in\Omega^{LB}, j=\gamma+1,...,m$; and if $q_0^{LB}+\sum_{i\in\Omega^{LB}:i\leq\gamma}q_i^{LB}\leq 1-\sum_{i\in\Omega^{UB}:i\geq\gamma+1}q_i^{UB}$, then the optimal $\mathbf{RM\text{-}S(\gamma)}$ solution is given by $x_i^{*R}(\gamma)=1$, for $i\in\Omega^{UB}, i=1,2,...,\gamma$, and $x_j^{*R}(\gamma)=0$, for $j\in\Omega^{UB}, j=\gamma+1,...,m$.

Sub-case 2(b): $Z_{UB}(\gamma)<1-\hat{\mu}_0$:
$Z_{UB}(\gamma)<1-\hat{\mu}_0\Leftrightarrow\Delta<\gamma\leq\left\lfloor g^{-1}\left(B'\right)\right\rfloor$. In this case, there exists at least one solution $\boldsymbol{x}$ that is not feasible with respect to Constraint (9), and one needs to solve $\mathbf{DM\text{-}S(\gamma)}$ and $\mathbf{RM\text{-}S(\gamma)}$ to determine their optimal solutions, $\boldsymbol{x}^{*D}(\gamma)$ and $\boldsymbol{x}^{*R}(\gamma)$. □

Proof of Theorem 2. The expression on γ_{LB} follows directly from the proof of Sub-case 2(a) in Lemma 1, which shows that for any $\gamma\leq\Delta$, Constraint (9) is satisfied for any $\boldsymbol{x}$. Therefore, for any $\gamma<\Delta$, we have that $Z^{*D}(\gamma)=\sum_{i\in\Omega:i=1}^{\gamma}\hat{\mu}_i<\sum_{i\in\Omega:i=1}^{\Delta}\hat{\mu}_i=Z^{*D}(\Delta)$, and $Z^{*R}(\gamma)=\max\left\{q_0^{LB}+\sum_{i\in\Omega^{LB}:i=1}^{\gamma}q_i^{LB},1-\sum_{i\in\Omega^{UB}:i=\gamma+1}^{m}q_i^{UB}\right\}<\max\left\{q_0^{LB}+\sum_{i\in\Omega^{LB}:i=1}^{\Delta}q_i^{LB},1-\sum_{i\in\Omega^{UB}:i=\Delta+1}^{m}q_i^{UB}\right\}=Z^{*R}(\Delta)$. Then, an optimal solution γ^{*k}, $k\in\{D,R\}$, satisfies $\gamma^{*k}\in\{\max\{0,\Delta\},...,\overline{m}\}$.
The expression on γ_{UB}^{k}, $k\in\{D,R\}$, follows because $E\left[Cost(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right]$ is strictly increasing in x_i, $i\in\Omega$ (Property 1), γ_{LB} leads to a feasible solution for both $\mathbf{DM}$ and $\mathbf{RM}$, as shown above in this proof, and γ^{*k}, $k\in\{D,R\}$, is optimal (hence feasible) for $\mathbf{DM}$ and $\mathbf{RM}$, respectively. Then, based on Constraint (5), we can write the following for $\mathbf{DM}$:

$$g\left(\gamma^{*D}\right)+K_1\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)-K_2\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)^2\leq g\left(\gamma^{*D}\right)+K_1\sum_{i\in\Omega}x_i^{*D}\hat{\mu}_i-K_2\left(\sum_{i\in\Omega}x_i^{*D}\hat{\mu}_i\right)^2\leq B'\Leftrightarrow$$

$$g\left(\gamma^{*D}\right)\leq B'-K_1\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)+K_2\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)^2\Leftrightarrow\gamma^{*D}\leq g^{-1}\left(B'-K_1\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)+K_2\left(\sum_{\substack{i\in\Omega:\\i\leq\gamma_{LB}}}\hat{\mu}_i\right)\right),$$

where the first inequality follows from Property A.1. In addition, we have the following results for **RM**:

$$g\left(\gamma^{*R}\right) + K_1\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right) - K_2\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right)^2$$

$$\le g\left(\gamma^{*R}\right) + K_1^l \sum_{i\in\Omega} x_i^{*R}\hat{\mu}_i - K_2^l\left(\sum_{i\in\Omega} x_i^{*R}\hat{\mu}_i\right)^2 \le B' \Leftrightarrow$$

$$g\left(\gamma^{*R}\right) \le B' - K_1\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right) + K_2\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right)^2 \Leftrightarrow$$

$$\gamma^{*R} \le g^{-1}\left(B' - K_1\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right) + K_2\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right)^2\right),$$

where the first inequality follows from Properties A.1 and A.2, completing the proof. □

Proof of Lemma 2. Consider $g(\sum_{\in\Omega} x_i) = c\left(\sum_{i\in\Omega} x_i\right)^n$, for some $c > 0$ and $n \in Z^+ : n \ge 2$, and therefore $g^{-1}(y) = \sqrt[n]{\frac{y}{c}}$. We first prove the result for any $y_1 \ge y_2$, y_1, $y_2 \in \mathbb{R}^+$, with superscript $k \in \{D, R\}$ added as needed to respectively denote the corresponding values for **DM** and **RM**.
Case 1. $\frac{y_1}{c} < 1 \Rightarrow \lfloor g^{-1}(y_1)\rfloor = \lfloor g^{-1}(y_2)\rfloor = 0$.
Case 2. $\frac{y_2}{c} \ge 1 \Rightarrow \frac{\partial g^{-1}(y_1)}{\partial n} = \frac{-\sqrt[n]{\frac{y_1}{c}}ln(\frac{y_1}{c})}{n^2} \le \frac{\partial g^{-1}(y_2)}{\partial n} = \frac{-\sqrt[n]{\frac{y_2}{c}}ln(\frac{y_2}{c})}{n^2} \le 0$. Thus, we have that $\lfloor g^{-1}(y_1)\rfloor$ and $\lfloor g^{-1}(y_2)\rfloor$ are non-increasing in n and that $\lfloor\sqrt[n]{\frac{y_1}{c}}\rfloor - \lfloor\sqrt[n+1]{\frac{y_1}{c}}\rfloor \ge \lfloor\sqrt[n]{\frac{y_2}{c}}\rfloor - \lfloor\sqrt[n+1]{\frac{y_2}{c}}\rfloor$.
Case 3. $\frac{y_2}{c} < 1 \le \frac{y_1}{c} \Rightarrow \lfloor g^{-1}(y_2)\rfloor = 0$, and $\lfloor\sqrt[n]{\frac{y_1}{c}}\rfloor - \lfloor\sqrt[n+1]{\frac{y_1}{c}}\rfloor \ge \lfloor\sqrt[n]{\frac{y_2}{c}}\rfloor - \lfloor\sqrt[n+1]{\frac{y_2}{c}}\rfloor = 0$.
Therefore, the result follows from the expressions in Theorem 2 on γ_{LB} and γ_{UB}^k, $k \in \{D, R\}$, by substituting $\gamma_{LB} = \lfloor g^{-1}(y_2)\rfloor$ and $\gamma_{UB}^k = \lfloor g^{-1}(y_1^k)\rfloor$, $k \in \{D, R\}$, where

$$y_1^R = B' - K_1\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right) + K_2\left(\max\left\{q_0^{LB} + \sum_{\substack{i\in\Omega^{LB}:\\ i\le\gamma_{LB}}} q_i^{LB}, 1 - \sum_{\substack{i\in\Omega^{UB}:\\ i\ge\gamma_{LB}+1}} q_i^{UB}\right\} - \mu_0\right)^2,$$

$y_1^D = B' - K_1\left(\sum_{\substack{i\in\Omega:\\ i\le\gamma_{LB}}} \hat{\mu}_i\right) + K_2\left(\sum_{\substack{i\in\Omega:\\ i\le\gamma_{LB}}} \hat{\mu}_i\right)^2$, and $y_2 = B' - K_1(1 - \hat{\mu}_0) + K_2(1 - \hat{\mu}_0)^2$, since $y_1^k \ge y_2$, for $k \in \{D, R\}$. □

Proof of Corollary 1 . Both parts follow directly from Theorem 2 and Lemma 1, which indicate that if $\overline{m} \le \gamma_{LB}$, any $\gamma \in [0, \overline{m}]$ satisfies Constraint (5), and therefore, we must have that $\gamma^{*k} = \overline{m}$, $k \in \{D, R\}$; and if $\overline{m} \ge \gamma_{UB}^k$, then $\gamma^{*k} \in \{\gamma_{LB}, \gamma_{UB}^k\}$, which automatically satisfies Constraint (6). □

Proof of Theorem 3. By Theorem 1, we have that an optimal **DM-S(**γ^{*D}**)** solution, $\boldsymbol{x}^{*D}(\gamma^{*D})$, corresponds to an optimal **DM** solution, $\boldsymbol{x}^{*D}$, and an optimal **RM-S(**γ^{*R}**)** solution, $\boldsymbol{x}^{*R}(\gamma^{*R})$, corresponds to an optimal **RM** solution, $\boldsymbol{x}^{*R}$. We next show that when $Z_{UB}(\gamma) \le Z^{*k}(\gamma - 1)$, $k \in \{D, R\}$, i.e., the stopping criterion of **Algorithm MSO**, an optimal solution is attained. This result follows because we have that, for $k \in \{D, R\}$, if $Z_{UB}(\gamma) \le Z^{*k}(\gamma - 1)$, then

$$Z^{*k}(\gamma + i) \le Z_{UB}(\gamma + i) < Z_{UB}(\gamma) \le Z^{*k}(\gamma - 1) \le Z_{UB}(\gamma - 1), \text{for all } i \in [1, \overline{m} - \gamma],$$

where the first and last inequalities follow by feasibility of an **DM-S(**γ**)** and **RM-S(**γ**)** solution, the third inequality follows by the assumed condition that $Z_{UB}(\gamma) \le Z^{*k}(\gamma - 1)$, and the second inequality follows because $Z_{UB}(\gamma)$ is strictly decreasing in γ (Theorem 1). Then, the optimality of **Algorithm MSO** directly follows by Theorem 2 and Lemma 1, because the algorithm enumerates over all values of $\gamma \in \left[\min\{\gamma_{LB}, \overline{m}\}, \min\{\gamma_{UB}^k, \overline{m}\}\right]$, $k \in \{D, R\}$ which contains the optimal number of variants in **DM**, i.e., γ^{*D}, or **RM**, i.e., γ^{*R}.
Regarding the complexity of **Algorithm MSO**, there exists an algorithm with complexity $O(mr')$ that

solves sub-problem **DM-S(γ)**, and an algorithm with complexity $O(mr)$ that solves sub-problem **RM-S(γ)** for a fixed value of $u \in \{0,1\}$, hence, the complexity of solving **RM-S(γ)** is $O(2mr)$ (see Remark 2). Then, the complexity result follows, because **Algorithm MSO** solves Sub-problem **DM-S(γ)** at most $\min\{\gamma^D_{UB}, \overline{m}\} - \min\{\gamma_{LB}, \overline{m}\}$ times, and solves Sub-problem **RM-S(γ)** at most $\min\{\gamma^R_{UB}, \overline{m}\} - \min\{\gamma_{LB}, \overline{m}\}$ times. □

Proof of Theorem 4. By Eq. (3) and Constraint (5), for any solution $\boldsymbol{x}$ that is feasible with respect to both one-tier and two-tier processes (i.e., (K_1, K_2) replaced by $\left(C_{SC}, \frac{C_{SC}}{4}\right)$ for $l = 1$, and by $\left(\frac{2C_{GS}+C_{SC}(1-\hat{\mu}_0)}{2}, \frac{2C_{GS}+C_{SC}}{4}\right)$ for $l = 2$), we have: $K_1^1 - K_1^2 = \frac{C_{SC}(1+\hat{\mu}_0)-2C_{GS}}{2}$ and $K_2^1 - K_2^2 = \frac{C_{GS}}{2}$. Then,

$$\begin{aligned} E\left[Cost^1(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] - E\left[Cost^2(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] =& (K_1^1 - K_1^2)\left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right) - (K_2^1 - K_2^2)\left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)^2 \\ =& \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)\left[(K_1^1 - K_1^2) - (K_2^1 - K_2^2)\left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)\right] \\ =& \left(\sum_{i\in\Omega} x_i\hat{\mu}_i\right)\left[\frac{C_{SC}\,(1+\hat{\mu}_0)}{2} - \frac{C_{GS}}{2}\left(2 - \sum_{i\in\Omega} x_i\hat{\mu}_i\right)\right]. \end{aligned} \tag{33}$$

Since $\sum_{i\in\Omega} x_i\hat{\mu}_i \geq 0$, $\forall \boldsymbol{x}$, we have that:

$$E\left[Cost^1(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] - E\left[Cost^2(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right] \geq 0 \Leftrightarrow \sum_{i\in\Omega} x_i\hat{\mu}_i \geq 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0). \tag{34}$$

Therefore, we have the following:

1. $\frac{C_{SC}}{C_{GS}} \geq \frac{2}{1+\hat{\mu}_0} \Leftrightarrow 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0) \leq 2 - \frac{2(1+\hat{\mu}_0)}{1+\hat{\mu}_0} = 0$, and the inequality in Eq. (34) is satisfied for any $\boldsymbol{x}$. Hence, an optimal solution of the one-tier process is always feasible for the two-tier process.
2. $\frac{C_{SC}}{C_{GS}} \leq 1 \Leftrightarrow 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0) \geq 2 - (1+\hat{\mu}_0) = 1 - \hat{\mu}_0 \geq \sum_{i\in\Omega} x_i\hat{\mu}_i$, and the inequality in Eq. (34) can not be satisfied for any $\boldsymbol{x}$. Hence, an optimal solution of the two-tier process is always feasible for the one-tier process.
3. $1 < \frac{C_{SC}}{C_{GS}} < \frac{2}{1+\hat{\mu}_0}$:
   If there exists a feasible solution $\boldsymbol{x}$ with respect to Constraints (5), (6), and (7), and with $(K_1, K_2) = \left(C_{SC}, \frac{C_{SC}}{4}\right)$, i.e., the parameters for the one-tier process, which further satisfies the condition provided in the theorem, i.e., $2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0) \leq \sum_{i\in\Omega} x_i\hat{\mu}_i$, then the optimal solution must satisfy: $\sum_{i\in\Omega} x_i^{*D1}\hat{\mu}_i \geq \sum_{i\in\Omega} x_i\hat{\mu}_i \geq 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0)$, and it follows, by Eq. (34), that $E\left[Cost^1(\boldsymbol{x}^{*D1};\hat{\boldsymbol{\mu}})\right] \geq E\left[Cost^2(\boldsymbol{x}^{*D1};\hat{\boldsymbol{\mu}})\right]$, i.e., an optimal solution of the one-tier process is feasible for the two-tier process, and hence, $\sum_{i\in\Omega} x_i^{*D2}\hat{\mu}_i \geq \sum_{i\in\Omega} x_i^{*D1}\hat{\mu}_i$, or equivalently by Property 1, $Pr\left(FN(\boldsymbol{x}^{*D2};\hat{\boldsymbol{\mu}})\right) \leq Pr\left(FN(\boldsymbol{x}^{*D1};\hat{\boldsymbol{\mu}})\right)$.

   For **RM**, if there exists a feasible solution $\boldsymbol{x}$ such that $\sum_{i\in\Omega} x_i\hat{\mu}_i \geq 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0)$, then there are two possible cases:

   (a) $\sum_{i\in\Omega} x_i^{*R1}\hat{\mu}_i \geq 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0)$. In this case, we automatically have that, $E\left[Cost^1(\boldsymbol{x}^{*R1};\hat{\boldsymbol{\mu}})\right] - E\left[Cost^2(\boldsymbol{x}^{*R1};\hat{\boldsymbol{\mu}})\right] \geq 0$.

   (b) $\sum_{i\in\Omega} x_i^{*R1}\hat{\mu}_i < 2 - \frac{C_{SC}}{C_{GS}}(1+\hat{\mu}_0)$. In this case, it follows, by Property 1, that $\boldsymbol{x}^{*R1}$ is feasible for the two-tier process.

   Therefore, by optimality of x_i^{*R2} for the two-tier process for **RM**, we have that, $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i^{*R2}\theta_i\right\} \geq \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{\theta_0 + \sum_{i\in\Omega} x_i^{*R1}\theta_i\right\}$, or equivalently, $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})}\left\{Pr\left(FN(\boldsymbol{x}^{*R2};\boldsymbol{\theta})\right)\right\} \leq$

$\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{ Pr\left(FN(\boldsymbol{x}^{*R1};\boldsymbol{\theta})\right)\right\}$, completing the proof. □

Proof of Corollary 2. The result follows by Theorem 4, because by Theorem 2 and Lemma 1, the solution $\{x_i = 1,\ i = 1, ..., \min\{\gamma_{LB}, \overline{m}\}$, and $x_j = 0,\ j = \min\{\gamma_{LB}, \overline{m}\} + 1, ..m$, is a feasible solution for the one-tier process for both **DM** and **RM**. □

Proof of Theorem 5. Since **DM** and **RM** have the same feasible region, an optimal **DM** solution is feasible for **RM**. Then, by Property A.2, the following inequalities hold, $\forall \boldsymbol{\mu} \in S(\boldsymbol{Q})$:

$$\mu_0 + \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i \geq \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\theta_i\right\} \geq \min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\theta_i\right\}, \tag{35}$$

where the first inequality follows because the expression, $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\theta_i\right\}$, is defined over all $\boldsymbol{\theta} \in S(\boldsymbol{Q})$, and we have that $\boldsymbol{\mu} \in S(\boldsymbol{Q})$; and the second inequality follows by the optimality of $\boldsymbol{x}^{*R}(\boldsymbol{\mu})$ for **RM**.

By Property A.2, $\min_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{Q})} \left\{\theta_0 + \sum_{i\in\Omega} x_i\theta_i\right\} = \max\left\{q_0^{LB} + \sum_{i\in\Omega} x_i q_i^{LB}, 1 - \sum_{i\in\Omega}(1-x_i)\, q_i^{UB}\right\}$. Thus, recalling that $q_i^{LB} = \mu_i(1 - h_i^{LB})$, and $q_i^{UB} = \mu_i(1 + h_i^{UB})$, $i \in \Omega \cup \{0\}$, the following inequalities hold (see (Eq. 35)):

$$\mu_0 + \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i \geq \mu_0(1 - h_0^{LB}) + \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB}),$$

$$\frac{1}{4}\left(1 - \mu_0 - \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i\right)^2 \leq \frac{1}{4}\left(1 - \mu_0(1 - h_0^{LB}) - \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB})\right)^2,$$

and at least one of the following inequalities should hold (see Eq. (35)):

$$\begin{aligned}
&\text{either } \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB}) \geq \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB}), \text{ or} \\
&1 - \sum_{i\in\Omega}(1 - x_i^{*R}(\boldsymbol{\mu}))\mu_i(1 + h_i^{UB}) \geq 1 - \sum_{i\in\Omega}(1 - x_i^{*D}(\boldsymbol{\mu}))\mu_i(1 + h_i^{UB}), \\
\Rightarrow\ &\text{either } (1 - h_{min}^{LB})\sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i \geq \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB}) \geq \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB}), \text{ or} \\
&(1 + h_{max}^{UB})\sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i \geq \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i(1 + h_i^{UB}) \geq \sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 + h_i^{UB}), \\
\Leftrightarrow\ &\text{either } \frac{1}{4}\left(1 - \mu_0 - \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i\right)^2 \leq \frac{1}{4}\left(1 - \mu_0 - \frac{\sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 - h_i^{LB})}{(1 - h_{min}^{LB})}\right)^2, \text{ or} \\
&\frac{1}{4}\left(1 - \mu_0 - \sum_{i\in\Omega} x_i^{*R}(\boldsymbol{\mu})\mu_i\right)^2 \leq \frac{1}{4}\left(1 - \mu_0 - \frac{\sum_{i\in\Omega} x_i^{*D}(\boldsymbol{\mu})\mu_i(1 + h_i^{UB})}{(1 + h_{max}^{UB})}\right)^2.
\end{aligned}$$

This completes the proof. □

E Case Study Details

E.1 Construction of the Uncertainty Set for RM

To construct the uncertainty set $S(\boldsymbol{Q}) = \left\{\boldsymbol{\theta} : \theta_i \in [q_i^{LB}, q_i^{UB}], i \in \Omega \cup \{0\}, \sum_{i\in\Omega\cup\{0\}} \theta_i = 1\right\}$ for the robust model **RM**, we first derive the bounds for population-level prevalences, q_i^{LB} and $q_i^{UB}, i \in \Omega \cup \{0\}$, from their group-level counterparts, $q_i^{r,LB}$ and $q_i^{r,UB}, r \in R$. We do this via the Wilson Score Method with continuity correction (Newcombe (1998)). In particular, following Newcombe (1998), we derive, for each $i \in \Omega \cup \{0\}$ and $r \in R$:

$$\begin{aligned} q_i^{r,LB} =& \frac{2n^r\hat{\mu}_i^r + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n^r} + 4\hat{\mu}_i^r(n^r(1-\hat{\mu}_i^r)+1)}}{2(n^r+z^2)}, \text{ and} \\ q_i^{r,UB} =& \frac{2n^r\hat{\mu}_i^r + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n^r} + 4\hat{\mu}_i^r(n^r(1-\hat{\mu}_i^r)-1)}}{2(n^r+z^2)}, \end{aligned}$$

where z is the 95th percentile of the standard normal distribution, and n^r is the sample size for group r based on the data in Schrijver et al. (2016), leading to:

$$q_i^{LB} = \sum_{r\in R} f^r q_i^{r,LB} \text{ and } q_i^{UB} = \sum_{r\in R} f^r q_i^{r,UB}, i \in \Omega \cup \{0\},$$

that is, the weighted average of lower and upper limits.

E.2 Estimated Variant Prevalences for New York

Table 9: Point Prevalences, Lower Bounds, Upper Bounds, and Rankings for the top 85 CF-causing variants for NY, and the combined prevalence of the remaining variants (based on Schrijver et al. (2016))

Variant Ranking-Name	Point Prevalence (NY)	Lower Bound	Upper Bound
Mutation-free	0.8780172	0.8739933	0.8819137
1-F508del	0.0832589	0.0796051	0.0873500
2-G542X	0.0034355	0.0026819	0.0046990
3-G551D	0.0026506	0.0021655	0.0036606
4-R117H;5T	0.0018443	0.0014586	0.0027651
5-N1303K	0.0016919	0.0012453	0.0026652
6-W1282X	0.0015948	0.0012282	0.0025002
7-3849+10kbC->T	0.0012358	0.0008256	0.0021764
8-R553X	0.0011630	0.0008028	0.0020509
9-3120+1G->A	0.0010138	0.0004909	0.0020367
10-1717-1G->A	0.0010018	0.0007026	0.0018403
11-621+1G->T	0.0009916	0.0006921	0.0018282
12-2789+5G->A	0.0007746	0.0005303	0.0015672
13-I507del	0.0006976	0.0003958	0.0015353
14-R334W	0.0006232	0.0003247	0.0014585
15-G85E	0.0005215	0.0003031	0.0012869
16-R1162X	0.0004853	0.0002372	0.0012685
17-1898+1G->A	0.0004007	0.0002642	0.0010998
18-3876delA	0.0003973	0.0001866	0.0011562
19-3659delC	0.0003858	0.0002677	0.0010734
20-S549N	0.0003517	0.0001226	0.0011144
21-3272-26A->G	0.0003427	0.0001903	0.0010489
22-R560T	0.0003151	0.0001873	0.0010033

Variant Ranking-Name	Point Prevalence (NY)	Lower Bound	Upper Bound
23-L206W	0.0003054	0.0001448	0.0010195
24-2184insA	0.0003021	0.0001891	0.0009806
25-R347P	0.0002936	0.0001914	0.0009656
26-A455E	0.0002936	0.0001914	0.0009656
27-Q493X	0.0002936	0.0001914	0.0009656
28-2183AA->G	0.0002588	0.0001266	0.0009489
29-R1066C	0.0002282	0.0000735	0.0009303
30-W1089X	0.0002109	0.0000704	0.0009047
31-406-1G->A	0.0002109	0.0000704	0.0009047
32-394delTT	0.0002014	0.0001180	0.0008551
33-E60X	0.0002014	0.0001180	0.0008551
34-2184delA	0.0002014	0.0001180	0.0008551
35-P67L	0.0002014	0.0001180	0.0008551
36-1154insTC	0.0002014	0.0001180	0.0008551
37-3905insT	0.0002014	0.0001180	0.0008551
38-R347H	0.0002014	0.0001180	0.0008551
39-S945L	0.0001900	0.0000671	0.0008691
40-A559T	0.0001892	0.0000361	0.0008850
41-2307insA	0.0001762	0.0000308	0.0008653
42-R75X	0.0001728	0.0000437	0.0008531
43-2105-2117del13insAGAAA	0.0001643	0.0000461	0.0008382
44-1811+1643G->T	0.0001643	4.61E-05	8.38E-04
45-Y1092X	0.0001518	5.02E-05	8.11E-04
46-R1158X	0.0001439	4.60E-05	7.99E-04
47-3120G->A	0.0001439	4.60E-05	7.99E-04
48-711+1G->T	0.0001434	5.25E-05	7.96E-04
49-R352Q	0.0001434	5.25E-05	7.96E-04
50-W1204X	0.0001410	3.51E-05	8.04E-04
51-663delT	0.0001410	3.51E-05	8.04E-04
52-2055del9->A	0.0001410	3.51E-05	8.04E-04
53-935delA	0.0001177	2.53E-05	7.69E-04
54-1288insTA	0.0001177	2.53E-05	7.69E-04
55-M1101K	0.0001093	5.01E-05	7.40E-04
56-V520F	0.0001093	5.01E-05	7.40E-04
57-R117C	0.0001093	5.01E-05	7.40E-04
58-2622+1G->A	0.0001093	5.01E-05	7.40E-04
59-1811+1634A->G or 1811+1.6kbA->G	0.0001029	1.46E-05	7.48E-04
60-Q890X	0.0000944	1.69E-05	7.33E-04
61-712-1G->T	0.0000944	1.69E-05	7.33E-04
62-G330X	0.0000782	4.50E-06	7.10E-04
63-S466X	0.0000717	4.10E-06	7.00E-04
64-S1255X	0.0000717	4.10E-06	7.00E-04
65-1078delT	0.0000711	1.06E-05	6.97E-04
66-P205S	0.0000711	1.06E-05	6.97E-04
67-1248+1G->A	0.0000711	1.06E-05	6.97E-04

Variant Ranking-Name	Point Prevalence (NY)	Lower Bound	Upper Bound
68-H199Y	0.0000711	1.06E-05	6.97E-04
69-F311L	0.0000711	1.06E-05	6.97E-04
70-H609R	0.0000711	1.06E-05	6.97E-04
71-3271delGG	0.0000711	1.06E-05	6.97E-04
72-1812-1G->A	0.0000651	3.98E-06	6.88E-04
73-3791delC	0.0000651	3.98E-06	6.88E-04
74-R1066H	0.0000586	4.21E-06	6.77E-04
75-L467P	0.0000521	4.83E-06	6.66E-04
76-Q98X	0.0000521	4.83E-06	6.66E-04
77-Y913X	0.0000521	4.83E-06	6.66E-04
78-405+3A->C	0.0000521	4.83E-06	6.66E-04
79-E585X	0.0000455	5.94E-06	6.55E-04
80-444delA	0.0000455	5.94E-06	6.55E-04
81-3500-2A->G	0.0000455	5.94E-06	6.55E-04
82-1548delG	0.0000455	5.94E-06	6.55E-04
83-S549R	0.0000371	8.27E-06	6.40E-04
84-Y122X	0.0000371	8.27E-06	6.40E-04
85-3849+4A->G	0.0000371	8.27E-06	6.40E-04
$\sum_{i\in\Omega:i=86}^{312}\hat{\mu}_i$	0.0084174		