Novel Pooling Strategies for Genetic Testing, with Application to Newborn Screening

Hussein El Hajj[1]*, Douglas R. Bish[2], Ebru K. Bish[2], Denise M. Kay[3]

[1] Department of Management Sciences, University of Waterloo
Waterloo, ON N2L 3G1, Canada

[2] Department of Information Systems, Statistics, and Management Science, University of Alabama
Tuscaloosa, AL 35487, United States

[3] Wadsworth Center, New York State Department of Health
Albany, NY 12208, United States

August 2020; Revised: July 2021; Accepted: September 2021

Abstract

Newborn screening (NBS) is a state-level initiative that detects life-threatening genetic disorders for which early treatment can substantially improve health outcomes. Cystic fibrosis (CF) is among the most prevalent disorders in NBS. CF can be caused by a large number of mutation *variants* to the CFTR gene. Most states use a multi-test CF screening process that includes a genetic test (*DNA*). However, due to cost concerns, *DNA* is used only on a small subset of newborns (based on a low-cost biomarker test with low classification accuracy), and only for a small subset of CF-causing variants.

To overcome the cost barriers of expanded genetic testing, we explore a novel approach, of multi-panel pooled *DNA* testing. This approach leads not only to a novel optimization problem (variant selection for screening, variant partition into multi-panels, and pool size determination for each panel), but also to novel CF NBS processes. We establish key structural properties of optimal multi-panel pooled *DNA* designs; develop a methodology that generates a *family* of optimal designs at different costs; and characterize the conditions under which a 1-panel versus a multi-panel design is optimal. This methodology can assist decision-makers to design a screening process, considering the cost versus accuracy trade-off. Our case study, based on published CF NBS data from the state of New York, indicates that the multi-panel and pooling aspects of genetic testing work synergistically, and the proposed NBS processes have the potential to substantially improve both the efficiency and accuracy of current practices.

Keywords: Genetic testing, newborn screening, resource allocation, pooled testing, partition problem

*Corresponding Author: *hme35@vt.edu*

1 Introduction and Motivation

Newborn screening (NBS) is a state-level initiative that routinely screens newborns for life-threatening genetic disorders for which early treatment can substantially improve health outcomes. NBS has saved thousands of newborns from disability and death in the United States (US) [44]. While the Advisory Committee on Heritable Disorders in Newborns and Children recommends thirty-five genetic disorders for NBS [3], the number of disorders included in a state's NBS program can be less [8], as cost is a major barrier to the inclusion of disorders, e.g., [16, 44, 68, 71, 78, 85, 86]. Thus, screening must be both accurate and efficient. NBS is performed via laboratory tests on dried blood spots routinely collected from newborns for this purpose.

One of the most prevalent NBS disorders is cystic fibrosis (CF), which is caused by harmful mutations to the CFTR gene [14, 54]. Currently, there are 352 well-characterized CF-causing *variants* (specific types of mutations), most of them very rare [18, 80]. Like most NBS disorders, CF is a recessive disorder; everyone has two copies of the CFTR gene, one inherited from each parent, and to have CF a newborn must inherit a mutation, of any CF-causing variant, from *each* parent. That is, a CF-positive newborn has two CF-causing mutations, one on each CFTR gene. Newborns that inherit only one CF-causing mutation are asymptomatic CF-carriers (the type of variant does not alter the newborn's CF-positivity nor carrier status). CF manifests primarily with respiratory and digestive symptoms, and is potentially life-threatening [33], but starting treatment before symptoms appear can reduce hospitalization and complications [2, 12, 17, 39, 45]. In 2004, the Centers for Disease Control and Prevention recommended CF for NBS [45], and since 2009 every state's NBS program has included CF screening [25].

NBS for CF is accomplished via a *screening process*, i.e., a sequence of tests and decision rules, that classifies newborns as *screen-negative* or *screen-positive* as accurately as possible, under limited resources. The two most commonly used screening tests, both of which are performed on the dried blood spots routinely collected from the newborn and sent to the NBS laboratory, include:
Immunoreactive trypsinogen (IRT) test is a biomarker test that measures *IRT* levels, which are generally elevated in CF-positive newborns [22]. However, the range of *IRT* levels for CF-positive and CF-negative (mutation-free or CF-carrier) groups overlap [58, 77], resulting in relatively low classification accuracy, e.g., depending on the decision rule (threshold), sensitivity and specificity vary between 85-97% and 95-99%, respectively [24, 45, 53, 55, 62].

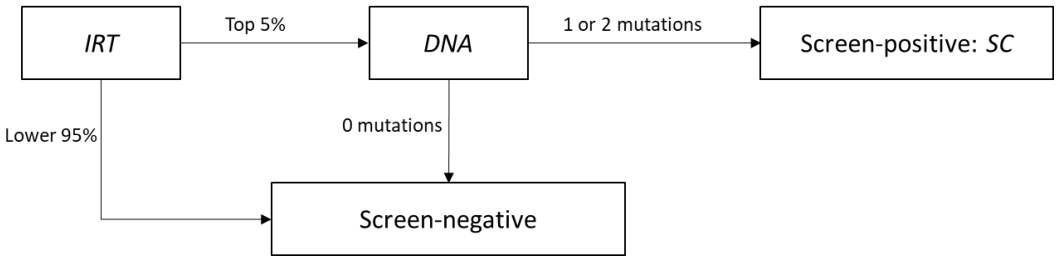
Deoxyribonucleic acid test (DNA) is a genetic test that typically uses variant-specific molecular probes, which bind to the variant, amplify it via polymerase chain reaction (PCR), and provide a signal if binding occurs (implying the presence of the specific variant) [84].[1] Probe-based *DNA* has a technological limit on the number of variants included in the *panel* (i.e., set of variants searched for), e.g., a panel limit of around 90 variants is common [64]. *DNA* has almost perfect *analytical sensitivity*, that is, it detects the variants in its panel with almost perfect reliability [50, 51, 52, 59]. However, its *clinical sensitivity*, that is, the likelihood that it will detect a CF-positive subject, is dependent on the variants included in its panel. Because there is a large number of variants, adding more variants to the panel improves the test's clinical sensitivity, but comes at the expense of a higher testing cost, and is restricted by the technological limit on panel size. Consequently, for the practical case when the *DNA* panel does not contain all CF-causing variants, its *clinical sensitivity* will be less than perfect. Thus, it is the test's clinical sensitivity that impacts the likelihood of a missed CF case, and we explicitly model this accuracy versus cost trade-off.

While each state designs its own CF screening process, all states use *IRT* as the first test. The most commonly used process is the *IRT*/*DNA* process (Fig. 1), in which the *IRT* test is followed by a *DNA* test for newborns with *IRT* levels exceeding a given threshold. Fig. 1 depicts the commonly used 5% *daily IRT threshold* (§4.2.2): newborns with *IRT* levels in the top 5%, of all *IRT* levels for each testing day, undergo *DNA* testing [53]. For *DNA*, newborns with at least one mutation detected (of any CF-causing variant) are typically classified as *screen-positive*, and all other newborns are classified as *screen-negative* (e.g., [53, 70, 89]). A *false-negative* occurs when a CF-positive newborn is classified as screen-negative, and a *false-positive* occurs when a CF-negative newborn is classified as screen-positive. For example, for the *IRT*/*DNA* process, a false-negative occurs when either the CF-positive newborn's *IRT* level is below the *IRT* threshold, or the *DNA* panel does not include any of their specific mutation variants (i.e., it has less-than-perfect clinical sensitivity). On the other hand, a false-positive occurs when the newborn is a CF-carrier, whose single mutation variant is included in the *DNA* panel (hence detected due to the test's perfect analytical sensitivity), and whose *IRT* level is higher than or equal to the *IRT* threshold.

By the FDA guidelines, "genetic testing is not intended for stand-alone diagnostic" [40]. Consequently, all current processes end with the referral of the screen-positive newborns for diagnostic

[1] The probe-based *DNA* test differs from next-generation sequencing, which is another genetic testing technology that determines the genetic code for larger sections of the CFTR gene, and then known variants are detected via bioinformatics [9]. However, the probe-based *DNA* tests are less expensive, and require shorter processing times than next-generation sequencing tests. More importantly for our purposes, the *DNA* tests allow for pooled (group) testing, whereas next-generation sequencing tests, for technical reasons, do not.

Figure 1: The commonly used *IRT*/*DNA* process for CF NBS



testing, conducted via the gold standard *Sweat Chloride (SC)* test in all US states, e.g., [26, 38, 66]. *SC* has perfect sensitivity and specificity, and thus, correctly classifies the newborns, fixing any false-positives from NBS. However, an *SC* referral of a false-positive comes at the expense of unnecessary testing [88], which requires the newborn to be taken to a specialized testing facility, potentially causing parental anxiety and out-of-pocket costs (e.g., travel, missed work) [26, 83, 87]. On the other hand, any CF-positive newborn missed in the screening process (i.e., a false-negative) is not referred to *SC*, leading to late diagnosis, which often results in poor health outcomes. Consequently, maximizing the accuracy of CF NBS becomes equivalent to minimizing the false-negatives; and the testing process is naturally constrained by testing cost considerations, and the unnecessary testing (*DNA* and/or *SC*) of false-positives contribute to the testing cost.

The common *IRT*/*DNA* process suffers from two major drawbacks: **(1)** *IRT* test leads to some false-negatives (missed CF cases); for example, between 2007-2012, all false-negatives reported for New York's CF NBS program, and half of all false-negatives reported for California's CF NBS program, stemmed from the *IRT* test [53, 55]; and **(2)** *DNA* testing is expensive, and has a technological limit on panel size; as a result, variant selection for *DNA* revolves around the trade-off between clinical sensitivity and cost, e.g., see [35], further the cost of *DNA* limits the number of newborns that can be genetically tested.

With the goal of improving current CF NBS practices, we explore a **novel approach** of pooled *DNA*, which we refer to as *P-DNA*. Under *P-DNA*, we use Dorfman pooling ([31]), where samples from multiple subjects (extracted from their dried blood spots) are combined into a single testing pool and tested with one *DNA* test. If the pool tests positive (i.e., at least one variant in the panel is detected), then each subject is individually tested (with an additional *DNA* test per subject, using a new sample, extracted from the same dried blood spot) to identify subject(s) with CF-causing variants; and if the pool tests negative, then all subjects in the pool are classified as screen-negative. Pooling can substantially reduce the number of *DNA* tests required for NBS over

the current individual *DNA* testing paradigm. While Dorfman pooling is used in various health screening contexts (e.g., [6, 67], and the references therein), NBS introduces a *new application area* and a *new pooling design problem*. First, panel composition (i.e., variant set) and pool size (i.e., the number of subjects tested in each pool) must be determined jointly, because panel composition impacts the optimal pool size: adding more variants to the panel increases the clinical sensitivity of *DNA* testing, which is a driver of the optimal pool size [7]. We also explore another novel idea, of using multiple panels, where each panel has its own pool size, and requires its own *DNA* test, leading to the decision of how to optimally partition the selected variant set into multiple panels. Specifically, in the case of a multi-panel *P-DNA*, each panel has its own testing pool, and each newborn's dried blood spot (through samples extracted) is tested, in parallel, in every panel's pool, followed by the individual testing of those newborns (through additional samples) in any positive-testing pool according to Dorfman pooling. Then, a newborn is classified as screen-negative only if no mutations are detected in any of the panels, and is classified as screen-positive otherwise, see Fig. 2. The multi-panel approach allows the testing of more variants, in numbers comparable to some next-generation sequencing tests (thus increasing the clinical sensitivity of *DNA*), but importantly, as we show in this paper, it can increase efficiency substantially when used in conjunction with pooling, beyond the efficiency gains of pooling alone. Therefore, the *P-DNA* design problem incorporates this variant partition decision into variant selection (using only one panel is thus a special case).

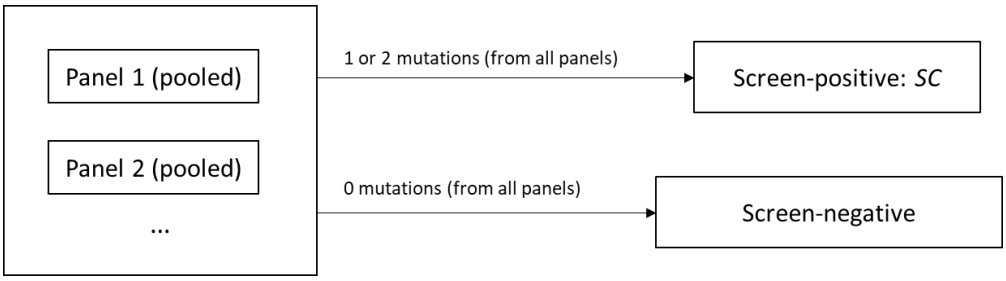
This research falls under the general umbrella of public health screening and its intersection with operations research (OR), e.g., see [15, 42] for references on the application of OR and statistics models to disease screening. In the following, we provide an overview of the related literature. Given the vast literature, our discussion is not exhaustive, but rather indicative of how this paper complements, and contributes to the literature. Literature that is the most closely related to this work is also cited throughout the paper. Because genetic testing is performed *in vitro*, i.e., on specimens (in our setting, dried blood spots from newborns), pooled testing is viable, and this aspect differentiates this research from studies that focus on *in vivo* testing, i.e., directly on the subject. From this perspective, this work is closer to donated blood screening, e.g., [34], infectious disease screening, e.g., [5, 37, 49, 67], *in vitro* cancer screening, e.g., [47, 61, 65, 81], among others. However, the unique characteristics of CF and genetic testing give rise to a novel decision problem in our setting, i.e., the integrated variant selection, variant partition, and pool size determination problem discussed above.

Regarding the CF NBS literature that examines screening processes, a majority of these studies are in the domain of medical/health systems literature; see [35, 36] for a more detailed overview, and §4.2.1 for additional literature about current CF NBS practices. Many of these papers use descriptive analyses based on historical data, e.g., [23, 48, 55], with only a few studies that utilize predictive analyses to compare a small number of given CF NBS processes, e.g., [71, 87]. To our knowledge, there are only two papers that use prescriptive analyses to optimize different parts of CF NBS; [35, 36] both design a single-panel *DNA* for individual testing (i.e., without pooling), that is, following the current CF testing paradigm, given a testing budget, and focusing, respectively, on the robustness and equity aspects of screening. The partitioning of variants into multiple panels, and pooling aspects of this work are unique elements that differentiate it from this previous research.

Much of the pooled testing literature studies Dorfman pooling and its variations, mostly under a deterministic and known prevalence rate; only a few papers consider a stochastic rate with a given distribution, see the references in [7]. Other pooling methods, such as array pooling, are also explored, e.g., [56, 72]. Pooled testing is one component of our decision problem, which has unique characteristics in our setting, including a pool prevalence that depends on variant selection and partition components, and a limited testing budget. We consider Dorfman pooling because it is relatively easy to implement, yet highly efficient, and hence is adopted in many screening applications, including screening of donated blood, sexually-transmitted diseases and other infectious diseases, and many others, e.g., [1, 4, 30, 57, 63, 73, 74]. Thus, Dorfman pooling is a viable first step for CF NBS, where pooling is currently not utilized. We also briefly discuss *P-DNA* under array pooling, but the research question, of which pooling method to use in different settings, is beyond the scope of this paper.

The contributions of this paper are multi-fold. From a *modeling* perspective, we introduce a novel decision problem that arises not only in NBS, but in genetic testing in general. From a *theoretical* perspective, we establish key structural properties of optimal *P-DNA* designs (variant selection, variant partition into multiple panels, pool size determination). The unique pooling and multi-panel dimensions of this work give rise to a new decision problem in genetic testing, contributing to the small number of mathematical models on genetic testing, including [35, 36], both of which design a single-panel *DNA* under individual testing and for a given testing budget, as discussed above. The structural properties of optimal designs lead to an efficient methodology for generating a *family* of optimal *P-DNA* designs, along with their corresponding budgets (i.e., *budget breakpoints*, which are complex functions of the overall process design and parameters).

Figure 2: Representative *P-DNA* process for CF NBS



We also characterize the conditions under which a 1-panel versus a multi-panel *P-DNA* design is optimal. From a *practical* perspective, the family of designs generated by our methodology allows the decision-maker to design an optimal *P-DNA*, considering the trade-off between clinical sensitivity and cost. Further, the increased efficiency in the genetic testing component, made possible by the reduced cost per newborn of *P-DNA*, allows for improved CF NBS processes. The case study, which uses published CF NBS data from the state of New York [53], indicates that the multi-panel and pooling aspects of genetic testing work synergistically, and the proposed NBS processes have the potential to substantially improve both the efficiency and accuracy of current practices. These findings have important implications on NBS practices, for instance, our modeling framework and methodology can be used by state laboratories to make decisions on new screening processes, e.g., testing platforms, testing kits, and/or subcontracting decisions, and by test manufacturers to develop new, innovative testing kits.

The remainder of this paper is organized as follows. §2 presents the notation, assumptions, decision problem, and some preliminaries. §3 derives key structural properties of optimal *P-DNA* designs, which lead to an efficient, optimization-based methodology for generating a family of optimal *P-DNA* designs and their corresponding budgets. §4 performs a case study and applies the new methodology for *P-DNA* design to published CF NBS data from the state of New York. Finally, §5 provides a summary of the main takeaways from this work, along with the limitations of the study, which motivate future research directions. To facilitate the presentation, Appendix A provides a summary of the acronyms and the mathematical notation, Appendix B includes many of the derivations, and Appendices C-E include supplementary results and the mathematical proofs.

2 The Decision Problem and Preliminaries

We first provide the notation, assumptions, and the decision problem (§2.1), followed by some preliminaries that will be used in the subsequent analysis (§2.2).

2.1 Notation, and Assumptions, and Decision Problem

Let Ω denote the set of CF-causing variants, with cardinality m (currently $m = 352$) and variant frequency vector $\boldsymbol{q} = (q_i)_{i\in\Omega}$, with $q_i, i \in \Omega$, denoting the conditional probability that a CFTR gene has a mutation of variant i, given that the gene has a mutation, thus $\sum_{i\in\Omega} q_i = 1$. Because each subject has two CFTR genes, the probability that a CF-positive subject (i.e., with one mutation in each CFTR gene) has two mutations of variant $i \in \Omega$ is q_i^2, and one mutation of variant i and one mutation of variant j, $i, j \in \Omega, i \neq j$, is given by $2q_iq_j$. We define random variable N as the number of mutations (of any variant in set Ω) a random newborn has, with sample space $\mathcal{S}(N) = \{0, 1, 2\}$, respectively denoting the newborn's status as mutation-free, CF-carrier, and CF-positive, and with probability mass function (pmf), $P_N(n) = Pr(N = n), n \in \mathcal{S}(N)$. Without loss of generality, we arrange the variants in set Ω such that $q_i \geq q_j, \forall i, j \in \Omega, i < j$. In the remainder of the paper, we refer to an individual *DNA* test simply as *DNA*, to pooled-*DNA* as *P-DNA*, and use the term genetic testing to refer to both.

We denote vectors in boldface. Let $\mathbf{0}$ denote the m-dimensional 0 vector, and $\mathbf{1}_l, l \in Z^+, l \leq m$, denote the m-dimensional binary vector having a value of 1 for the first l elements, and a value of 0 for the remaining m-l elements. We use the subscript i to refer to the variants in set Ω, the superscript k to refer to each *P-DNA* panel (Fig. 2), and omit the indices when it is clear from context, or when a result or an expression applies independently of the related index.

The goal of the *P-DNA* **design problem** is to minimize the probability of a false-negative classification (FN) by making three inter-dependent decisions: **(i)** selecting the *set of variants* for screening; **(ii)** *partitioning* the selected variants into panels; and **(iii)** selecting the *pool size* for each panel. This design problem has technological limitations on panel size, $\overline{z}$, pool size, $\overline{t}$, and number of panels used, $\overline{\eta}$. Then, the decision variables include, for each panel $k = 1, \cdots, \overline{\eta}$, the binary vector $\boldsymbol{x}^k = \left(x_i^k\right)_{i\in\Omega}$, where $x_i^k = 1$ if variant i is included in panel k, and $x_i^k = 0$ otherwise (equivalently represented by set $S(\boldsymbol{x}^k) \equiv \{i \in \Omega : x_i^k = 1\}$); and pool size $t^k \in Z^+$. Observe that panel $k, k = 1, \cdots, \overline{\eta}$, is empty (i.e., not used) if $\boldsymbol{x}^k = \mathbf{0}$. Further, there is a budget restriction, B, on the total cost of genetic and diagnostic (SC) testing per newborn. We use the budget, B,

as a modeling convenience; in reality, this is also a decision (§1), therefore, we develop an efficient methodology for generating a family of optimal *P-DNA* designs for a range of budget values, to allow the decision-maker to make trade-offs between budget and accuracy, i.e., the costs of testing and false-negatives.

Define the binary vector $\boldsymbol{x}^{12\cdots\overline{\eta}} \equiv \sum_{k=1}^{\eta} \boldsymbol{x}^k$, and set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) \equiv \{i \in \Omega : x_i^{12\cdots\overline{\eta}} = 1\} = \cup_{k=1}^{\eta} S(\boldsymbol{x}^k)$, i.e., respectively the combined variant vector and set covered by *P-DNA*. As will become clear in the sequel, it is informative to consider the three components of the *P-DNA* design problem: the *variant selection* component determines set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, and the *variant partition* component splits the selected variant set into vectors $(\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}$ (equivalently, sets $S(\boldsymbol{x}^k), k = 1, \cdots, \overline{\eta}$), which, in turn, impact pool sizes $(t^k)_{k=1,\cdots,\overline{\eta}}$. For a given panel vector $\boldsymbol{x}$, we denote the *panel size* by $z(\boldsymbol{x}) \equiv \sum_{i\in\Omega} x_i$ and *panel coverage* by $y(\boldsymbol{x}) \equiv \sum_{i\in\Omega} x_i q_i$ (with superscript $k = 1, \cdots, \overline{\eta}$ added for panel index as needed); and for a given variant partition, we let $\eta((\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}) \equiv \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}}$ denote the *number of panels used*, where the indicator variable $I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}} = 1$ if $\boldsymbol{x}^k > \boldsymbol{0}$, and $I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}} = 0$ otherwise, $k = 1, \cdots, \overline{\eta}$. To simplify the notation, in places we drop the arguments in parantheses when clear from context. Finally, without loss of generality, we relabel the panels in any variant partition such that panels $1, \cdots, \eta$ are non-empty, and panels $\eta + 1, \cdots, \overline{\eta}$ are empty. We refer to a variant partition that uses exactly η panels as an η-panel design or an η-partition.

At the conclusion of genetic testing, newborns who do not have any variants in set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ are classified as screen-negative, while newborns with at least one variant in set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ are referred for the diagnostic *SC* test, and diagnosed with CF or not according to the outcome (Fig. 2). Specifically, *SC* distinguishes between CF-positives, who have two mutations, and CF-carriers, who have a single mutation (independent of which variant(s)). On the other hand, if a CF-positive newborn's specific variants are not included in set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, a false-negative occurs, and these missed CF cases are a primary concern in NBS.

We make the following assumptions, which we briefly discuss here; §5 provides a more thorough discussion.

Assumption ($\boldsymbol{A_1}$). *All CF-causing variants are known and included in set Ω.*

Assumption ($\boldsymbol{A_2}$). *Both the genetic test, DNA, and the diagnostic test, SC, are perfectly reliable, i.e., the DNA analytical sensitivity (the probability of detecting a variant in the panel) and specificity (the true negative probability) are 1 for each variant included in its panel(s), and SC can identify all CF-positive and CF-negative subjects accurately.*

Assumption ($\boldsymbol{A_3}$)**.** *Pooling does not alter the analytical sensitivity of P-DNA.*

Assumption ($\boldsymbol{A_1}$) is a practical assumption; there are additional variants that are suspected of being CF-causing, and undoubtedly, more unknown variants, but the set of 352 variants [18] is well-characterized, that is, there is high confidence that these variants are CF-causing. We expect this set to expand as more variants are discovered and characterized, but any variant that has not yet been identified in CF-positive individuals is likely to be very rare. Assumption ($\boldsymbol{A_2}$) is supported by various clinical studies on CF genetic testing, which indicate that the analytical sensitivity and specificity of the *DNA* test is almost perfect for variants included in its panel (e.g., [50, 51, 52, 59]); and the *SC* is considered the gold standard for CF diagnosis [26]. However, the clinical sensitivity of the *DNA* test is naturally less than 1 when its panel does not include all CF-causing variants, i.e., the clinical sensitivity is a function of the variant selection decision, and a less-than-perfect clinical sensitivity is the main driver of a missed CF case for the *DNA* test. For Assumption ($\boldsymbol{A_3}$), pooling might lower the analytical sensitivity for large pool sizes, due to dilution of positive samples with negative samples in the pool, e.g., [32, 69, 76]. Thus, while Assumption ($\boldsymbol{A_3}$) is supported up to certain pool sizes (which we model via the pool size limit, $\bar{t}$) in other contexts [60], this assumption would need to be validated for CF NBS, as we discuss in §5. Pooling does not affect the test's specificity.

On the cost side, each *SC* test incurs a cost of c_{SC}, and each genetic test incurs a fixed cost (e.g., consumables, labor) of c_f, plus a variable cost (e.g., dNTPs, enzymes, variant-specific reagents) of $c_v\left(z(\boldsymbol{x})\right)$, which is non-decreasing in panel size, $z(\boldsymbol{x})$; we make no other assumptions on the functional form of $c_v(.)$. Thus, $c_v(.)$ depends only on panel size $z(\boldsymbol{x})$, and not on the specific variants in the panel. This is a good assumption for probe-based genetic testing technologies (see §1), where each variant has a unique probe (i.e., a unique sequence of *DNA*), but all probes are similar in structure, i.e., a segment of *DNA* and a signaling molecule. Besides the probes, which contribute to a linear form, the test has other reagents (e.g., dNTPs, enzymes), and the amount of these reagents (hence cost) may increase in a concave manner. As a result, in practice we expect $c_v(.)$ to be either (approximately) linear or concave increasing in the number of variants.

2.2 Preliminaries

We provide expressions on the expected testing cost and false-negative probability as a function of panel composition $\boldsymbol{x}$, and pool size t (these expressions expand their counterparts in [7] and [35] to the *P-DNA* scheme); see Appendix B for all derivations in this section.

To this end, let $p(\boldsymbol{x})$ denote the probability that a random newborn has at least one mutation, of any variant covered by the panel, and $D(\boldsymbol{x}, t)$ and $C(\boldsymbol{x}, t)$ respectively denote the per newborn expected number of tests and per newborn expected cost for *P-DNA*. We can write:

$$p(\boldsymbol{x}) = P_N(1)\ y(\boldsymbol{x}) + P_N(2)\ \left(2y(\boldsymbol{x}) - (y(\boldsymbol{x}))^2\right), \quad \forall \boldsymbol{x} \geq \mathbf{0} \tag{1}$$

$$D(\boldsymbol{x}, t) = \frac{1}{t} + 1 - (1 - p(\boldsymbol{x}))^t, \qquad \forall \boldsymbol{x} > \mathbf{0}, \forall t \geq 2 \tag{2}$$

$$C(\boldsymbol{x}, t) = D(\boldsymbol{x}, t) \times (c_f + c_v\,(z(\boldsymbol{x}))) \qquad \forall \boldsymbol{x} > \mathbf{0}, \forall t \geq 2. \tag{3}$$

Observe that Eq. (2) follows because one test for t subjects suffices if the pooled test's outcome is negative (i.e., no subject in the pool has any variant covered by the panel), and t additional individual tests (one per each subject in the pool) are needed if the pooled test's outcome is positive (i.e., at least one subject in the pool has a variant covered by the panel), see Assumptions ($\boldsymbol{A_1}$)-($\boldsymbol{A_2}$). (Note that for the individual *DNA*, $D(\boldsymbol{x}) = 1$, and pool size is irrelevant.) Then, the expected total testing cost per newborn, denoted by $TC(.)$, is the *SC* cost multiplied by the probability that a newborn is referred for *SC*, plus the cost of *P-DNA*:

$$TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}}) = c_{SC} \times p(\boldsymbol{x}^{12\cdots\overline{\eta}}) + \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k > \mathbf{0}\}} \times C(\boldsymbol{x}^k, t^k), \tag{4}$$

As discussed above, the primary concern for NBS is to minimize the probability of a false-negative (FN). An FN occurs if the newborn is CF-positive (i.e., has two mutations) and no mutations are detected by genetic testing (i.e., when the specific variants of the CF-positive newborn are not covered by any panel, or equivalently by set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, see Assumptions ($\boldsymbol{A_1}$)-($\boldsymbol{A_3}$)). This follows because if at least one mutation is found in *any* panel, then the newborn will undergo *SC* testing, which eliminates an FN (Fig. 1), leading to the following expression:

$$Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}})) = \left(1 - \sum_{i \in \Omega} x_i^{12\cdots\overline{\eta}}\ q_i\right)^2\ P_N(2). \tag{5}$$

3 Optimal Genetic Testing Design: Model and Properties

The *P-DNA Genetic Testing Design Problem* (**GP**) minimizes the probability of a false-negative ($Pr(FN)$) by making variant selection, variant partition, and pool size decisions, under a testing budget (B), and limits on panel size ($\overline{z}$), number of panels ($\overline{\eta}$), and panel pool size ($\overline{t}$):

GP Model:

$$\underset{(\boldsymbol{x}^k,t^k)_{k=1,\cdots,\overline{\eta}}}{\text{minimize}} \quad Pr\left(FN(\boldsymbol{x}^{12\cdots\overline{\eta}})\right) = \left(1-\sum_{i\in\Omega} q_i \sum_{k=1}^{\overline{\eta}} x_i^k\right)^2 P_N(2) \tag{6}$$

$$\text{subject to} \quad TC((\boldsymbol{x}^k,t^k)_{k=1,\cdots,\overline{\eta}}) = c_{SC}\times p\left(\sum_{k=1}^{\overline{\eta}}\boldsymbol{x}^k\right) + \sum_{k=1}^{\overline{\eta}}\left(\max_{i\in\Omega}\{x_i^k\}\right)\times\left(\frac{1}{t^k}+1-(1-p(\boldsymbol{x}^k))^{t^k}\right)\times\left(c_f+c_v\left(z(\boldsymbol{x}^k)\right)\right)\le B \tag{7}$$

$$\sum_{i\in\Omega} x_i^k \le \overline{z}, \qquad k=1,\cdots,\overline{\eta} \tag{8}$$

$$\sum_{k=1}^{\overline{\eta}} x_i^k \le 1, \qquad i\in\Omega \tag{9}$$

$$\sum_{k=1}^{\overline{\eta}}\left(\max_{i\in\Omega}\{x_i^k\}\right)\le\overline{\eta} \tag{10}$$

$$t^k \le \overline{t}, \qquad k=1,\cdots,\overline{\eta} \tag{11}$$

$$x_i^k \text{ binary}, \qquad i\in\Omega,\ k=1,\cdots,\overline{\eta} \tag{12}$$

$$t^k \ge 0,\ \text{integer}, \qquad k=1,\cdots,\overline{\eta}. \tag{13}$$

We let $\boldsymbol{x}^{k*}$, t^{k*}, $k=1,\cdots,\overline{\eta}$, and $\eta^*((\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}) = \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^{k*}>\mathbf{0}\}}$ respectively denote the panel vector, pool size, and number of panels used in an optimal solution to **GP**, where the indicator variable $I_{\{\boldsymbol{x}^k>\mathbf{0}\}} = \max_{i\in\Omega}\{x_i^k\}$, equivalently, it equals 1 if $\boldsymbol{x}^k > \mathbf{0}$, and 0 otherwise, for $k=1,\cdots,\overline{\eta}$, as defined in §2.1.

GP is a difficult optimization problem: the variant selection component, on its own, is NP-hard even for a special case of individual *DNA* (i.e., $D(\boldsymbol{x})=1$) [35]. That is, even without the variant partition and pool size components, which further increase the difficulty of the problem. Therefore, in the following, we establish key structural properties of optimal solutions to **GP**. These properties allow us to not only generate a family of optimal solutions and their corresponding budgets in an efficient manner, but also construct an effective approximation procedure that generates a solution for any specific budget and bound its deviation from an optimal solution.

Observe that the **GP** objective function (minimization of the FN probability) depends *only* on the variants selected, and is independent of both the variant partition and panel pool size decisions, where the latter decisions impact only the feasible region (Eqs. (7)-(11)), as formally stated below.

Remark 1.

1. Given any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, the false-negative probability $Pr\left(FN(\boldsymbol{x}^{12\cdots\overline{\eta}})\right)$ is independent of both the variant partition and panel pool sizes, $(\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}}$.

2. Given any panel $\boldsymbol{x}$, there exists an optimal pool size such that:

$$t^*(\boldsymbol{x}) = \underset{\{t\in Z^+,\ t\leq\overline{t}\}}{\operatorname{argmin}} \big\{C(\boldsymbol{x},t)\big\} = \underset{\{t\in Z^+,\ t\leq\overline{t}\}}{\operatorname{argmin}} \big\{D(\boldsymbol{x},t)\big\}.$$

   Thus, the panel's optimal pool size is a function only of the panel's own variant composition, $\boldsymbol{x}$, and is independent of other panel compositions, hence, $t^{k*} = t^*(\boldsymbol{x}^k), k = 1,\cdots,\overline{\eta}$.

3. Given any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, there exists an optimal variant partition such that:

$$(\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}} = \underset{\left\{(\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}:\sum_{k=1}^{\overline{\eta}}\boldsymbol{x}^k=\boldsymbol{x}^{12\cdots\overline{\eta}}\right\}}{\operatorname{argmin}} \big\{TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})\big\}.$$

Without loss of optimality, in the remainder of the paper we focus on the optimal pool size vector and optimal variant partition characterized in Remark 1. In particular, we expand the characterization of an optimal Dorfman pool size, established in [7], to consider panel composition (Appendix C). We then study the variant partition (§3.1) and variant selection (§3.2) components. All mathematical proofs can be found in Appendix E.

3.1 Variant Partition

In this section we establish structural properties of an optimal variant partition. Without loss of generality, for any η-panel design (η-partition), we relabel the non-empty panels as $k = 1,\cdots,\eta$, and for any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, with size $z(\boldsymbol{x}^{12\cdots\overline{\eta}}) = l$, we relabel the variants as $\{1,\cdots,l\}$, following a non-increasing order of their frequency, i.e., $q_1 \geq q_2 \geq \cdots \geq q_l$. In the following, we focus on the optimal variant partition characterized in Remark 1, that is, a minimizer of the expected total cost.

Definition 1. A partition of a given variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) = \{1,\cdots,l\}$ is said to be an *ordered partition* if $S(\boldsymbol{x}^1) = \{1,\cdots,z(\boldsymbol{x}^1)\}$, $S(\boldsymbol{x}^2) = \{z(\boldsymbol{x}^1)+1,\cdots,z(\boldsymbol{x}^1)+z(\boldsymbol{x}^2)\},\cdots,S(\boldsymbol{x}^\eta) = \{\sum_{k=1}^{\eta-1} z(\boldsymbol{x}^k)+1,\cdots,l\}$, for any panel size vector $z(\boldsymbol{x}^k)_{k=1,\cdots,\eta}$ and $\eta = 1,\cdots,\overline{\eta}$.

Thus, an ordered partition for given a variant set can be constructed in a greedy manner, with a number of the highest frequency variants assigned to one panel, a number of the next highest frequency variants assigned to another panel, and so on, and there exist multiple ordered partitions

having the same panel size vector even when symmetric partitions are excluded. The concept of an ordered partition will play an important role in the search for an optimal variant partition, as indicated by the following set of results.

Lemma 1. *Among all variant partitions of set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, there exists an ordered partition that minimizes the expected total cost, $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})$.*

In light of Lemma 1, we are able to formulate the variant partition problem for any *given variant set* as a shortest path problem, for which there exist known polynomial-time algorithms.

Corollary 1. The variant partition problem for variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) = \{1, \cdots, l\}$ (i.e., with variants relabeled following a non-increasing order of variant frequency) can be formulated as a shortest path problem (denoted $\mathbf{SP}(S(\boldsymbol{x}^{12\cdots\overline{\eta}}))$), defined on an acyclic directed graph, with vertex set $V(l) = \{1, \cdots, l, l+1\}$ (i.e., a vertex for each variant, plus a dummy vertex $l+1$); and edge set $E(l) = \{(i,j) : i < j, i, j \in V(l)\}$, where each edge (i,j) corresponds to a panel comprised of variants $\{i, i+1, \cdots, j-1\}$, that is, $\boldsymbol{x}(i,j) : (x_r = 1)_{r=i,i+1,\cdots,j-1}, (x_r = 0)_{r=1,\cdots,i-1,j,\cdots,\overline{\eta}}$, and edge cost $d(i,j)$ represents the panel's expected cost, that is, $d(i,j) = C(\boldsymbol{x}(i,j), t^*(\boldsymbol{x}(i,j)))$ (Eq. (3)) if $j - i \leq \overline{z}$, and $d(i,j) = \infty$ otherwise, where $t^*(\boldsymbol{x}(i,j))$ is computed via Property C.1 using $p(\boldsymbol{x}(i,j)) = P_N(1)\ y\big(\boldsymbol{x}(i,j)\big) + P_N(2)\ \Big(2y(\boldsymbol{x}(i,j)) - (y(\boldsymbol{x}(i,j)))^2\Big)$ (Eq. (1)), where $y(\boldsymbol{x}(i,j)) = \sum_{r=i}^{j-1} q_r$. The complexity of $\mathbf{SP}(S(\boldsymbol{x}^{12\cdots\overline{\eta}}))$ is $O(\overline{\eta}l^2)$ for $\overline{\eta} < m$ (e.g., Bellman-Ford Algorithm, [11, 19, 41]); and $O(l^2)$ for $\overline{\eta} \geq m$ (e.g., topological sorting algorithm, Dijkstra's Algorithm [19, 21, 29]).

By construction of the graph in Corollary 1, each path from vertex 1 to vertex $l+1$ corresponds to an ordered partition of set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$. For example, path $1 \to i \to j \to l+1$ corresponds to an ordered 3-partition, consisting of sets $S(\boldsymbol{x}^1) = \{1, \cdots, i-1\}, S(\boldsymbol{x}^2) = \{i, \cdots, j-1\}, S(\boldsymbol{x}^3) = \{j, \cdots, l\}$. Thus, the set of paths from vertices 1 to $l+1$ includes all ordered partitions of the given variant set, and the shortest path corresponds to the ordered partition with the lowest expected total cost, which, by Lemma 1, is an optimal partition. Lemma 1 and the shortest path problem formulation in Corollary 1 follow based on the concavity of the $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ function in $y(\boldsymbol{x})$, along with various properties established in [19, 21] (Appendix E). For a special case of the variant partition problem with $\overline{\eta} = 2$ (i.e., at most 2 panels are allowed), we also develop an algorithm that solves the variant partition problem with improved complexity compared to Corollary 1.

Lemma 2. *For the special case of the variant partition problem with $\overline{\eta} = 2$, there exists an algorithm that solves the variant partition problem with complexity $O(l)$ for a variant set with l variants, $l = 2, \cdots, \min\{2\overline{z}, m\}$.*

3.2 Variant Selection

In this section we study properties of an optimal variant set.

Definition 2. For any $l = 1, \cdots, m$, the set $S(l) \equiv \{1, 2, \cdots, l\}$, which is comprised of the l highest frequency variants in set Ω (i.e., without the relabeling of the variants), is said to be an *ordered variant set.*

By this definition, ordered variant set $S(l), l = 1, \cdots, m$, corresponds to variant selection vector $\boldsymbol{x}^{12\cdots\overline{\eta}} = \mathbf{1}_l$, and we denote its optimal variant partition by $\boldsymbol{x}^{k*}(l), k = 1, \cdots, \overline{\eta} : \sum_{k=1}^{\overline{\eta}} \boldsymbol{x}^{k*}(l) = \mathbf{1}_l$, i.e., the ordered partition that minimizes the expected total cost among all partitions of set $S(l)$, which is characterized in Lemma 1. Then, $\eta^*(l) = \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^{k*}(l) > \mathbf{0}\}}$.

Definition 3. For any variant set $S(l)$, $l = 1, \cdots, m$, the per newborn expected total cost corresponding to its optimal partition and optimal pool sizes is said to be a *budget breakpoint (B^l)*:

$$B^l \equiv TC((\boldsymbol{x}^{k*}(l), t^*(\boldsymbol{x}^{k*}(l)))_{k=1,\cdots,\overline{\eta}}), \tag{14}$$

and the solution, $(\boldsymbol{x}^{k*}(l), t^*(\boldsymbol{x}^{k*}(l))_{k=1,\cdots,\overline{\eta}})$, is said to be an *optimal breakpoint design.*

Theorem 1.

1. *When $B = B^l$, $l = 1, \cdots, m$, the optimal variant set is given by $\boldsymbol{x}^{12\cdots\overline{\eta}*} = \mathbf{1}_l$. Equivalently, $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}) = S(l)$ is an optimal variant set.*

2. *For any budget $B < B^l$, $l = 1, \cdots, m$, $\nexists$ any non-ordered or ordered variant set $S \subseteq \Omega$ such that $Pr(FN(S)) < Pr(FN(S(l)))$.*

Corollary 2.

1. At budget $B = B^l$, $l = 1, \cdots, m$, an optimal variant partition corresponds to the optimal ordered partition for set $S(l)$ that is characterized in Lemma 1, that is, $\boldsymbol{x}^{k*}(l), k = 1, \cdots, \overline{\eta}$; and B^l is the lowest budget at which $Pr(FN(S(l)))$ can be attained.

2. If set Ω has variants with distinct frequencies, i.e., no two variants have the same frequency, then $\boldsymbol{x}^{k*}(l), k = 1, \cdots, \overline{\eta}$, is the unique optimal solution at budget $B = B^l$.

3. The optimal solution at $B = B^m$ remains optimal for all budgets $B \geq B^m$.

Theorem 1 is a key result. It establishes that when the testing budget corresponds to any budget breakpoint B^l, $l = 1, \cdots, m$, the optimal variant set consists of the l highest frequency variants, i.e., set $S(l)$, and hence, can be determined in a greedy manner. The theorem also states that each budget breakpoint is the lowest budget at which the corresponding FN probability can be attained; this is also significant, because it applies to *all* (i.e., non-ordered and ordered) variant sets. Thus, if the goal is to generate the entire family of optimal breakpoint designs and budget breakpoints, then it is sufficient to consider only the m ordered variant sets, $S(1), \cdots, S(m)$, instead of the total 2^m variant sets. Thus, Theorem 1 not only leads to an efficient methodology for generating the entire family of optimal breakpoint designs and budgets in polynomial-time, but also motivates the development of an effective approximation procedure at *any* budget level (i.e., not necessarily one of the breakpoints), the optimality gap of which is analytically characterized.

In the presence of multiple optimal designs at a budget breakpoint, Corollary 3 generates those with variant set, variant partition, and pool sizes respectively characterized in Theorem 1, Lemma 1, and Property C.1. In particular, Corollary 3 follows because, in the process of computing the shortest path from vertex 1 to the dummy (terminal) vertex, Bellman-Ford Algorithm, Dijkstra's Algorithm, and the topological sorting algorithm all compute the shortest path from vertex 1 to every other vertex in the vertex set [10, 21, 29].

Corollary 3. To generate the entire family of optimal breakpoint designs and budget breakpoints, it is sufficient to solve the shortest path problem defined in Corollary 1 only once, for variant set $S(m)$. Then, for each $l = 1, \cdots, m$, the optimal variant set corresponds to $S(l)$; the optimal variant partition $(\boldsymbol{x}^{k*}(l))_{k=1,\cdots,\overline{\eta}}$, can be retrieved from the shortest path from vertex 1 to vertex $l+1$; and the optimal pool sizes, $(t^*(\boldsymbol{x}^{k*}(l)))_{k=1,\cdots,\overline{\eta}}$ (computed *a priori*, via Property C.1, for construction of the graph in Corollary 1) and budget breakpoint B^l can be retrieved from the graph, based on the edges on the shortest path from vertex 1 to vertex $l+1$. The complexity is $O(\overline{\eta}m^2)$ for $\overline{\eta} < m$, and $O(m^2)$ for $\overline{\eta} \geq m$.

In general, however, the problem of finding an optimal solution to **GP** at an arbitrary (non-breakpoint) budget B remains difficult, as discussed at the beginning of §3. Nevertheless, the family of breakpoint designs and budget breakpoints, generated by Corollary 3, enable the decision-maker

to design the *P-DNA* process considering the trade-off between the testing cost and the false-negative probability. Further, the family of breakpoint designs can be used to select an *approximate* solution at an arbitrary budget B. For example, a "good" approximate solution is given by the highest breakpoint design feasible at budget B, whose optimality gap, $\varepsilon(B)$, is bounded in the following theorem.

Theorem 2. *Consider any budget B such that $B^l < B < B^{l+1}$, for some $l = 1, \cdots, m-1$. The optimality gap of the approximate solution at budget B, given by $(\boldsymbol{x}^{k*}(l), t^*(\boldsymbol{x}^{k*}(l))_{k=1,\cdots,\overline{\eta}}$, can be bounded from above as follows:*

$$\begin{aligned}\varepsilon(B) &< Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}} = \mathbf{1}_l)) - Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}} = \mathbf{1}_{l+1})) \\ &= \left[\left(1 - \sum_{i=0}^{l} q_i\right)^2 - \left(1 - \sum_{i=0}^{l+1} q_i\right)^2\right] P_N(2) \equiv \varepsilon_{UB}(B).\end{aligned}$$

Remark 2. By Theorem 1, $\nexists B : B^l < B < B^{l+1}$, $l = 1, \cdots, m-1$, for which the upper bound on the optimality gap, $\varepsilon_{UB}(B)$, provided in Theorem 2, is tight.

The following lemma leads to insight on how the upper bound on the optimality gap behaves in the budget.

Lemma 3. *The upper bound on the optimality gap, $\varepsilon_{UB}(B)$, is non-increasing in budget B.*

Thus, at higher non-breakpoint budgets, the approximate solution (i.e., the highest breakpoint design that is feasible) is likely to yield a false-negative probability that is closer to the optimal solution.

3.3 Design Insights

In this section our goal is to provide design insights on when the decision-maker should use a 1-panel (i.e., $\eta = 1$) or a multi-panel (i.e., $\eta \geq 2$) design, and in the latter case, the number of panels, η. We do this by analytically characterizing the conditions under which each design type dominates. Recall that an η-panel design, $\eta = 1, \cdots, \overline{\eta}$, refers to a design that uses exactly η panels. To establish the structural properties in this section, we use the term *optimal η-panel design* to refer to an optimal design among the class of η-panel designs, that is, an optimal solution to **GP** when the number of non-empty panels is constrained to equal η, $\eta = 1, \cdots, \overline{\eta}$ (through the modification of Constraint (10)). The following definition will be used in our analysis.

Definition 4. A function f is said to be *sub-additive* if $f(x+y) \leq f(x)+f(y)$, and *super-additive* if $f(x+y) \geq f(x)+f(y)$, $\forall x, y \in \mathbb{R}$. [43]

In the following, we first fix the variant set for both 1-panel and multi-panel designs, hence their false-negative probability remains equal (Remark 1). We say a design type "cost-dominates" its counterpart if it yields the same false-negative probability as its counterpart, but at a lower or equal expected total cost, for a given variant set. Theorem 3 derives the conditions under which each design type cost-dominates.

Theorem 3. *Consider any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ with size $z(\boldsymbol{x}^{12\cdots\overline{\eta}}) \geq 2$. The following properties hold $\forall \eta = 2, \cdots, \min\{\overline{\eta}, z(\boldsymbol{x}^{12\cdots\overline{\eta}})\}$:*

1. *If $\exists\ c_f \geq 0$ such that the optimal η-panel (1-panel) design cost-dominates the optimal 1-panel (η-panel) design, then the η-panel (1-panel) design will continue to cost-dominate for all lower (higher) values of c_f.*

2. *Let $c_v(z(\boldsymbol{x})) = \epsilon \times c'_v(z(\boldsymbol{x}))$, where $c'_v(z(\boldsymbol{x}))$ is non-decreasing in $z(\boldsymbol{x})$, and $\epsilon > 0$. Then, $\exists\ \overline{\epsilon}(S(\boldsymbol{x}^{12\cdots\overline{\eta}}), \eta) > 0$ such that the optimal 1-panel design cost-dominates for all $\epsilon \leq \overline{\epsilon}(S(\boldsymbol{x}^{12\cdots\overline{\eta}}), \eta)$, and the optimal η-panel design cost-dominates otherwise.*

Thus, if the fixed cost c_f is sufficiently low, or the variable cost $c_v(.)$ (equivalently, ϵ) is sufficiently high, a multi-panel design cost-dominates for a given variant set, but a sufficiently low c_f value may not exist for some variant sets, that is, if the optimal 1-panel design cost-dominates the optimal η-panel design at $c_f = 0$, then the 1-panel design will cost-dominate, $\forall c_f \geq 0$. On the other hand, the threshold on the variable cost function, $\overline{\epsilon}(.)$, is always strictly positive.

Next, we discuss the optimality, i.e., in terms of minimizing the false negative probability, of each design type. To this end, we say a design type "FN-dominates" its counterpart if it leads to a lower or equal FN probability at a given budget.

Theorem 4. *For any budget B:*

1. *If $c_f + c_v\left(z(\boldsymbol{x})\right)$ is super-additive in $z(\boldsymbol{x})$, then $\exists\ \overline{k}(B, \eta)$ such that the optimal $\eta + k$-panel design FN-dominates the optimal η-panel design for any $\eta = 1, \cdots, \overline{\eta}$, $k = 0, \cdots, \overline{k}(B, \eta)$.*

2. *If $c_v(z(\boldsymbol{x})) = c$, $\forall z(\boldsymbol{x}) \in Z^+$, for some $c \geq 0$, then the optimal η-panel design FN-dominates the optimal $\eta + k$-panel design for any $\eta = 1, \cdots, \overline{\eta} - 1$, $k = 1, \cdots, \overline{\eta} - \eta$.*

Because the expected total cost, $TC(.)$, is a function of both the fixed and variable costs, c_f and $c_v(.)$, of the genetic test, and covers both the pool testing and individual retesting components, the cost thresholds in Theorem 3 clearly depend on all problem parameters. Theorem 4 makes the dependence on the genetic test cost parameters explicit. In particular, when the cost per genetic test is super-additive in panel size (Definition 4), a 1-panel design incurs a higher cost per genetic test than the combined cost of all panels of a multi-panel design. This, combined with the property that the expected number of tests per panel, $D(.)$, is monotone increasing in panel coverage, $y(\boldsymbol{x})$ (Lemma E.1), implies that the expected cost, $TC(.)$, is higher for a 1-panel design compared to a multi-panel design, for any variant set. Secondly, when the variable cost is a constant, i.e., independent of panel size, $TC(.)$ depends only on the expected number of tests, $D(.)$, which is concave in panel coverage, $y(\boldsymbol{x})$ (Lemma E.1), and hence the cost is lower when the variants are combined into a single panel, compared to a multi-panel design.

In general, while higher values of c_f favor a lower number of panels (Theorem 3), a convex $c_v(.)$ function would behave in the opposite way, and restrictions on both c_f and $c_v(.)$ are needed in Theorem 4 for the FN-dominance of one design type over another. We provide two illustrative examples below. In both examples, a budget is given, which may not correspond to a budget breakpoint for an optimal η-panel design (that is, Theorem 1 may not necessarily hold). Hence, we obtain the optimal variant set for each η-panel design via enumeration; the optimal pool size and variant partition for a given variant set are then obtained by Lemma 1 and Property C.1. Tables 1 and 2 display each optimal η-panel design (variant set, with variant partition given in $\{.\}$ – the 5-panel design is not feasible in either example), and the optimal solution to $\mathbf{GP}(\overline{\eta} = 5)$, i.e., without any restriction on number of panels, is bolded.

Common parameters for Examples 1-2: $\Omega = \{1, 2, 3, 4, 5\}$, with normalized frequency vector $\boldsymbol{q} = (0.58, 0.19, 0.11, 0.08, 0.04)$; cost per SC test, $c_{SC} = \$500$; pmf of N: $P_N(0) = 0.97125, P_N(1) = 0.02857, P_N(2) = 0.00018$.

Example 1. Consider that $c_f = \$1$, $c_v(z(\boldsymbol{x})) = \$(z(\boldsymbol{x}))^{0.9}$ (a concave function, thus violating the conditions in Theorem 4, see Definition 4), and $B = \$15.00$. The optimal 1-, 3-, and 4-panel designs all share the same non-ordered variant set $S = \{1, 2, 3, 5\}$ (the ordered variant set $S(4) = \{1, 2, 3, 4\}$ is not feasible for these designs at the given budget), and thus have the same FN probability. The 2-panel design feasibly uses the ordered variant set $S(4)$, lowering the FN probability (Table 1). Thus, the 2-panel design FN-dominates all other designs, and the FN-dominance relationship from Theorem 4 does not hold.

Table 1: Optimal η-panel, $\eta = 1, \cdots, 4$, *P-DNA* designs for Example 1.

η-panel design	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
Optimal variant set (each partition in {.}), $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))$	{1,2,3,5}	**{1}{2,3,4}**	{1}{2}{3,5}	{1}{2}{3}{5}
Expected total cost, $TC((\boldsymbol{x}^{k*}(\eta), t^*(\boldsymbol{x}^{k*}(\eta)))_{k=1,\cdots,\overline{\eta}})$	\$14.68	**\$14.91**	\$14.44	\$14.43
False-negatives per 10^6 newborns, $Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))) \times 10^6$	1.04	**0.26**	1.04	1.04

Example 2. Consider that $c_f = 0$, $c_v(z(\boldsymbol{x})) = \$(z(\boldsymbol{x}))^{1.1}$ (a convex function, thus satisfying the condition in part 1 of Theorem 4, see Definition 4), and $B = \$13.25$. The 1-panel design is FN-dominated by both 2- and 3-panel designs, but not by the 4-panel design, where the latter requires a set with at least 4 variants, and because the ordered set $S(4)$ is not feasible, it is forced to select a non-ordered variant set, and does not perform well (Table 2). Thus, Theorem 4 holds with $\overline{k}(B, 1) = 3$.

Table 2: Optimal η-panel, $\eta = 1, \cdots, 4$, *P-DNA* designs for Example 2.

η-panel design	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
Optimal variant set (each partition in {.}), $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))$	{1,2,5}	{1}{2,4}	**{1}{2}{3}**	{1}{3}{4}{5}
Expected total cost, $TC((\boldsymbol{x}^{k*}(\eta), t^*(\boldsymbol{x}^{k*}(\eta)))_{k=1,\cdots,\overline{\eta}})$	\$12.61	\$12.79	**\$13.23**	\$12.15
False-negatives per 10^6 newborns, $Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))) \times 10^6$	6.47	4.14	**2.33**	6.47

Theorem 4 leads to the following corollary.

Corollary 4. For any budget B:

1. If $c_f + c_v(z(\boldsymbol{x}))$ is super-additive in $z(\boldsymbol{x})$, then the optimal solution to **GP** must correspond to an $\eta + \overline{k}(B, \eta)$-design, for some $\eta = 1, \cdots, \overline{\eta}$.

2. If $c_v(z(\boldsymbol{x})) = c$, $\forall z(\boldsymbol{x}) \in Z^+$, for some $c \geq 0$, then the 1-panel design must be optimal for **GP**.

In general, we do not expect the genetic test cost function to be super-additive, which is satisfied, for example, by a convex non-decreasing $c_v(.)$ *and* $c_f = 0$ (Definition 4), see §2.1 for discussion. When the super-additivity condition is not satisfied, there exists no clear FN-dominance of the 1-panel design over a multi-panel design, as Example 1 demonstrates for a concave $c_v(.)$ function. We discuss the impact of the form of the variable cost function further in §4.3.2.

4 Case Study

We start with an outline of the objectives and scope of the case study (§4.1), followed by data sources and calibration (§4.2), and a discussion of key findings (§4.3-§4.4).

4.1 Objectives and Scope

We perform a case study based on the state of New York (NY), using published or publicly available data to illustrate the properties and benefits of the *P-DNA* test.

To this end, we first discuss the *P-DNA* test in isolation (§4.3), and then when integrated into two novel CF NBS processes, *IRT/P-DNA* and *P-DNA* (§4.4). For comparison, we also consider *IRT/DNA* (i.e., with individual single-panel genetic testing), which is representative of most current CF NBS processes (Fig. 1), and is the process described in [53] for NY[2]. These are not the only processes that can effectively utilize *P-DNA*, and it is possible to improve upon their performance through process-level optimization and process design, which are interesting problems beyond the scope of this paper (§5). Because our focus in this work is on the optimization of the genetic testing component (*P-DNA* or *DNA*), we use decision rules from practice to govern the rest of the process (§4.2.2).

We generate a family of optimal breakpoint designs for both *P-DNA* and *DNA*, i.e., for each ordered variant set $S(l)$, along with their budget breakpoint B^l, indexed by l (Definitions 2 and 3). Specifically, for *P-DNA*, we generate the optimal breakpoint designs characterized in Lemma 1, Theorem 1, and Property C.1, using Corollaries 1 and 3. For comparison purposes, we allow *DNA* to use multiple panels. However, *DNA* does not use pooling, and therefore, unless the genetic test cost function is super-additive, the only advantage of a multi-panel design for *DNA* is a higher coverage, and not the efficiency benefit that a multi-panel design brings to *P-DNA*. Then, under the linear $c_v(.)$ that we consider (§4.2.1), the optimal number of panels used in *DNA* always equals the minimum number of panels possible for a given variant set (i.e., index l), and the optimal *DNA* design at index l simply follows from Theorem 1, as there is no pooling, and the variant partition among panels is immaterial.

For both *P-DNA* and *DNA*, the number of panels affects only the expected testing cost, and not the false-negative probability (Remark 1); we report the latter in terms of the expected false-

[2]With the exception that newborns without any detected mutations but ultra-high *IRT* levels were also referred for SC testing during the study period in NY.

negatives per 1,485,358 newborns (the number of newborns in the NY study, see §4.2.3), using the notation FN. Therefore, for an index l, both the *P-DNA* and *DNA* will yield the same FNs *given the same testing population*. However, the characteristics of the testing population that undergoes genetic testing changes depending on the process used (i.e., depending on whether or not *IRT* is used, §4.2.4). To simplify the presentation, we refer to a budget breakpoint (i.e., expected cost per newborn) simply as *budget* in this section. All cost and budget terms are in US dollars ($).

4.2 Data Sources and Model Calibration

We next provide the case study data, and the decision rules for the three CF NBS processes.

4.2.1 Genetic Testing Data

In practice, states negotiate prices for testing kits and related material, and infrequently, some states subcontract some portion of the testing process. Further, there are numerous testing platforms to choose from, hence numerous cost structures to consider. We use representative cost data for genetic testing, loosely derived from the literature. Specifically, for the *base case*, we consider that the variable and fixed costs per genetic test are, $c_v(z(\boldsymbol{x})) = \$z(\boldsymbol{x})$ (i.e., linear increasing in panel size) and $c_f = \$10$, respectively, and the costs per *IRT* and *SC* tests are, $c_{IRT} = \$1.5$ and $c_{SC} = \$161.4$, respectively [75]. Because genetic testing is a generic, and relatively new, technology, with many applications beyond newborn screening and much current development, the cost of genetic testing is, in general, trending downward. Therefore, we also consider scenarios with different fixed and variable costs (§4.4).

Regarding the technological limits, we do not use a pool size limit $\bar{t}$ due to a lack of rigorous clinical studies showing the relationship of pool size and accuracy for *DNA* (§5). Based on current probe-based technologies, we use a technological panel size limit of $\overline{z} = 90$ variants [64]. We vary the limit on number of panels, $\overline{\eta}$, to study optimal designs with different number of panels (with $\overline{\eta} = m$ representing the setting with no limit on number of panels). If multiple panels are used, they are tested in parallel; and an η-panel design, $\eta = 1, \cdots, \overline{\eta}$, can test up to $\eta\overline{z}$ variants.

To provide perspective, consider that the American College of Medical Genetics (ACMG) recommends testing, using, at a minimum, the 23 most common variants in the US [28], but current CF panels differ among states. For instance, New York's process, as described in [53], used 39 variants (until 2019), and California's process uses 40 variants [55], while states like North Carolina and Wisconsin both use a panel of 139 variants (using a next-generation sequencing technology)

[70, 89]. No state currently uses Dorfman pooling or multiple panels as part of CF NBS. To our knowledge, all current *DNA* panels are derived solely based on descriptive analyses. Typically, a number of the highest frequency variants are added to the *DNA* panel, based on frequencies from historical screening data and budget considerations, e.g., [28, 55].

4.2.2 Process Decision Rules

For the *IRT/DNA* and *IRT/P-DNA* processes, we follow the 5% *daily threshold* for *IRT*, reported in [53], in which newborns with *IRT* levels in the top 5% of all *IRT* levels each testing day are classified as *IRT*-positive, and sent to the next test in the process. This daily percentile threshold is a common decision rule for the *IRT* test due to seasonal, biological, and demographic variations in *IRT* levels [53, 58, 77, 82], with 4% and 5% threshold values commonly used, dictated mainly by limited testing budgets [45]. In addition, following current practices, for genetic testing (*DNA* or *P-DNA*), we classify any newborn with at least one mutation detected (of any variant and in any panel) as screen-positive, and refer them for *SC* testing for diagnosis (Fig. 1).

4.2.3 Population-level Data

The CFTR-2 data set [18] is the most comprehensive data set on CF-causing variants, and includes 352 well-characterized CF-causing variants from more than 88,000 CF-positive subjects. We include all 352 variants in set Ω (hence $m = 352$), and derive the variant frequency vector, $\boldsymbol{q} = (q_i)_{i\in\Omega}$, based on variant frequencies reported in the data set, normalized so that $\sum_{i\in\Omega} q_i = 1$. The CFTR-2 data set indicates that variant frequencies can vary significantly, for example, the most frequent variant, F508Del, has a (normalized) frequency of 0.741741, which is around 27 times the frequency of the second most frequent variant, G542X, which has a frequency of 0.027031. Some variants are very rare, e.g., 248Del2515, the least frequent variant, has a frequency of 1.5×10^{-5}.

Only the *P-DNA* process uses genetic testing on all newborns (*general newborn population*), and both *IRT/DNA* and *IRT/P-DNA* processes use genetic testing only on the subset of newborns that are *IRT*-positive (*post-IRT population*). Importantly, CF and carrier prevalence likely differ in these two populations. Therefore, we derive the proportion of mutation-free, CF-carrier, and CF-positive newborns for both the general population (i.e., the pmf of random variable N), and the post-*IRT* population (which we denote by the pmf of random variable N_{post}). For this purpose, we use the CF NBS data reported in [53], for 1,485,358 newborns screened in NY between 2007-2012. In this study cohort, 260 newborns were CF-positive, 9 of whom were missed by the *IRT*

test (these 9 FNs at the *IRT* level constitute all the FNs reported for NY NBS in this cohort); 79,973 newborns were screened post-*IRT* (i.e., with individual 1-panel *DNA*), based on the 5% *IRT* threshold; and of those 79,973 post-*IRT* newborns, all 251 were correctly identified as CF-positive via the process (excluding the 9 FNs from *IRT*), and 3,850 newborns, who were referred for *SC* testing with only one mutation detected, were confirmed to be CF-carriers by *SC*. Thus, for the general population, we compute:

$$P_N(2) = \frac{\#\text{ CF-positive newborns}}{\#\text{ screened newborns}} = \frac{260}{1,485,358} = 0.00018.$$

We do not have data on the number of carriers and mutation-free newborns in this cohort, therefore we use estimates from the literature that suggest that the proportion of CF-carriers in the general US population is around 1:35 [27, 46], that is, $P_N(1) = \frac{1}{35} = 0.02857$, yielding $P_N(0) = 0.97125$. For the post-*IRT* newborn population, based on the NY data set, we compute:

$$\begin{aligned} P_{N_{post}}(2) &= \frac{\#\text{ CF-positive newborns in the post-}IRT\text{ population}}{\#\text{ post-}IRT\text{ newborns}} = \frac{251}{79,973} = 0.00314, \\ P_{N_{post}}(1) &= \frac{\#\text{ CF-carriers in the post-}IRT\text{ population}}{\#\text{ post-}IRT\text{ newborns}} = \frac{3,850}{79,973} = 0.04814, \end{aligned}$$

yielding $P_{N_{post}}(0) = 0.94872$.

4.2.4 Process-level Data

When a process uses *IRT*, we modify the budget breakpoint (Eq. (14)) and the false-negative probability (Eq. (5)) to consider the 5% *IRT* threshold, and respectively include the *IRT* cost per newborn, denoted by c_{IRT}, and the rate of FNs stemming from *IRT*; and for the genetic test (*DNA* or *P-DNA*), we use the pmf of the post-*IRT* random variable, N_{post}, as discussed below. Note that variant set $S(l), l = 1, \cdots, m$, corresponds to $\boldsymbol{x}^{12\cdots\overline{\eta}*}(l) = \mathbf{1}_l$ (for both *DNA* and *P-DNA*).

IRT/DNA($\overline{\eta}$) process, with *DNA* panel vector $(\boldsymbol{x}^{k*}(l)_{k=1,\cdots\overline{\eta}})$:
We have that, $\boldsymbol{x}^{k*}(l) = \mathbf{1}_{\min\{l,k\overline{z}\}} - \mathbf{1}_{\min\{l,(k-1)\overline{z}\}}$, $k = 1, \cdots, \overline{\eta}$, leading to:

$$B^l = c_{IRT} + 0.05 \times \left(k \times c_f + \sum_{k'=1}^{k} c_v \left(z(\boldsymbol{x}^{k'*}(l)) \right) + c_{SC} \times p \left(\boldsymbol{x}^{12\cdots\overline{\eta}*}(l) \right) \right),\ l = 1, \cdots, m. \tag{15}$$

IRT/P-DNA($\overline{\eta}$) process, with *P-DNA* panel vector $(\boldsymbol{x}^{k*}(l)_{k=1,\cdots,\overline{\eta}})$:

$$B^l = c_{IRT} + 0.05 \times TC((\boldsymbol{x}^{k*}(l), t^*(\boldsymbol{x}^{k*}(l)))_{k=1,\cdots,\overline{\eta}}),\ l = 1, \cdots, m, \tag{16}$$

where $TC(.)$ is as defined in Eq. (4).

We compute the FN probability for variant set $S(l)$, equivalently, $\boldsymbol{x}^{12\cdots\overline{\eta}*}(l) = \mathbf{1}_l$, as follows (with the subscript, *post*, used to denote the post-IRT genetic testing):

$$\begin{aligned} Pr(FN(\mathbf{1}_l)) &= (1-0.05) \times Pr(N=2|\text{not top 5\% } IRT) + \ 0.05 \times Pr(FN_{post}(\mathbf{1}_l)|\text{top 5\% } IRT) \\ &= 0.95 \times Pr(N=2|\text{not top 5\% } IRT) + 0.05 \times \left(1 - \sum_{i=1}^{l} q_i\right)^2 P_{N_{post}}(2), \end{aligned}$$

where $Pr(FN_{post}(\mathbf{1}_l)|\text{top 5\% } IRT)$ is the conditional probability of FN from genetic testing, given that the newborn is IRT-positive, and is calculated via Eq. (5) (with the pmf of random variable N replaced by the pmf of random variable N_{post}); and $Pr(N=2|\text{not top 5\% } IRT)$ is the conditional probability that a newborn is CF-positive, given that they are not in the top 5% IRT group (i.e., IRT-negative), and is estimated based on the data in [53]. In particular, the cohort of $1,485,358$ contained 9 FNs from IRT, and $79,973$ IRT-positives, hence we derive:

$$Pr(N=2|\text{not top 5\% } IRT) = \frac{\#\ IRT\text{-negative and CF-positive newborns}}{\#\ IRT\text{-negative newborns}} = \frac{9}{(1,485,358-79,973)} = 6.4 \times 10^{-6}.$$

4.3 Properties of the *P-DNA* Test

In this section we illustrate key properties of the *P-DNA* test using the data for the *P-DNA* process, that is, variant frequencies in the general newborn population. To this end, we first explore the two novel aspects of a *P-DNA* design, pooling and multi-panel testing (§4.3.1), and then discuss some operational issues for *P-DNA* implementation (§4.3.2).

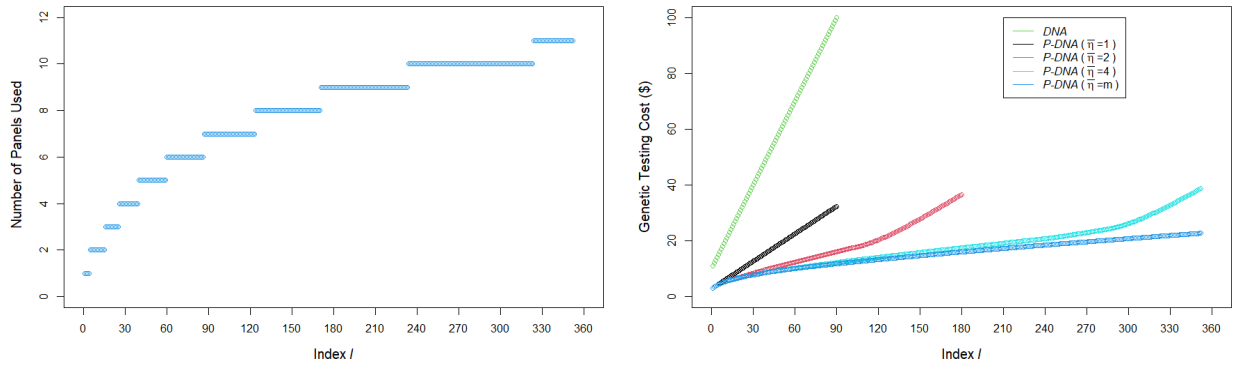
4.3.1 Efficiency and Efficacy of *P-DNA*

We first consider *P-DNA*($\overline{\eta} = m$) (i.e., no limit on number of panels), and study the optimal number of panels ($\eta^*(l)$) for all ordered variant sets $S(l), l = 1, \cdots, m(= 352)$, see Fig. 3(a). A 1-panel design is optimal only for small variant sets, e.g., it is optimal for $l = 1, \cdots, 4$ when $c_f = \$10$ (base case), for $l = 1, 2$ when $c_f = \$5$, and for $l = 1, \cdots, 10$ when $c_f = \$20$. A 2-panel design is optimal for $l = 5, \cdots, 15$ in the base case, after which an optimal partition always utilizes more than two panels. Thus, even if we consider only the 23 most common variants, the minimum recommended by the ACMG [28], a design with more than two panels is optimal. The optimal number of panels, $\eta^*(l)$, increases as variant set $S(l)$ expands, e.g., when the variant set contains all the variants in Ω ($l = 352$), the optimal number of panels is 11. In addition, as l increases, $\eta^*(l)$ remains unchanged over a larger range of l values (e.g., 2-panel is optimal for $l = 5, \cdots, 15$, whereas 10-panel is optimal

for $l = 234, \cdots, 323$), because at larger values of l, adding one more variant has only a small impact on the overall prevalence of the variant set.

We next study the cost efficiency of a multi-panel *P-DNA* design. Fig. 3(b) displays the genetic testing cost as a function of index l for *P-DNA*($\overline{\eta}$), for $\overline{\eta} = 1, 2, 4, m$, and for the current 1-panel *DNA*, indicating that all variations of the *P-DNA* process dominate the *DNA* process. Interestingly, testing all 352 variants in Ω with 11 panels leads to a similar budget ($27.34) as testing 132 variants with 2 panels ($27.53), or testing only 14 variants with a 1-panel *DNA* ($28.11), highlighting the benefits stemming from the pooling and multi-panel aspects of *P-DNA*. Due to capacity restrictions of genetic testing machines, using more panels may reduce the throughput; this aspect should be considered in practical applications. We illustrate how a *P-DNA* achieves these benefits with a simple example.

Figure 3: Optimal number of panels used ($\eta^*(l)$) by *P-DNA*($\overline{\eta} = m$), and the genetic testing cost for *P-DNA*($\overline{\eta} = 1, 2, 4, m$) and *DNA*, as a function of index l



(a) Number of panels vs design index

(b) Genetic testing cost vs design index

Example 3. Consider *P-DNA*($\overline{\eta} = 1$), *P-DNA*($\overline{\eta} = 2$), and *DNA* designs for $l = 90$ (Table 3):

- *P-DNA*($\overline{\eta} = 2$): Panel 1 has 12 variants (the highest frequency variants), a pool size of 7, and a retest (i.e., individual testing) probability of 0.164, while Panel 2 has 78 variants, a pool size of 19, and a retest probability of 0.053, leading to an expected genetic testing cost per newborn of $16.05 ($= \$9.30 + \$6.75$).

- *P-DNA*($\overline{\eta} = 1$): The single, 90-variant panel has a pool size of 7 and a retest probability of 0.181, leading to an expected genetic testing cost per newborn of $32.35.

- *DNA*: The single, 90-variant panel incurs, in the absence of pooling, an expected genetic testing cost per newborn of $100.

For this instance, the 1- and 2-panel *P-DNA* designs reduce the genetic testing cost by 67.6% and 83.9%, respectively, over *DNA*. Going from one to two panels reduces the *P-DNA* cost by 50.3% (from \$32.35 to \$16.05) due to further efficiencies gained through multi-panel testing. To see this, consider that the *P-DNA* testing cost has two components: the initial pooled testing cost, and the expected (individual) retesting cost. The 2-panel design reduces both costs (Table 3). For the pooled test(s), the 1- and 2-panel designs respectively incur costs of \$14.29 and \$3.14+\$4.63 = \$7.77: the 2-panel design is able to reduce this cost through the use of a larger pool size for the more expensive Panel 2 test (Table 3). On the other hand, for retesting, the 1- and 2-panel designs respectively incur expected costs of \$18.06 and \$3.61 + \$4.67 = \$8.28: this cost reduction follows because a 2-panel design offers a main advantage, in that if a panel tests positive (which happens with probability $1-(1-p(\boldsymbol{x}))^t$ for each panel $\boldsymbol{x}$ with pool size t), then individual retesting is required only for the positive-testing panel, instead of the entire variant set, as happens in the 1-panel design. Indeed, the more expensive Panel 2 has a lower retest probability, leading to significant cost savings.

Table 3: Details for *P-DNA*($\overline{\eta}=1$) and *P-DNA*($\overline{\eta}=2$) designs for $l=90$.

$l=90$		*P-DNA*($\overline{\eta}=1$)	*P-DNA*($\overline{\eta}=2$)	
		Panel 1	Panel 1	Panel 2
Panel Size ($z(\boldsymbol{x})$)		90	12	78
Pool Size ($t^*(\boldsymbol{x})$)		7	7	19
Retest Probability $(1-(1-p(\boldsymbol{x})^{t^*(\boldsymbol{x})})$		0.181	0.164	0.053
Cost Per Genetic Test ($c_f + c_v(z(\boldsymbol{x}))$)		\$10 + \$90 = \$100	\$10 + \$12 = \$22	\$10 + \$78 = \$88
Genetic Testing Cost per Newborn	Pooled Test $\frac{1}{t}(c_f + c_v(z(\boldsymbol{x})))$	\$100/7 = \$14.29	\$22/7 = \$3.14	\$88/19 = \$4.63
	Individual Retest $\left(1-(1-p(\boldsymbol{x})^{t^*(\boldsymbol{x})}\right)\left(c_f + c_v(z(\boldsymbol{x}))\right)$	0.181 × \$100 = 18.06	0.164 × \$22 = 3.61	0.053 × \$88 = 4.67
	Total $(C(\boldsymbol{x}, t^*(\boldsymbol{x})))$	\$14.29 + \$18.06 = \$32.35	\$3.14 + \$3.61 = \$6.75	\$4.63 + \$4.66 = \$9.30

To gain further insight, for *P-DNA*($\overline{\eta}=2$) we display the panel sizes ($z(\boldsymbol{x}^{k*}), k=1,2$) and panel coverage ($y(\boldsymbol{x}^{k*}), k=1,2$), as well as pool sizes ($t^{k*}, k=1,2$) and individual retest probabilities for the family of optimal designs (Figs. 9-10, Appendix D.2). These figures indicate that the optimal variant partition is such that one panel typically contains a smaller number of higher frequency variants (Panel 1 in this case) than the other panel. Relegating those rare variants to one panel (Panel 2) allows for a larger pool size (of 19), and even with a larger panel size (of 78), it still has a lower retest probability, which in turn reduces the testing cost for rare variants. Observe, in Fig. 3 that around index 109, the 2-panel design starts to lose efficiency, this corresponds to

the point where Panel 2 reaches its technological limit, forcing Panel 1 to grow at a quicker pace (interestingly, as l increases, the additional variant goes into Panel 2, and the highest frequency variant in Panel 2 transfers to Panel 1).

In summary, the *P-DNA* test utilizes the testing budget more efficiently (through pooling and multi-panel testing); further, the use of multi-panels allows the detection of more variants compared to the current single-panel, individual testing practice (*DNA*).

4.3.2 Practical Considerations

Next, we briefly discuss some practical considerations for the *P-DNA* test.

- **Common pool sizes:** We characterize an optimal common pool size for *P-DNA* (Appendix C.1). Our numerical study indicates that while moving from a 1-panel to a 2-panel design is highly effective in reducing costs (as discussed above), allowing panel pool sizes to differ yields much smaller benefits, thus using common pool sizes preserves most of the efficiency gains of *P-DNA*, while reducing process complexity (Appendix C.2).

- **Variable cost function:** In addition to the linear case, we also consider *P-DNA* designs under convex and concave variable cost functions, $c_v(.)$. A concave (convex) $c_v(.)$ function makes it more (less) favorable to combine variants in one panel, and not surprisingly, lowers (raises) the budget breakpoint for each index l, compared to the linear case (Fig. 8, Appendix D.1). Interestingly, the 1-panel *P-DNA* cost-dominates the 2-panel *P-DNA* for small variant sets (Theorem 3) under all cost functions, i.e., for $l = 1, \cdots, 4$ for linear cost, for $l = 1, \cdots, 12$ for concave cost, and for $l = 1, 2$ for convex cost functions (Appendix D.1).

- **Equity considerations:** Variant frequencies can differ across demographic groups, e.g., [36], and larger panels are in general more equitable. For example, considering the most common variants for the White, Hispanic, and Black groups [79], we find that the 25 (50) most common variants for each of these groups are included in the top 48, 118, and 162 (61, 216, and 195) most common variants, respectively, in the CFTR-2 data set. Thus, by simply expanding the variant set, *P-DNA* allows for the screening of some rare variants that may be found predominantly in minority demographic groups, thus contributing to a more equitable solution, see §5.2 for further discussion.

4.4 Process Comparison

In this section we compare the three testing processes, *IRT/DNA*, *IRT/P-DNA*, and *P-DNA*: We consider two variations of the latter two processes, with $\overline{\eta} = 2$ and 4. For each process, we generate the family of optimal breakpoint designs for $l = 1, \cdots, 2\overline{z}(= 180)$ (above which a 2-panel design is no longer feasible), and compute the corresponding budget breakpoint B^l, and expected FNs in a cohort of 1,485,358 newborns (similar to the NY cohort [53]). This allows us to study the screening cost versus accuracy trade-off at a process level. Due to the declining cost of genetic testing, we also look at a second cost structure that is half of the base case costs, i.e., $c_f = \$5$ and $c_v(z(\boldsymbol{x})) = \$0.5 \times z(\boldsymbol{x})$.

For both cost structures, *P-DNA*($\overline{\eta} = 2$) is the most expensive, followed by *P-DNA*($\overline{\eta} = 4$), *IRT/DNA*, *IRT/P-DNA*($\overline{\eta} = 2$), and *IRT/P-DNA*($\overline{\eta} = 4$) for each l (Fig. 4). Comparing *IRT/DNA* and *IRT/P-DNA*, we see that for the same index l (where both processes incur the same FNs), either *IRT/P-DNA* variation ($\overline{\eta} = 2, 4$) cost-dominates *IRT/DNA*, by reducing the budget at the same FN level. Further, the more panels are allowed, the lower the budget.

We next compare *P-DNA* and *IRT/P-DNA*. Because *P-DNA* genetically tests all newborns, as opposed to only 5% of newborns in *IRT/P-DNA*, it naturally incurs a higher budget: This higher *P-DNA* budget happens through an increase in not only the genetic tests, but also *SC* referrals. At the same time, the elimination of the *IRT* test in *P-DNA* lowers the FNs for each l, compared to *IRT/P-DNA*. We illustrate this with an example.

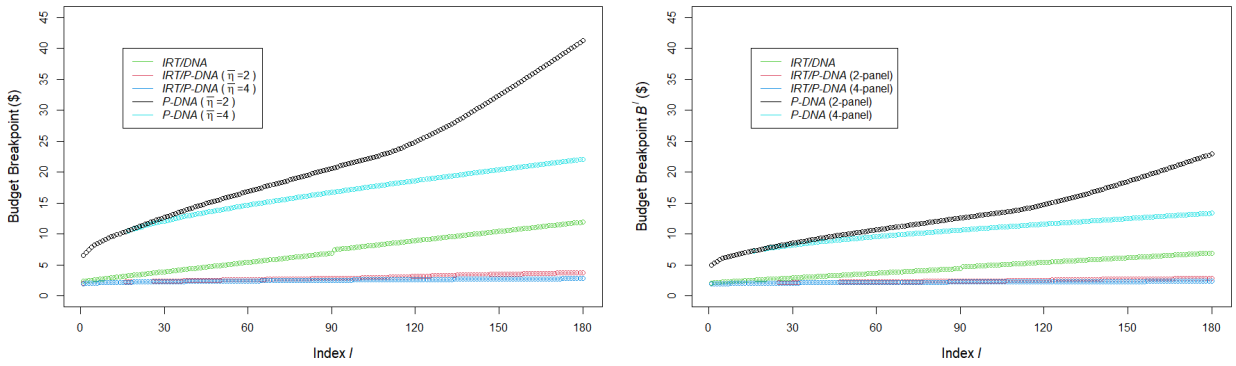
Example 4. Consider *IRT/P-DNA*($\overline{\eta} = 2$) and *P-DNA*($\overline{\eta} = 2$) designs at $l = 5$, for $c_f = \$10$ and $c_v(z(\boldsymbol{x})) = \$z(\boldsymbol{x})$ (Table 4):

- *IRT/P-DNA*($\overline{\eta} = 2$): The expected FNs is 16.52, and the budget is \$2.07, of which \$1.50 is for *IRT*, \$0.23 is for *P-DNA* (used only on 5% of newborns), and \$0.34 is for *SC* testing.
- *P-DNA*($\overline{\eta} = 2$): The expected FNs is 8.34, and the budget is \$8.25, of which \$4.44 is for *P-DNA* (used on all newborns), and \$3.81 is for *SC* testing.

In this example, the diagnostic *SC* testing contributes to almost half of the budget of *P-DNA*, which refers over ten times as many carriers for *SC* testing as the other processes (Table 4). The *SC* component of cost grows slower than the genetic testing component as l rises, because the number of carriers detected grows at a fairly slow rate, and index l does not affect the per test cost of *SC* (Fig. 11, Appendix D.2).

Finally, we study the FN versus budget trade-off. Because the commonly used *IRT/DNA* is always cost-dominated by *IRT/P-DNA*, we compare *IRT/P-DNA* and *P-DNA*. *IRT/P-DNA*($\overline{\eta} = 2$) attains its lowest FNs at $l = 180$ (i.e., at its technological limit), with 9.05 FNs and a budget of \$3.74 (Fig. 5(a) and Table 4). On the other hand, *P-DNA*($\overline{\eta} = 2$) reduces FNs from 9.05 for all variant sets $l \geq 5$, e.g., at $l = 5$, it incurs 8.34 FNs at a budget of \$8.25. Thus, for budgets less than \$3.74, *IRT/P-DNA* is the best process, but by investing more for testing, FNs can be further reduced. To explore the cost-effectiveness of *P-DNA*($\overline{\eta} = 2$) for $l \geq 5$, we compute the cost per each additional FN reduction possible in this process for $l \geq 5$, over the baseline *IRT/P-DNA*($\overline{\eta} = 2$) at $l = 180$ (i.e., the index at which this process yields its lowest FNs), that is, *Cost per FN*$=\frac{B^l(\textit{P-DNA}(\overline{\eta}=2))-3.74}{9.05-FN(S(l),\textit{P-DNA}(\overline{\eta}=2))}$. We repeat the same analysis for the $\overline{\eta} = 4$ case, using *IRT/P-DNA*($\overline{\eta} = 4$) at $l = 180$ as the baseline, which incurs 9.05 FNs at a budget of \$2.78.

Figure 4: Budget breakpoints, B^l, for various processes for optimal designs for two genetic testing cost structures, as a function of index l

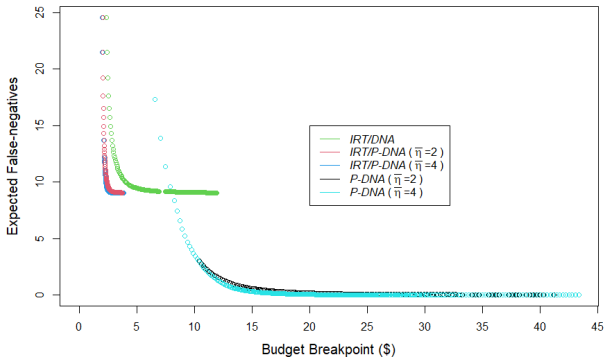


(a) $c_f = \$10$, $c_v(z(\boldsymbol{x})) = \$z(\boldsymbol{x})$

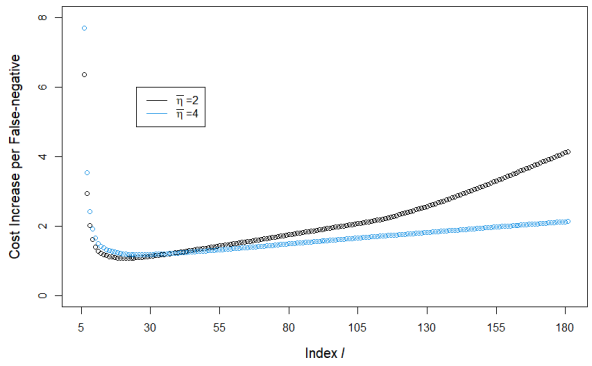
(b) $c_f = \$5$, $c_v(z(\boldsymbol{x})) = \$0.5 \times z(\boldsymbol{x})$

We discuss the cost-effectiveness of *P-DNA*($\overline{\eta} = 2$) (the observations for $\overline{\eta} = 4$ are similar). At $l = 5$, the cost per FN reduction is high, with a value of 6.35. This is due to the relatively large gap between the budgets of *IRT/P-DNA* at $l = 180$, and *P-DNA* at $l = 5$, and the small difference in their respective FNs (Fig. 5(a)-(b)). However, the cost per FN reduces as more variants are added to *P-DNA*, that is, FN decreases at a faster rate than the testing cost increases, until index $l = 21$. At this point, the cost per FN reaches its minimum value of 1.08, before starting to increase again with each additional variant, which happens because the variants are arranged in a non-increasing order of their frequency, thus, adding one more variant brings a diminishing return.

Figure 5: Expected false-negatives for the three processes under *P-DNA*($\overline{\eta} = 2$) and ($\overline{\eta} = 4$) for budget breakpoints, B^l, as a function of index l



(a) Expected false-negatives vs budget breakpoint



(b) Cost increase per FN for *P-DNA*($\overline{\eta}$) compared to *IRT/P-DNA*($\overline{\eta}$) vs design index

Table 4: Results for select optimal designs (indexed by l) for the three processes, for two genetic testing cost structures, where $\overline{\eta} = 2$.

CF NBS Process	Index (l)	Budget Breakpoint (B^l)		# of FN (per 1,485,358)		# of SC Tests (per 1,485,358)	
		$c_f = \$5$, $c_v(z(\boldsymbol{x})) = \$0.5 \times z(\boldsymbol{x})$	$c_f = \$10$, $c_v(z(\boldsymbol{x})) = \$z(\boldsymbol{x})$	*IRT*	*DNA*	CF-carrier	CF-positive
IRT/DNA	5	\$2.22	\$2.59	9	7.52	3,160.35	243.48
IRT/P-DNA		\$1.96	\$2.07				
P-DNA		\$6.03	\$8.25	-	8.34	34,836.79	251.66
IRT/DNA	20	\$2.63	\$3.37	9	2.06	3,491.27	248.94
IRT/P-DNA		\$2.06	\$2.25				
P-DNA		\$7.64	\$11.08	-	2.26	38,484.47	257.74
IRT/DNA	40	\$3.14	\$4.39	9	0.69	3,646.30	250.41
IRT/P-DNA		\$2.17	\$2.46				
P-DNA		\$9.30	\$14.21	-	0.73	40,193.44	259.27
IRT/DNA	90	\$4.40	\$6.90	9	0.17	3,756.94	250.83
IRT/P-DNA		\$2.36	\$2.81				
P-DNA		\$12.57	\$20.61	-	0.15	41,413.04	259.85
IRT/DNA	180	\$6.91	\$11.91	9	0.05	3,817.86	250.95
IRT/P-DNA		\$2.83	\$3.74				
P-DNA		\$22.90	\$41.20	-	0.02	42,084.49	259.98

5 Conclusions, Limitations, and Future Work

In this paper, we introduce and study a novel pooled *DNA* (*P-DNA*) test, which cleverly incorporates pooling and multi-panel testing to improve the efficiency of standard *DNA* tests, which are used in many applications, including NBS. While pooling is a common strategy for increasing test efficiency, its use in CF NBS is novel. More importantly, our work shows that pooling and multi-panel testing work in a synergistic way to improve the efficiency of testing, and the benefits

go beyond what pooling alone can achieve. Further, multi-panel testing also expands the technological limit on the number of variants the genetic test can screen for, thus allowing for a higher clinical sensitivity. Thus, both aspects provide substantial benefit, especially in the NBS setting where the goal is to detect genetic disorders, which can be caused by a large number of harmful, and potentially rare, genetic variants. We incorporate *P-DNA* into two novel CF NBS processes, namely *IRT/P-DNA* and *P-DNA*, which we compare with the most commonly used *IRT/DNA* process (Fig. 1). We develop a family of optimal designs and budget breakpoints for each process, and derive key design insights. Considering the implementation of these processes, pooling is already used in high-volume public health screening (using genetic testing to detect pathogens), while multi-panel testing is essentially just using multiple tests, and each newborn's dried blood spot already undergoes multiple tests for multiple disorders. Hence, the methods discussed in this paper are viable from an implementation perspective. In general, less complex testing protocols are desirable, and this should be considered when determining the number of pools, or whether common pool sizes should be used. The level of implementation difficulty is also a function of the specific testing platforms used in the testing laboratory; but this topic is beyond the scope of this paper.

It is our hope that this work inspires both academicians and practitioners to further explore these novel approaches to NBS. For this purpose, we summarize the main takeaways from this work, followed by some limitations and future research directions.

5.1 Summary of Insights Gained and Further Discussion

This study leads to the following insights:

- *The IRT/P-DNA process outperforms the current IRT/DNA process.* This follows due to the efficiencies from pooling and multi-panel testing, allowing *IRT/P-DNA* to reduce the false-negative rate at the same budget (through the use of a larger variant set), or attain the same false-negative rate at a lower budget, e.g., when both processes use the same variant set comprised of the 90 highest frequency variants, which offers high coverage (97.58%), the *IRT/P-DNA* process cost is around 40.7% of the the *IRT/DNA* process cost. This savings can be used to provide genetic testing to more newborns, i.e., by lowering the *IRT* threshold, thus potentially decreasing the false-negative rate attributed to *IRT*, increasing the social benefits of screening.

- *The high efficiency of P-DNA allows it to be used as a single-test process.* This has the advantage of reducing the false-negative rate by eliminating the *IRT* test. Indeed, our analysis indicates that at higher testing budgets, the *P-DNA* process outperforms the other processes in terms of the false-negative rate. While the *P-DNA* process might still be too expensive, we expect it to become more viable as genetic testing costs go down. For example, our cost-effectiveness analysis indicates that if the fixed and variable costs of genetic testing reduce by 50%, the *P-DNA* process budget (cost) reduces between 24% (for a variant set containing only the most frequent variant) and 45% (for a variant set containing the 180 most frequent variants).
- *A practical strategy, of using a common pool size for a multi-panel P-DNA, preserves most of the efficiency gains.* In particular, the use of a common pool size in *P-DNA*, with at most 2 panels allowed, increases the testing cost by at most 9% compared to the unconstrained pool size setting.
- *A family of optimal P-DNA designs (i.e., at specific budgets) can be generated in polynomial time.* While the general problem, of determining an optimal design at an arbitrary budget, is NP-hard, we exploit structural properties of optimal designs, which allow us to generate a family of optimal designs, and their budgets, in polynomial time. *This not only allows policy-makers to consider the cost versus accuracy trade-off when designing an NBS process, but also provides a basis for an approximate solution, at any budget, with a performance guarantee that improves as the budget increases.*

5.2 Limitations and Future Work

Next we discuss some limitations of this work, which also provide opportunities for future research:

- We assume that all CF-causing variants are in set Ω (Assumption ($\boldsymbol{A_1}$)). Practically speaking, set Ω will grow in size as new CF-causing variants are discovered/characterized. However, we believe this is a minor limitation. New variants are likely to be very rare, and most newborns with CF are likely to have at least one variant in the panel, and would thus be referred for diagnostic testing. Moreover, while this implies that our analysis likely underestimates the rate of false-negatives, this underestimation is similar for each process at each design index, thus our overall findings should remain valid even when this assumption is relaxed.
- Our assumption, that pooling does not affect analytical sensitivity (Assumption ($\boldsymbol{A_3}$)), should be validated in a laboratory setting, and an appropriate pool size limit should be determined.

- Variant frequencies can differ across demographic groups, e.g., see [36]. Thus, considering equity issues when selecting variants is important, but not trivial, given the complexity of the **GP** Model. Still, the proposed *P-DNA* framework, with pooling and multi-panel testing, not only utilizes the testing budget more efficiently, but also expands the technological limit on the number of variants that can be screened for, over the current single-panel, individual testing paradigm. This allows the *P-DNA* to expand the variant set, and include more rare variants found predominantly in underrepresented demographic groups. In addition, this work provides a benchmark for an equity-based model, which is an important future research direction.
- Optimal breakpoint designs are based on the ordering of the variants with respect to their frequency. As a result, an optimal design may be robust to small estimation errors in variant frequencies, as long as the ordering is preserved. Consider that in the CFTR-2 data set, the most common variant (variant 1) has a (normalized) frequency of 74.17%, which exceeds the combined frequency of all other variants; and the next ten common variants $(2, \cdots, 11)$ have a combined frequency of 12.93%, which again exceeds the combined frequency of all lower frequency variants $(12, \cdots, 352)$. This analysis suggests that if the budget permits, these twelve most common variants should always be included in an optimal variant set. On the other hand, as expected, many rare variants are clustered together, but these variants are less likely to be in an optimal variant set unless the budget is quite large. It is an important future research direction to explore rigorous approaches for constructing a robust variant set.
- We study Dorfman pooling due to its simplicity and efficiency, which make it a commonly used pooling method in public health screening. Other pooling methods have also received attention in the literature (§1). For example, array pooling, e.g., [56, 72], utilizes overlapping pools (to form a testing array), as opposed to the non-overlapping pools used in Dorfman pooling. While our structural results do not necessarily extend when *P-DNA* is implemented with array pooling, a preliminary numerical study indicates that *P-DNA* with array pooling can be effective in certain cases (Appendix C.3). In general, neither array pooling nor Dorfman pooling always dominates, and identifying the optimal pooling method for a given setting is an important research direction.
- For the *IRT/DNA* and *IRT/P-DNA* processes, our model optimizes only the genetic testing component, and not the overall process. For instance, rather than expanding the genetic testing panel, it might be preferable in some cases to modify the *IRT* threshold beyond the

current 5%. This would increase the number of newborns undergoing genetic testing, hence reducing the false-negatives from *IRT*. This is a difficult design problem, in part because the distribution of *IRT* levels for newborns is noisy in general, and for CF-positive newborns, data sparsity is a problem. The model and methodology developed in this paper can be used as a building block to analyze this difficult design problem.

- Alternative processes that utilize *P-DNA* should be studied. For instance, it would be interesting to study a *P-DNA*/*IRT* process that uses *IRT* only for those newborns with one mutation detected: in this case, the *SC* referral rule could be based on *both* the newborn's mutational status and the *IRT* level, which is potentially more accurate than declaring newborns CF-negative based solely on *IRT* levels. Further, California, and now New York, follow *DNA* with a more complete sequencing, i.e., next-generation sequencing test, which is used to reduce the number of newborns referred for *SC* testing [55], and a similar concept could be explored for *P-DNA*.
- The *P-DNA* test makes universal genetic testing more viable for NBS, which could allow other genetic disorders to be included into this test as well. As long as the number of CF-causing variants in the panel, or panels, is below the technological limit, probes for other disorders can be added, and this "disorder bundling" can potentially reduce NBS costs, especially as the prevalence rates of many other NBS disorders are much lower than that of cystic fibrosis. Further, with universal genetic testing, reliable carrier detection becomes more viable, which could be of value. Currently, NBS programs do not provide this information, partially for ethical/privacy concerns, and partially because there is not universal genetic testing, thus only a limited portion of the newborns would receive this information. Looking into carrier detection issues is beyond the scope of this research, but this is an important research direction.
- Pricing structures for genetic testing are not readily available from the manufacturers, and involve contracts and negotiations. Therefore, further exploration of the genetic testing cost structure is another important avenue for future research.

Acknowledgments: We are grateful to Professor Stefan Scholtes, 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.

References

[1] B. Abdalhamid, C. R. Bilder, E. L. McCutchen, S. H. Hinrichs, S. A. Koepsell, and P. C. Iwen. Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *American Journal of Clinical Pathology*, 153(6):715–718, 2020. doi: https://doi.org/10.1093/ajcp/aqaa064.

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

[3] Advisory Committee on Heritable Disorders in Newborns and Children (ACHDNC). Recommended uniform screening panel core conditions, 2018. *https://www.hrsa.gov/sites/default/files/hrsa/advisory-committees/heritable-disorders/rusp/rusp-uniform-screening-panel.pdf*, Accessed in June 2020.

[4] American Red Cross. What happens to donated blood, 2021. *https://www.redcrossblood.org/donate-blood/blood-donation-process/what-happens-to-donated-blood.html*, Accessed in June 2021.

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

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

[7] H. Aprahamian, D. R. Bish, and E. K. Bish. Optimal group testing: Structural properties and robust solutions, with application to public health screening. *INFORMS Journal on Computing*, 32(4):895–911, 2020.

[8] Baby's First Test, 2019. *http://www.babysfirsttest.org/newborn-screening/states*, Accessed in June 2020.

[9] M. W. Baker, A. E. Atkins, S. K. Cordovado, M. Hendrix, M. C. Earley, and P. M. Farrell. Improving newborn screening for cystic fibrosis using next-generation sequencing technology: A technical feasibility study. *Genetics in Medicine*, 18(3):231–238, 2016.

[10] J. Bang-Jensen and G. Z. Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.

[11] R. Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

[12] D. Borowitz, R. B. Parad, J. K. Sharp, K. A. Sabadosa, K. A. Robinson, M. J. Rock, P. M. Farrell, M. K. Sontag, M. Rosenfeld, S. D. Davis, et al. 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):S106–S116, 2009.

[13] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

[15] M. L. Brandeau. *Allocating Resources to Control Infectious Diseases*, pages 443–464. Springer US, Boston, MA, 2004.

[16] California Department of Public Health (CDPH). Genetic disease screening program, 2019. *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 June 2020.

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

[18] CFTR2, 2020. *https://www.cftr2.org/*, Accessed in April 2020.

[19] A. Chakravarty, J. Orlin, and U. Rothblum. A partitioning problem with additive objective with an application to optimal inventory groupings for joint replenishment. *Operations Research*, 30(5):1018–1022, 1982.

[20] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth. On the LambertW function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

[21] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.

[22] J. Cunningham and L. Taussig. *An Introduction to Cystic Fibrosis for Patients and their Families*. Cystic Fibrosis Foundation, 6th edition, 2013.

[23] R. J. Currier, S. Sciortino, R. Liu, T. Bishop, R. A. Koupaei, and L. Feuchtbaum. Genomic sequencing in cystic fibrosis newborn screening: what works best, two-tier predefined CFTR mutation panels or second-tier CFTR panel followed by third-tier sequencing? *Genetics in Medicine*, 19(10):1159–1163, 2017.

[24] G. R. Cutting. Chapter 58 - cystic fibrosis. In D. Rimoin, R. Pyeritz, and B. Korf, editors, *Emery and Rimoin's Principles and Practice of Medical Genetics*, pages 1 – 54. Academic Press, Oxford, 2013.

[25] Cystic Fibrosis Foundation. All fifty states to screen newborns for cystic fibrosis by 2010, 2009. *https://www.cff.org/About-Us/Media-Center/Press-Releases/All-Fifty-States-to-Screen-Newborns-for-Cystic-Fibrosis-by-2010/*, Accessed in June 2020.

[26] Cystic Fibrosis Foundation. Sweat test, 2017. *https://www.cff.org/What-is-CF/Testing/Sweat-Test/*, Accessed in June 2021.

[27] Cystic Fibrosis Foundation. Carrier testing for cystic fibrosis, 2019. *https://www.cff.org/What-is-CF/Testing/Carrier-Testing-for-Cystic-Fibrosis/*, Accessed in June 2020.

[28] J. L. Deignan, C. Astbury, G. R. Cutting, D. Del Gaudio, A. R. Gregg, W. W. Grody, K. G. Monaghan, and S. Richards. CFTR variant testing: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, pages 1–8, 2020.

[29] E. W. Dijkstra et al. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1): 269–271, 1959.

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

[31] R. Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[32] S. Edouard, E. Prudent, P. Gautret, Z. A. Memish, and D. Raoult. Cost-effective pooling of DNA from nasopharyngeal swab samples for large-scale detection of bacteria by real-time PCR. *Journal of Clinical Microbiology*, 53(3):1002–1004, 2015.

[33] C. Ehre, C. Ridley, and D. J. Thornton. Cystic fibrosis: An inherited disease affecting mucin-producing organs. *The International Journal of Biochemistry & Cell Biology*, 52:136–145, 2014.

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

[35] H. El Hajj, D. R. Bish, and E. K. Bish. Optimal genetic screening for cystic fibrosis. *Operations Research*, Accepted, 2021.

[36] H. El Hajj, D. R. Bish, and E. K. Bish. Equity in genetic newborn screening. *Naval Research Logistics*, 68(1):44–64, 2021.

[37] H. El Hajj, D. R. Bish, E. K. Bish, and H. Aprahamian. Screening multi-dimensional heterogeneous populations for infectious diseases under scarce testing resources, with application to covid-19. *Naval Research Logistics*, 2021. doi: https://doi.org/10.1002/nav.21985.

[38] W. Eng, V. A. LeGrys, M. S. Schechter, M. M. Laughon, and P. M. Barker. Sweat-testing in preterm and full-term infants less than 6 weeks of age. *Pediatric Pulmonology*, 40(1):64–67, 2005.

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

[40] Food and Drug Administration, 2005. *https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/cftr-gene-mutation-detection-systems-class-ii-special-controls-guidance-industry-and-fda-staff*, Accessed in June 2021.

[41] L. R. Ford Jr. Network flow theory. Technical report, Rand Corp Santa Monica Ca, 1956.

[42] G.-G. P. Garcia, M. S. Lavieri, R. Jiang, M. A. McCrea, T. W. McAllister, S. P. Broglio, C. C. Investigators, et al. Data-driven stochastic optimization approaches to determine decision thresholds for risk estimation models. *IISE Transactions*, 52(10):1098–1121, 2020.

[43] A. Gilányi, editor. *Subadditive Functions*, pages 455–481. Birkhäuser Basel, Basel, 2009.

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

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

[46] L. Henneman, I. Bramsen, L. Van Kempen, M. B. Van Acker, G. Pals, H. E. Van Der Horst, H. J. Adèr, H. M. Van Der Ploeg, and P. Leo. Offering preconceptional cystic fibrosis carrier couple screening in the absence of established preconceptional care services. *Public Health Genomics*, 6(1):5–13, 2003.

[47] R. J. Hilsden, S. J. Heitman, B. Mizrahi, S. A. Narod, and R. Goshen. Prediction of findings at screening colonoscopy using a machine learning algorithm based on complete blood counts (colonflag). *PLOS ONE*, 13:1–9, 11 2018. doi: 10.1371/journal.pone.0207848.

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

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

[50] John Hopkins Medicine. Cystic fibrosis and CF-related disorders NGS panel, 2019. *https://www.hopkinsmedicine.org/dnadiagnostic/tests/tests/cystic-fibrosis-and-cf-related-disorders-ngs-panel*, Accessed in June 2020.

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

[52] A. Kammesheidt, M. Kharrazi, S. Graham, S. Young, M. Pearl, C. Dunlop, and S. Keiles. Comprehensive genetic analysis of the cystic fibrosis transmembrane conductance regulator from dried blood specimens–implications for newborn screening. *Genetics in Medicine*, 8(9):557–562, 2006.

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

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

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

[56] H.-Y. Kim, M. G. Hudgens, J. M. Dreyfuss, D. J. Westreich, and C. D. Pilcher. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, 63(4):1152–1163, 2007.

[57] S. Y. Kim, J. Lee, H. Sung, H. Lee, M. G. Han, C. K. Yoo, S. W. Lee, and K. H. Hong. Pooling upper respiratory specimens for rapid mass screening of COVID-19 by real-time RT-PCR. *Emerging Infectious Diseases*, 26(10):2469, 2020.

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

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

[60] A. Krook, I. M. Stratton, and S. O'Rahilly. Rapid and simultaneous detection of multiple mutations by pooled and multiplex single nucleotide primer extension: application to the study of insulin-responsive glucose transporter and insulin receptor mutations in non-insulin-dependent diabetes. *Human Molecular Genetics*, 1(6):391–395, 1992.

[61] E. J. Kuipers, T. Rösch, and M. Bretthauer. Colorectal cancer screening—optimizing current strategies and new directions. *Nature Reviews Clinical Oncology*, 10(3):130, 2013.

[62] V. A. LeGrys. Newborn Screening for Cystic Fibrosis. *Laboratory Medicine*, 33(3):212–213, 2002.

[63] J. L. Lewis, V. M. Lockary, and S. Kobic. Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia Trachomatis* and *Neisseria Gonorrhoeae*. *Sexually Transmitted Diseases*, 39(1):46–8, 2012.

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

[65] A. M. Lutz, J. K. Willmann, F. V. Cochran, P. Ray, and S. S. Gambhir. Cancer screening: A mathematical model relating secreted blood biomarker levels to tumor sizes. *PLOS Medicine*, 5:1–11, 08 2008.

[66] A. C. V. Mattar, C. Leone, J. C. Rodrigues, and F. V. Adde. Sweat conductivity: an accurate diagnostic test for cystic fibrosis? *Journal of Cystic Fibrosis*, 13(5):528–533, 2014.

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

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

[69] N. T. Nguyen, H. Aprahamian, E. K. Bish, and D. R. Bish. A methodology for deriving the sensitivity of pooled testing, based on viral load progression and pooling dilution. *Journal of Translational Medicine*, 17(1):252, 2019.

[70] North Carolina State Laboratory of Public Health (NCLPH). Newborn screening: Reporting - cystic fibrosis, 2016. *https://slph.ncpublichealth.com/newborn/Reporting.asp*, Accessed in June 2020.

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

[72] R. Phatarfod and A. Sudbury. The use of a square array scheme in blood testing. *Statistics in Medicine*, 13(22):2337–2343, 1994.

[73] C. D. Pilcher, D. Westreich, and M. G. Hudgens. Group testing for SARS-CoV-2 to enable rapid scale-up of testing and real-time surveillance of incidence. *The Journal of Infectious Diseases*, 2020.

[74] M. Rios, S. Daniel, C. Chancey, I. K. Hewlett, and S. L. Stramer. West nile virus adheres to human red blood cells in whole blood. *Clinical Infectious Diseases*, 45(2):181–186, 2007.

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

[76] G. Rours, R. Verkooyen, H. Willemse, E. van der Zwaan, A. van Belkum, R. de Groot, H. Verbrugh, and J. Ossewaarde. Use of pooled urine samples and automated DNA isolation to achieve improved sensitivity and cost-effectiveness of large-scale testing for Chlamydia trachomatis in pregnant women. *Journal of Clinical Microbiology*, 43(9):4684–4690, 2005.

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

[78] M. Schmidt, A. Werbrouck, N. Verhaeghe, E. De Wachter, S. Simoens, L. Annemans, and K. Putman. A model-based economic evaluation of four newborn screening strategies for cystic fibrosis in flanders, belgium. *Acta Clinica Belgica*, pages 1–9, 2019.

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

[80] P. R. Sosnay, K. R. Siklosi, F. Van Goor, K. Kaniecki, H. Yu, N. Sharma, A. S. Ramalho, M. D. Amaral, R. Dorfman, J. Zielenski, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nature Genetics*, 45(10):1160–1167, 2013.

[81] L. N. Steimle and B. T. Denton. *Markov Decision Processes for Screening and Treatment of Chronic Diseases*, pages 189–222. Springer International Publishing, Cham, Switzerland, 2017.

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

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

[84] M. L. Turgeon. *Linne & Ringsrud's Clinical Laboratory Science-E-Book: The Basics and Routine Techniques*. Elsevier Health Sciences, 2015.

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

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

[87] J. Wells, M. Rosenberg, G. Hoffman, M. Anstead, and P. M. Farrell. A decision-tree approach to cost comparison of newborn screening strategies for cystic fibrosis. *Pediatrics*, 129(2):e339–e347, 2012.

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

[89] Wisconsin State Laboratory of Hygiene (WSLH). Newborn blood screening panel of diseases: Cystic fibrosis, 2018. *http://www.slh.wisc.edu/clinical/newborn/health-care-professionals-guide/nbs-test-panel-of-diseases/#cf*, Accessed in June 2020.

APPENDIX

A Summary of Acronyms and Mathematical Notation

A.1 Summary of Acronyms

CF	: Cystic fibrosis
CFTR	: Cystic fibrosis transmembrane conductance regulator
NBS	: Newborn screening
IRT	: Immunoreactive trypsinogen
SC	: Sweat chloride test
DNA	: Deoxyribonucleic acid test (genetic test)
P-DNA	: Pooled *DNA* test (pooled genetic test)

A.2 Summary of Mathematical Notation

All vectors are denoted in bold-face.

Parameters:

Ω	: Set of all CF-causing variants (with cardinality m), labeled following a non-increasing order of variant frequency
$q_i,\ i \in \Omega$	: Conditional probability that a CFTR gene has a mutation of variant i, given that the gene has a mutation (i.e., the frequency of variant i), where $\sum_{i\in\Omega} q_i = 1$
$S(l) = \{1, \cdots, l\}, l = 1, \cdots, m$	: Ordered variant set, containing the l highest frequency variants in set Ω
c_f	: Fixed cost (e.g., consumables, labor) per genetic test
$c_v\left(z(\boldsymbol{x})\right)$	: Variable cost (e.g., dNTPs, enzymes, variant-specific reagents) per genetic test, which is a non-decreasing function of panel size, $z(\boldsymbol{x})$
c_{SC}	: Cost per *SC* test
c_{IRT}	: Cost per *IRT* test
B	: Testing budget per newborn
$\overline{z}$	: Panel size limit
$\overline{t}$	: Pool size limit
$\overline{\eta}$	: Limit on number of panels

Decision variables:

$\boldsymbol{x}^k = \left(x_i^k\right)_{i\in\Omega}, k = 1, \cdots, \overline{\eta}$	: Binary vector, where $x_i^k = 1$ if variant $i \in \Omega$ is included in panel k, and $x_i^k = 0$ otherwise
$t^k \in Z^+, k = 1, \cdots, \overline{\eta}$	: Integer pool size for panel k

Functions of decision variables (represented in terms of panel vector $\boldsymbol{x}$ and pool size t):

$I_{\{\boldsymbol{x}>\boldsymbol{0}\}}$	: Indicator variable, where $I_{\{\boldsymbol{x}>\boldsymbol{0}\}} = 1$ if $\boldsymbol{x} > \boldsymbol{0}$, and $I_{\{\boldsymbol{x}>\boldsymbol{0}\}} = 0$ otherwise
$\eta((\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}) = \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}}$	: Number of panels used in partition $(\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}$
$z(\boldsymbol{x}) = \sum_{i\in\Omega} x_i$	: Panel size
$y(\boldsymbol{x}) = \sum_{i\in\Omega} x_i q_i$	: Panel coverage
$p(\boldsymbol{x})$	: Probability that a random newborn has at least one mutation, of any variant covered by the panel
$D(\boldsymbol{x}, t)$	: Per newborn expected number of tests for *P-DNA*
$C(\boldsymbol{x}, t)$	: Per newborn expected cost for *P-DNA*
$TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})$	: Per newborn expected total cost, including the cost of *P-DNA* and the diagnostic *SC* test
$S(\boldsymbol{x}) = \{i \in \Omega : x_i = 1\}$	: Variants corresponding to panel vector $\boldsymbol{x}$, with complement $\overline{S}(\boldsymbol{x}) = \Omega \setminus S(\boldsymbol{x})$
$\boldsymbol{x}^{12\cdots\overline{\eta}} = \sum_{k=1}^{\overline{\eta}} \boldsymbol{x}^k$	: Combined variant vector for all panels of *P-DNA*, corresponding to variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$

Random event and random variables:

$FN(\boldsymbol{x}^{12\cdots\overline{\eta}})$	: Event that a random newborn is a false-negative, that is, the newborn is CF-positive (i.e., with two mutations, one on each CFTR gene) and both mutation variants are excluded from variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ (i.e., not screened for)
N	: Number of mutations (of any variant in set Ω) a random newborn in the general population has, with sample space $\mathcal{S}(N) = \{0, 1, 2\}$ (respectively denoting the mutation-free, CF-carrier, and CF-positive status) and pmf $P_N(.)$
N_{post}	: Number of mutations (of any variant in set Ω) a random newborn in the post-*IRT* population has, with sample space $\mathcal{S}(N_{post}) = \{0, 1, 2\}$ and pmf $P_{N_{post}}(.)$ (used only in the case study)

Optimal solutions and functions of optimal solutions:

$t^*(\boldsymbol{x})$	: Optimal integer pool size for panel $\boldsymbol{x}$
$(\boldsymbol{x}^{k*}(l))_{k=1,\cdots,\overline{\eta}},\ l = 1, \cdots, m$	: Optimal partition of set $S(l)$
$\eta^*((\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}) = \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^{k*}>\boldsymbol{0}\}}$	: Number of panels used in optimal partition $(\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}$
$B^l = TC((\boldsymbol{x}^{k*}(l), t^*(x^{k*}(l)))_{k=1,\cdots,\overline{\eta}}), l = 1, \cdots, m$	: Budget breakpoint corresponding to variant set $S(l)$

B Derivations for §2.2

The following expressions extend those in [35] to the *P-DNA* scheme,to consider a generic form of the pmf for random variable N, rather than the specific form in [35] that is derived based on certain assumptions on the parent population, including "uniform mixing" to produce an offspring. Recall that $q_i, i \in \Omega$, is the conditional probability that a CFTR gene has a mutation of variant i, given that the gene has a mutation, hence $\boldsymbol{q} = (q_i)_{i\in\Omega}$: $\sum_{i\in\Omega} q_i = 1$ (Appendix A.2). Then, an *FN* event happens when a newborn is CF-positive, that is, with two mutations (one on each CFTR gene), *and* both mutation variants are excluded from variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, that is, they are not screened for. We define random variable $V_i, i \in \Omega$, as the number of mutations of variant i a random newborn has (on their two CFTR genes), with sample space $\mathcal{S}(V_i) = \{0, 1, 2\}$, and $N = \sum_{i\in\Omega} V_i$. Then, $Pr(V_i = 1|N = 1) = q_i$, $Pr(V_i = 2|N = 2) = q_i^2$, and $Pr(V_i = 1, V_j = 1|N = 2) = 2q_iq_j, i, j \in \Omega : i < j$ (domain defined to eliminate symmetries). Then, the *FN* probability for any given binary vector $\boldsymbol{x}^{12\cdots\overline{\eta}}$ follows:

$$
\begin{aligned}
Pr\left(FN(\boldsymbol{x}^{12\cdots\overline{\eta}})\right) &= \left(\sum_{i,j\in\overline{S}(\boldsymbol{x}^{12\cdots\overline{\eta}}):i<j} Pr(V_i = 1, V_j = 1|N = 2) + \sum_{i\in\overline{S}(\boldsymbol{x}^{12\cdots\overline{\eta}})} Pr(V_i = 2|N = 2)\right) P_N(2) \\
&= \left(\sum_{i,j\in\overline{S}(\boldsymbol{x}^{12\cdots\overline{\eta}}):i<j} 2q_iq_j + \sum_{i\in\overline{S}(\boldsymbol{x}^{12\cdots\overline{\eta}})} q_i^2\right) P_N(2) \\
&= \left(\sum_{i,j\in\Omega:i<j} 2\left(1 - x_i^{12\cdots\overline{\eta}}\right)\left(1 - x_j^{12\cdots\overline{\eta}}\right) q_iq_j + \sum_{i\in\Omega}\left(1 - x_i^{12\cdots\overline{\eta}}\right) q_i^2\right) P_N(2) \\
&= \left(\sum_{i,j\in\Omega}\left(1 - x_i^{12\cdots\overline{\eta}}\right)\left(1 - x_j^{12\cdots\overline{\eta}}\right) q_iq_j\right) P_N(2) = \left(1 - \sum_{i\in\Omega} x_i^{12\cdots\overline{\eta}} q_i\right)^2 P_N(2).
\end{aligned}
$$

Next we derive an expression for $p(\boldsymbol{x})$, i.e., the probability that a random newborn has at least one mutation, of any variant covered by panel $\boldsymbol{x}$, which happens when the panel contains at least one of the mutation variant(s) of a CF-positive newborn (i.e., for a newborn with two mutations, $N = 2$), or the single mutation variant of a CF-carrier (i.e., for a newborn with one mutation, $N = 1$):

$$
\begin{aligned}
p(\boldsymbol{x}) = &\left(\sum_{i\in S(\boldsymbol{x})} Pr(V_i = 1|N = 1)\right) P_N(1) + \left(\sum_{i\in S(\boldsymbol{x}),j\in\overline{S}(\boldsymbol{x}):i<j} Pr(V_i = 1, V_j = 1|N = 2)\right) P_N(2) \\
&+ \left(\sum_{i\in\overline{S}(\boldsymbol{x}),j\in S(\boldsymbol{x}):i<j} Pr(V_i = 1, V_j = 1|N = 2)\right) P_N(2)
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\sum_{i,j\in S(\boldsymbol{x}):i<j} Pr(V_i=1,V_j=1|N=2)+\sum_{i\in S(\boldsymbol{x})} Pr(V_i=2|N=2)\right)P_N(2)\\
=&\left(\sum_{i\in S(\boldsymbol{x})} q_i\right)P_N(1)+\left(\sum_{i\in S(\boldsymbol{x}),j\in\overline{S}(\boldsymbol{x}):i<j} 2q_iq_j+\sum_{i\in\overline{S}(\boldsymbol{x}),j\in S(\boldsymbol{x}):i<j} 2q_iq_j+\sum_{i,j\in S(\boldsymbol{x}):i<j} 2q_iq_j+\sum_{i\in S(\boldsymbol{x})} q_i^2\right)P_N(2)\\
=&\left(\sum_{i\in\Omega} x_iq_i\right)P_N(1)+\left(\sum_{i,j\in\Omega} 2x_i(1-x_j)q_iq_j+\sum_{i,j\in\Omega:i<j} 2x_ix_jq_iq_j+\sum_{i\in\Omega} x_iq_i^2\right)P_N(2)\\
=&\left(\sum_{i\in\Omega} x_iq_i\right)P_N(1)+\left(\sum_{i,j\in\Omega} 2x_i(1-x_j)q_iq_j+\sum_{i,j\in\Omega:i\neq j} x_ix_jq_iq_j+\sum_{i\in\Omega} x_iq_i^2\right)P_N(2)\\
=&\left(\sum_{i\in\Omega} x_iq_i\right)P_N(1)+2\left(\sum_{i,j\in\Omega} x_i(1-x_j)q_iq_j\right)P_N(2)+\left(\sum_{i,j\in\Omega} x_ix_jq_iq_j\right)P_N(2)\\
=&\left(\sum_{i\in\Omega} x_iq_i\right)P_N(1)+P_N(2)\left(2\left(\sum_{i\in\Omega} x_iq_i\right)-\left(\sum_{i,j\in\Omega} x_ix_jq_iq_j\right)\right)\\
=&P_N(1)\ y(\boldsymbol{x})+P_N(2)\ \left(2y(\boldsymbol{x})-(y(\boldsymbol{x}))^2\right).
\end{aligned}
$$

Next we extend the pooled testing expressions in [7] to the *P-DNA* scheme. The *DNA* test has perfect analytical sensitivity and specificity (Assumption ($\boldsymbol{A_2}$)). Then, for any panel composition $\boldsymbol{x}>\boldsymbol{0}$ and pool size $t\in Z^+$, $t\geq 2$, the pooled test outcome will be positive only if at least one of the t subjects in the pool has at least one variant covered by the panel, i.e., with probability $1-(1-p(\boldsymbol{x}))^t$; and will be negative otherwise. Hence, $D(\boldsymbol{x},t)$, the per newborn expected number of tests for *P-DNA*, is $\frac{1}{t}$ per subject if the pooled test's outcome is negative (i.e., there is only one test for t subjects), and $\frac{1}{t}+1$ per subject if the pooled test's outcome is positive (i.e., each subject in the pool is retested via an individual test), leading to:

$$D(\boldsymbol{x},t)=\frac{1}{t}+1-(1-p(\boldsymbol{x}))^t,$$

and the expressions for $C(.)$ and $TC(.)$ (Eqs. (3)-(4)) follow by the testing cost structure and decision rules (Figs. 1-2).

C Optimal Pool Size: Characterization and Properties

Considering Dorfman pooling, §C.1 characterizes the optimal pool size, and §C.2 studies the impact of a common pool size constraint. Then §C.3 explores the use of an alternative pooling method, namely array pooling, in our problem setting.

C.1 Analytical Results on the Optimal Pool Size under Dorfman Pooling

We first extend a result from the literature to our problem setting, to condition on panel composition. To this end, we utilize the Lambert W function, $W(\alpha)e^{W(\alpha)}=\alpha$, $\alpha\in\Re$, and use $W_0(.)$ and $W_{-1}(.)$ to respectively refer to the principle and the secondary branches of the Lambert W function, which are well-characterized in the literature [20].

Property C.1. *(Theorem 1 from [7], expanded to consider panel composition $\boldsymbol{x}$) For a given panel $\boldsymbol{x}>\boldsymbol{0}$, an optimal continuous pool size, $\tilde{t}(\boldsymbol{x})$, can be obtained as follows:*

$$\tilde{t}(\boldsymbol{x})=\underset{t\geq 0}{\operatorname{argmin}}\{D(\boldsymbol{x},t)\}=\frac{2}{ln(1-p(\boldsymbol{x}))}W_0\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x})}\right)}\right).$$

Then, an optimal integer pool size, $t^*(\boldsymbol{x})$, constrained by pool size limit, $\bar{t}$, follows:

$$t^*(\boldsymbol{x}) = \min\left\{ \underset{t\in\left\{\lfloor \bar{t}(\boldsymbol{x})\rfloor, \lceil \bar{t}(\boldsymbol{x})\rceil, \bar{t}\right\}}{\operatorname{argmin}} \left\{D(\boldsymbol{x},t)\right\}, \bar{t} \right\},$$

where $t^*(\boldsymbol{x})$ is non-increasing in $p(\boldsymbol{x})$. Further, pooled testing (with optimal pool size) reduces the per subject expected number of tests over individual testing if and only if $p(\boldsymbol{x}) < 0.308$.

Next we characterize an optimal pool size for *P-DNA* under a common pool size constraint for all panels, which, by Remark 1, is the minimizer to the expression, $\sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}} \times C(\boldsymbol{x}^k, t)$.

Lemma C.1. *Consider a P-DNA scheme, under the additional constraint that all panels use a common pool size. For given panels $\boldsymbol{x}^k > \boldsymbol{0}$, $k = 1, \cdots, \overline{\eta}$, the optimal common, integer pool size, $t^{c*}((\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}})$, i.e., the minimizer to $\sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\boldsymbol{0}\}} \times C(\boldsymbol{x}^k, t)$, over $t \in \{2, \cdots, \bar{t}\}$, belongs to the set,*

$$t^{c*}((\boldsymbol{x}^k)_{k=1,\cdots,\overline{\eta}}) \in \left[\min_{k=1,\cdots,\overline{\eta}} \left\{t^*(\boldsymbol{x}^k)\right\}, \min\left\{ \max_{k=1,\cdots,\overline{\eta}} \left\{ \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}^k))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x}^k)}\right)}\right)\right\rceil \right\}, \bar{t} \right\}\right] \cup \{\bar{t}\}.$$

Proof. From [7], we have that, for a given $\boldsymbol{x}$, $D(\boldsymbol{x},t)$ is strictly decreasing in t, for $t \in Z^+ : 2 \le t \le t^*(\boldsymbol{x})$. Then, since the term, $c_f + c_v\left(z(\boldsymbol{x})\right)$, is independent of t, we have that $C(\boldsymbol{x},t)$ is strictly decreasing in t in this range. Further, since the remaining terms in $TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})$, namely, $C(\boldsymbol{x}^{k'}, t^{k'})$, $k' = 1, \cdots, k-1, k+1, \cdots, \overline{\eta}$, and $p\left(\boldsymbol{x}^{12\cdots\overline{\eta}}\right) \times c_{SC}$ (Eq. (4)), are independent of t^k, $k = 1, \cdots, \overline{\eta}$, we have that $TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})$ is strictly decreasing in t^k, $k = 1, \cdots, \overline{\eta}$, for $t^k \in Z^+ : 2 \le t^k \le t^*(\boldsymbol{x}^k)$. Thus, it follows that $TC((\boldsymbol{x}^k, t^c)_{k=1,\cdots,\overline{\eta}})$ is strictly decreasing in t^c, for $t^c \le \min_{k=1,\cdots,\overline{\eta}}\{t^*(x^k)\}$, and the lower bound of the range follows.

Regarding the upper bound, first note that $\frac{2}{ln(1-p(\boldsymbol{x}))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x})}\right)}\right)$ is a local maxima of $D(\boldsymbol{x},t)$, and $D(\boldsymbol{x},t)$ is strictly decreasing in t, for $t > \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right\rceil$ [7]. Then, it follows that $TC((\boldsymbol{x}^k, t^c)_{k=1,\cdots,\overline{\eta}})$ is strictly decreasing in t^c, for $t^c > \max_{k=1,\cdots,\overline{\eta}} \left\{ \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}^k))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x}^k)}\right)}\right)\right\rceil \right\}$.
There are two possible cases:
Case 1: $\bar{t} \le \max_{k=1,\cdots,\overline{\eta}} \left\{ \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}^k))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x}^k)}\right)}\right)\right\rceil \right\}$.
Because any pool size cannot exceed the pool size limit $\bar{t}$, the result trivially follows.

Case 2: $\bar{t} > \max_{k=1,\cdots,\overline{\eta}} \left\{ \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}^k))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x}^k)}\right)}\right)\right\rceil \right\}$.
Because $TC((\boldsymbol{x}^k, t^c)_{k=1,\cdots,\overline{\eta}})$ is strictly decreasing in t^c, for $t^c > \max_{k=1,\cdots,\overline{\eta}} \left\{ \left\lceil \frac{2}{ln(1-p(\boldsymbol{x}^k))} W_{-1}\left(-\frac{1}{2}\sqrt{ln\left(\frac{1}{1-p(\boldsymbol{x}^k)}\right)}\right)\right\rceil \right\}$,
it is sufficient to consider $t^c(\boldsymbol{x}^1, \cdots, \boldsymbol{x}^{\overline{\eta}}) = \bar{t}$, which is the optimal solution in this case, and the result follows. □

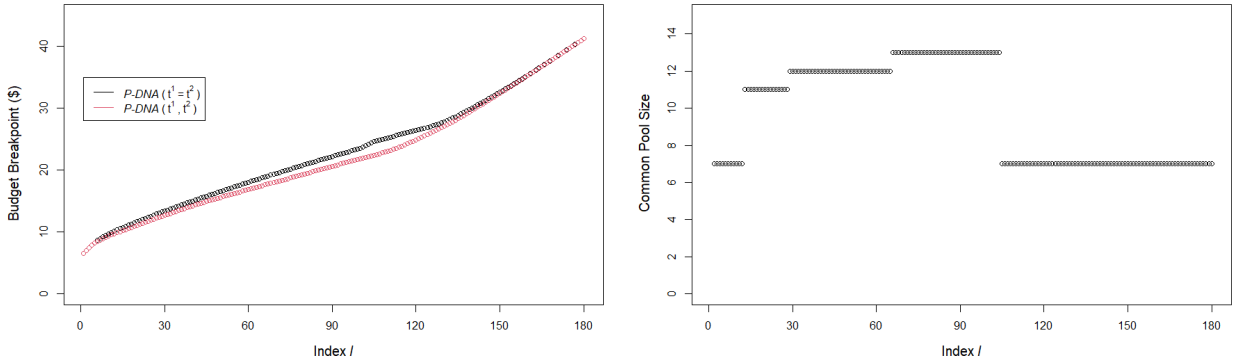
Thus, while an optimal common pool size is never smaller than the minimum of the "non-common" pool sizes for all the panels, it can be larger than their maximum.

Remark C.1. In the general setting without a common pool size constraint, the optimal integer pool size $t^*(\boldsymbol{x})$ for each panel $\boldsymbol{x} > \boldsymbol{0}$ is a step function of $p(\boldsymbol{x})$ only, independently of the other panels (Remark 1 and Property C.1). Then, optimal pool sizes for all panels will be equal in the general setting only when all $p(\boldsymbol{x}^k), k = 1, \cdots, \eta$, are sufficiently close, i.e., within some range dictated by the step function.

C.2 Impact of a Common Pool Size Constraint

We use the case study data (§4) to determine the optimal pool sizes for *P-DNA*($\overline{\eta} = 2$) in (i) the unconstrained pool size setting, where pool sizes t^{1*} and t^{2*} are derived from Property C.1, and (ii) the common pool size setting, where the common t^{c*} is derived from Lemma C.1. Fig. 6(a) depicts the budget breakpoints in both settings, and Fig. 6(b) displays the optimal common pool sizes. In this numerical study, common pool sizes do not turn out to be optimal for the unconstrained problem, because the most prevalent variant, which is included in the optimal variant set for each index $l = 1, \cdots, m$, has a frequency that exceeds the frequencies of all other variants combined (see Remark C.1). However, the numerical study indicates that allowing panel pool sizes to differ yields only a small benefit. Further, the common pool size is at most 13 over the family of optimal designs, while in the unconstrained setting the pool size can be as high as 47. This indicates that a pool size limit, $\bar{t}$, of 13 or higher will have only a minor impact on the efficiency of the *P-DNA* design. We also observe that an optimal variant partition under a common pool size restriction is similar to its unconstrained counterpart, in that the panel that contains more variants has lower frequency variants. Thus, most benefits of *P-DNA*($\overline{\eta} = 2$) can be attained through a simplified process with common pool sizes, which might be a beneficial strategy for a testing laboratory.

Figure 6: Budget breakpoints, B^l, for *P-DNA*($\overline{\eta} = 2$) with and without a common pool size constraint, and optimal common pool sizes, t^{c*}, for a family of optimal designs, as a function of index l



(a) Budget breakpoint vs design index

(b) Common pool size vs design index

C.3 An Alternative Pooling Method: Array Pooling

GP considers *P-DNA* integrated with the Dorfman pooling method, which is the most commonly studied and used pooling method (see §1). In this section, we provide some preliminary analysis on the potential benefits of another pooling method, namely array pooling, which utilizes non-overlapping pools, as opposed to the overlapping pools used in Dorfman pooling. In this section, we discuss whether the structural results for Dorfman pooling (§2-3) extend to array pooling, and compare the performance of array pooling and Dorfman pooling to derive insight. In particular, we consider the 2-stage square array pooling (e.g., [56, 72]), in which specimens from t^2 subjects are placed in a $t \times t$ array, and each row pool and each column pool (each containing t specimens) are separately tested with one *DNA* test per pool, resulting in $2t$ pooled tests. After pooled test outcomes are obtained, different individual retesting rules can be applied.

In our setting, with a test having perfect analytical sensitivity (hereafter, "sensitivity") and specificity (Assumptions ($\boldsymbol{A_2}$)-($\boldsymbol{A_3}$)), if a row (column) pool tests positive, then at least one column (row) pool must also test positive. In addition, if there is exactly one positive-testing row (column) and at least one positive-testing column (row), then the locations of all true-positive specimens are known with certainty, that is, there is no ambiguity. In other words, ambiguity, hence the need for individual retesting, arises only when there are at least two positive-testing rows, *and* at least two positive-testing columns. Deriving an expression for the expected number of retests is cumbersome because of the need to condition on both the number of

true-positives, and all possible locations of those true-positives, in the array. Hence, the relevant literature, e.g., [56], makes a simplifying assumption, that the outcomes of row pools and column pools are conditionally independent, given the true-positivity status of the subject at their intersection. In particular, [56] considers a retesting rule which, for a perfect test, reduces to individually retesting all subjects that lie at the intersection of each positive-testing row and each positive-testing column; and derives the expected number of tests under the conditional independence assumption. Because this specific retesting rule applies to both ambiguous and unambiguous subjects, the expression on the expected number of tests in [56], given in Remark C.2, provides an upper bound on the expected number of tests.

Remark C.2. (Eq. (9) in [56] at perfect sensitivity and specificity, expanded to consider panel composition $\boldsymbol{x}$) Assume that the outcomes of row pools and column pools are conditionally independent, given the true-positivity status of the subject at their intersection, and the test has perfect sensitivity and specificity. Then, for a given panel $\boldsymbol{x}$ and a $t \times t$ array, $t \in Z^+$, the per newborn expected number of tests under array pooling (i.e., pooled tests plus individual retests for all ambiguous subjects), denoted by $D^A(\boldsymbol{x}, t)$, is bounded from above as follows:

$$D^A(\boldsymbol{x}, t) \leq \frac{2}{t} + 1 - 2(1 - p(\boldsymbol{x}))^t + (1 - p(\boldsymbol{x}))^{2t-1}. \tag{17}$$

To study the *P-DNA* design problem under array pooling, one can replace the expression in Eq. (2) with the upper bound in Eq. (17), which serves as an approximation, and the **GP** formulation continues to hold, with the expected total cost expression in Eq. (4) updated to incorporate the upper bound in Eq. (17). However, some key structural properties established under Dorfman pooling (§2-3) do not necessarily extend to array pooling, as stated in Remark C.3.

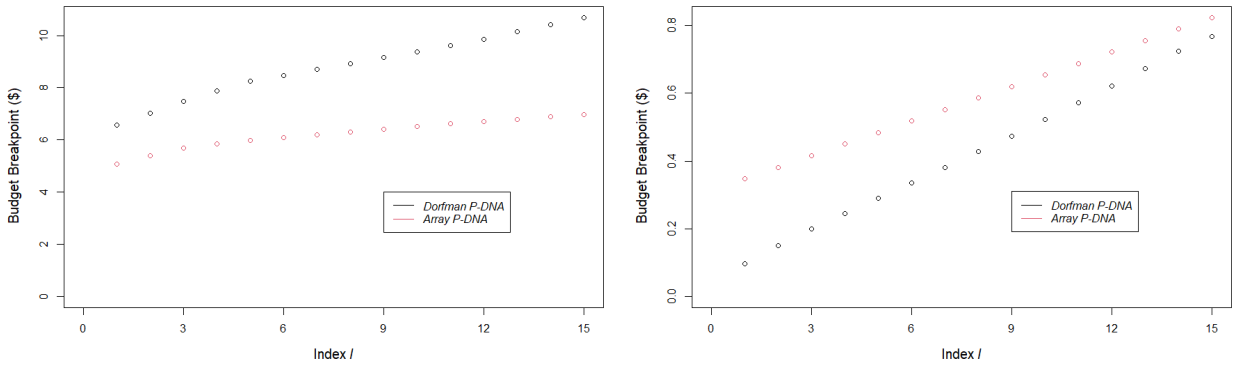
Remark C.3. The following results do not necessarily extend to *P-DNA* under array pooling:

1. Property C.1 (optimal pool size for panel $\boldsymbol{x}$): [7] establishes this result under Dorfman pooling by writing $\frac{\partial^2 D(\boldsymbol{x},t)}{(\partial t)^2}$ in the form of the Lambert W function, $W(\alpha)e^{W(\alpha)} = \alpha$, $\alpha \in \Re$, and using well-established properties of the Lambert function (e.g., [20]). In particular, for $p(\boldsymbol{x}) < 1 - e^{-e^{-1}}$, $\frac{\partial^2 D(\boldsymbol{x},t)}{(\partial t)^2}$ has exactly two real roots, hence $\frac{\partial D(\boldsymbol{x},t)}{\partial t}$ has exactly two stationary points, which respectively correspond to the principle and secondary branches of the Lambert W function, and the global minimizer of $D(\boldsymbol{x}, t)$, denoted by $\tilde{t}(\boldsymbol{x})$, can be characterized, as done in Property C.1 (see [7]). The Lambert function equivalence of the second derivative no longer holds for array pooling under the $D^A(\boldsymbol{x}, t)$ approximation in Eq. (17), and hence Property C.1 does not readily extend to array pooling.
2. Lemma 1 (optimality of an ordered partition for a given variant set). For array pooling, the expected number of tests $D^A(\boldsymbol{x}, t)$, hence $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})$, is not necessarily concave in panel coverage, $y(\boldsymbol{x})$. The variant partition problem with an arbitrary $TC(.)$ function is NP-hard [19].

Consequently, under array pooling, one needs to resort to enumeration over both all possible pool sizes, $t = 2, \cdots, \bar{t}$, and all possible variant partitions for each variant set.

Next we perform a preliminary numerical study to compare the performance of *P-DNA*($\overline{\eta} = 2$), integrated with either Dorfman pooling or array pooling. We use the data in §4, but due to the need for enumeration under array pooling, we construct two problem instances having only 15 of the variants in set Ω: (i) the most frequent variants, $\{1, \cdots, 15\}$, and (ii) medium frequency variants, $\{61, \cdots, 75\}$.

Figure 7: Budget breakpoints, B^l, for *P-DNA*($\overline{\eta} = 2$) under Dorfman pooling vs array pooling



(a) Problem instance with variants $\{1, \cdots, 15\}$ (b) Problem instance with variants $\{61, \cdots, 75\}$

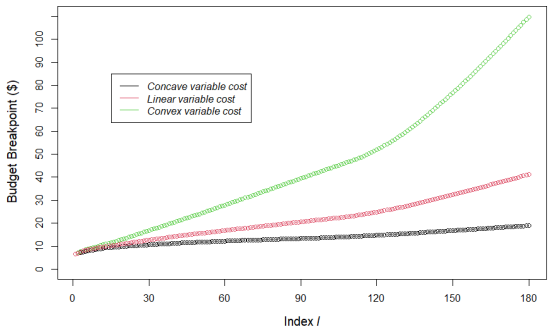
Whether Dorfman pooling or array pooling leads to a lower cost (i.e., a lower budget breakpoint for the same index l) when integrated with *P-DNA* depends on variant frequencies. While a lower cost is achieved by array pooling in the high frequency variant setting (Fig. 7 (a)), this is achieved by Dorfman pooling in the lower frequency variant setting (Fig. 7 (b)). Further analysis indicates that both Dorfman pooling and array pooling benefit from the use of a 2-panel design. In particular, the 2-panel design cost-dominates the 1-panel design for all $l \geq 5$ under Dorfman pooling, and for all $l \geq 4$ under array pooling in the high frequency variant setting; and for all $l \geq 15$ under both Dorfman and array pooling in the low frequency variant setting.

D Supporting Numerical Analysis

D.1 Sensitivity on the Form of the Variable Cost Function

To study the sensitivity of *P-DNA* budget breakpoints and design to the form of the variable cost function, $c_v(.)$, we consider *P-DNA*($\overline{\eta} = 2$) with a fixed cost of $c_f = \$10$, and three variable cost functions, governed by parameter r: $c_v(z(\boldsymbol{x})) = \$(z(\boldsymbol{x}))^r$, for $r = 1$ (linear), $r = 1.25$ (convex), and $r = 0.75$ (concave). All other data come from the case study (§4). Fig. 8 displays the budget breakpoints for a family of optimal designs for the *P-DNA*($\overline{\eta} = 2$) process for each variable cost function.

Figure 8: Budget breakpoints for the *P-DNA*($\overline{\eta} = 2$) process for a family of optimal designs under linear, convex, and concave functions for $c_v(.)$, as a function of index l

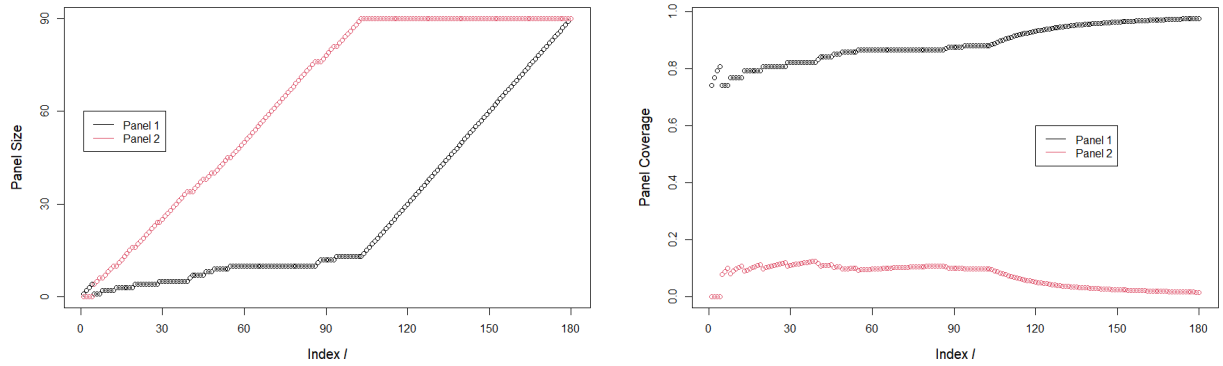


Not surprisingly, the concave (convex) cost function lowers (raises) the budget breakpoint for each index l, over the linear case (Fig. 8). For example, for index $l = 90$, the budget breakpoint reduces by 35.1%, and increases by 91.6%, respectively, over the linear case, under the concave and convex functions. Interestingly,

while the concave (convex) function expands (shrinks) the cost-dominance region for the 1-panel design compared to the linear case, it does not alter the order of cost-dominance: 1-panel design remains optimal for small variant sets (i.e., $l = 1, \cdots, 4$ for the linear case, $l = 1, \cdots, 12$ for the concave case, and $l = 1, 2$ for the convex case), and 2-panel design remains optimal for larger variant sets.

D.2 Additional Figures

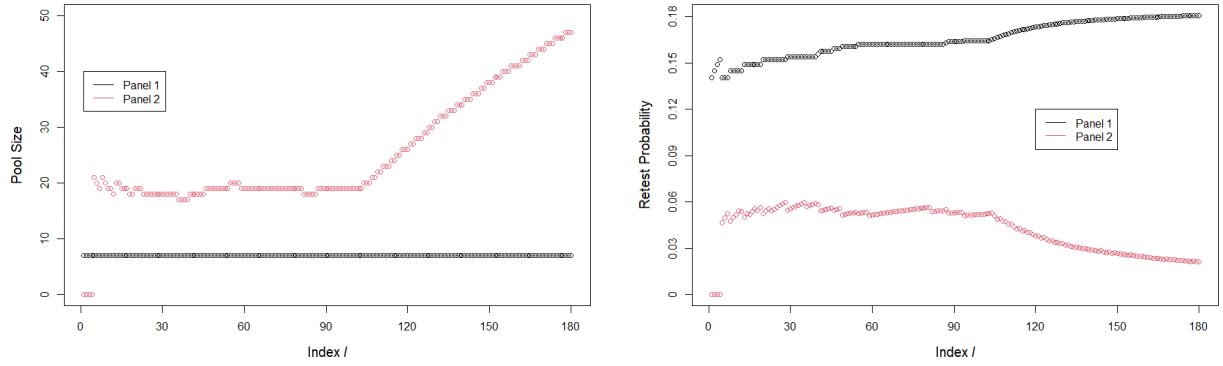
Figure 9: Optimal size, $z(\boldsymbol{x})$, and coverage, $y(\boldsymbol{x})$, of each panel of *P-DNA*($\overline{\eta} = 2$) for a family of optimal designs, as a function of index l



(a) Panel size vs design index

(b) Panel coverage vs design index

Figure 10: Optimal pool size, $t^*(\boldsymbol{x})$, and retest probability of each panel, $p(\boldsymbol{x})$, of *P-DNA*($\overline{\eta} = 2$) for a family of optimal designs, as a function of index l



(a) Panel pool size vs design index

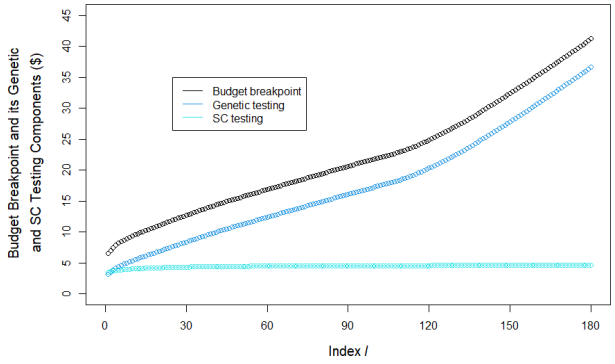
(b) Panel's retest probability vs design index

E Proofs

E.1 Supporting Derivations and Results

Without loss of generality, we assume that $q_i > 0$, $\forall i \in \Omega$ (by construction, set Ω includes those variants that have been detected in CF-positive subjects); and $P_N(n) > 0$, for $n = 0, 1, 2$.

Figure 11: Budget breakpoint, B^l, and its genetic and SC testing components for $P\text{-}DNA(\overline{\eta}=2)$ process, as a function of index l



We first provide some derivations and key properties that will be used subsequently in the proofs of the analytical results. We derive:

$$\frac{\partial p(\boldsymbol{x})}{\partial x_i} = q_i\left(P_N(1)+2P_N(2)\right) - 2q_i\left(\sum_{i\in\Omega} x_i q_i\right) P_N(2) = P_N(1)q_i + 2P_N(2)q_i\ \left(1-\sum_{i\in\Omega} x_i q_i\right) > 0,\ i\in\Omega.$$

$$\frac{\partial p(\boldsymbol{x})}{\partial y(\boldsymbol{x})} = P_N(1)+2P_N(2)-2y(\boldsymbol{x})P_N(2) > 0, \qquad \frac{\partial^2 p(\boldsymbol{x})}{\left(\partial y(\boldsymbol{x})\right)^2} = -2P_N(2) < 0.$$

$$\frac{\partial D(\boldsymbol{x},t)}{\partial p(\boldsymbol{x})} = t(1-p(\boldsymbol{x}))^{t-1} > 0, \qquad \frac{\partial^2 D(\boldsymbol{x},t)}{\left(\partial p(\boldsymbol{x})\right)^2} = -t(t-1)(1-p(\boldsymbol{x}))^{t-2} < 0,\ t\in Z^+, t\geq 2.$$

Lemma E.1. *Consider any panel composition $\boldsymbol{x}\neq \mathbf{1}_m$ and any pool size $t\in Z^+, t\geq 2$ (with panel indices $k,k'=1,\cdots,\overline{\eta}$, added as needed):*

1. *Each of $D(\boldsymbol{x},t)$, $D(\boldsymbol{x},\tilde{t}(\boldsymbol{x}))$, and $D(\boldsymbol{x},t^*(\boldsymbol{x}))$ is strictly concave increasing in $y(\boldsymbol{x})$.*
2. *(i) $C(\boldsymbol{x},t)$ is strictly increasing in x_i, for all $i\in\Omega$.*
   (ii) $TC((\boldsymbol{x}^k,t^k)_{k=1,\cdots,\overline{\eta}})$ is strictly increasing in $x_i^{k'}$, for all $i\in\Omega, k'=1,\cdots,\overline{\eta}$.
3. *Define the conditional panel coverage as $y(\boldsymbol{x};\gamma)\equiv y(\boldsymbol{x}:z(\boldsymbol{x})=\gamma)$, that is, with domain consisting of all $\boldsymbol{x}$ vectors with panel size γ, for an arbitrary $\gamma\in Z^+, \gamma\leq m-1$:*
   (i) Each of $C(\boldsymbol{x},t)$ and $C(\boldsymbol{x},t^*(\boldsymbol{x}))$ is strictly concave increasing in $y(\boldsymbol{x};\gamma)$.
   (ii) Each of $TC((\boldsymbol{x}^k,t^k)_{k=1,\cdots,\overline{\eta}})$ and $TC((\boldsymbol{x}^k,t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})$ is strictly concave increasing in $y(\boldsymbol{x}^{k'};\gamma^{k'})$, for all $k'=1,\cdots,\overline{\eta}$.
4. *$Pr(FN(\boldsymbol{x}))$ is strictly convex decreasing in $y(\boldsymbol{x})$.*

Proof. 1. Strict concavity of $D(\boldsymbol{x},t)$: We use the chain rule and the expressions derived at the beginning of Appendix E.1 to derive:

$$\frac{\partial D(\boldsymbol{x},t)}{\partial x_i} = \frac{\partial D(\boldsymbol{x},t)}{\partial p(\boldsymbol{x})}\frac{\partial p(\boldsymbol{x})}{\partial x_i} = t(1-p(\boldsymbol{x}))^{t-1}\left(q_i\left(P_N(1)+2P_N(2)\right) - 2q_i\left(\sum_{i\in\Omega} x_i q_i\right)P_N(2)\right) > 0,\ i\in\Omega$$

$$\frac{\partial D(\boldsymbol{x},t)}{\partial y(\boldsymbol{x})} = \frac{\partial D(\boldsymbol{x},t)}{\partial p(\boldsymbol{x})}\frac{\partial p(\boldsymbol{x})}{\partial y(\boldsymbol{x})} = t(1-p(\boldsymbol{x}))^{t-1}\left(P_N(1)+2P_N(2)-2y(\boldsymbol{x})P_N(2)\right) > 0,$$

$$\begin{aligned}\frac{\partial^2 D(\boldsymbol{x},t)}{\left(\partial y(\boldsymbol{x})\right)^2} &= \frac{\partial^2 D(\boldsymbol{x},t)}{\left(\partial p(\boldsymbol{x})\right)^2}\left(\frac{\partial p(\boldsymbol{x})}{\partial y(\boldsymbol{x})}\right)^2 + \frac{\partial D(\boldsymbol{x},t)}{\partial p(\boldsymbol{x})}\frac{\partial^2 p(\boldsymbol{x})}{\left(\partial y(\boldsymbol{x})\right)^2} \\ &= -t(t-1)(1-p(\boldsymbol{x}))^{t-2}\left(P_N(1)+2P_N(2)-2y(\boldsymbol{x})P_N(2)\right)^2 - 2t(1-p(\boldsymbol{x}))^{t-1}P_N(2) < 0,\end{aligned}$$

and the result follows.

Strict concavity of $D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))$: From [7], $\widetilde{t}(\boldsymbol{x})$ exists if and only if $p(\boldsymbol{x}) \leq 1-e^{-e^{-1}}$. Then, the following must hold:

$$\begin{aligned} &-0.5 < ln\left(1-\left(1-e^{-e^{-1}}\right)\right) \leq ln\left(1-p(\boldsymbol{x})\right) \leq 0 \\ \Leftrightarrow &-1 \leq -2ln(1-p(\boldsymbol{x}))\left(1+W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right)-1 < 0, \end{aligned}$$

where the last inequality follows because $-1 \leq W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right) \leq 0$.

Then, we must have:

$$\begin{aligned} \frac{\partial D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\partial p(\boldsymbol{x})} =& \frac{-1}{2(1-p(\boldsymbol{x}))W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)} > 0, \text{ and} \\ \frac{\partial^2 D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\left(\partial p(\boldsymbol{x})\right)^2} =& \frac{-2W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right) + \frac{-W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)(1-p(\boldsymbol{x}))}{ln(1-p(\boldsymbol{x}))(1-p(\boldsymbol{x}))\left(1+W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right)}}{4(1-p(\boldsymbol{x}))^2\left(W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right)^2} \\ =& \frac{-2ln(1-p(\boldsymbol{x}))\left(1+W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right)-1}{4(1-p(\boldsymbol{x}))^2 W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right) ln(1-p(\boldsymbol{x}))\left(1+W_0\left(-\frac{1}{2}\sqrt{\left(ln\frac{1}{1-p(\boldsymbol{x})}\right)}\right)\right)} < 0. \end{aligned}$$

Therefore, by the chain rule, and from the expressions at the beginning of Appendix E.1:

$$\begin{aligned} \frac{\partial D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\partial y(\boldsymbol{x})} =& \frac{\partial D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\partial p(\boldsymbol{x})}\frac{\partial p(\boldsymbol{x})}{\partial y(\boldsymbol{x})} > 0, \quad \text{and} \\ \frac{\partial^2 D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\left(\partial y(\boldsymbol{x})\right)^2} =& \frac{\partial^2 D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\left(\partial p(\boldsymbol{x})\right)^2}\left(\frac{\partial p(\boldsymbol{x})}{\partial y(\boldsymbol{x})}\right)^2 + \frac{\partial D(\boldsymbol{x}, \tilde{t}(\boldsymbol{x}))}{\partial p(\boldsymbol{x})}\frac{\partial^2 p(\boldsymbol{x})}{\left(\partial y(\boldsymbol{x})\right)^2} < 0. \end{aligned}$$

Strict concavity of $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$: By Remark 1, $t^*(\boldsymbol{x})$ satisfies, $D(\boldsymbol{x}, t^*(\boldsymbol{x})) = \min_{t\in\{2,\cdots,\bar{t}\}}\{D(\boldsymbol{x}, t)\}$. Also, from part 1(i) of this lemma, we have that $D(\boldsymbol{x}, t)$ is strictly concave increasing in $y(\boldsymbol{x})$, $t = 2, \cdots, \bar{t}$. Then, $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is the point-wise minimum of strictly concave increasing functions, and hence is also strictly concave increasing in $y(\boldsymbol{x})$ [13].

2-3. To prove the results for $C(\boldsymbol{x}, t) = D(\boldsymbol{x}, t) \times (c_f + c_v(z(\boldsymbol{x}))) = \left(\frac{1}{t} + 1 - (1-p(\boldsymbol{x}))^t\right) \times (c_f + c_v(z(\boldsymbol{x})))$, we derive:

$$\begin{aligned} \frac{\partial C(\boldsymbol{x}, t)}{\partial x_i} =& \frac{\partial D(\boldsymbol{x}, t)}{\partial x_i}(c_f + c_v(z(\boldsymbol{x}))) + (D(\boldsymbol{x}, t))\frac{\partial c_v(z(\boldsymbol{x}))}{\partial x_i} > 0, \ i \in \Omega, \\ \frac{\partial C(\boldsymbol{x}, t)}{\partial y(\boldsymbol{x}; \gamma)} =& \frac{\partial D(\boldsymbol{x}, t)}{\partial y(\boldsymbol{x}; \gamma)}(c_f + c_v(\gamma)) > 0, \\ \frac{\partial C(\boldsymbol{x}, t^*(\boldsymbol{x}))}{\partial y(\boldsymbol{x}; \gamma)} =& \frac{\partial D(\boldsymbol{x}, t^*(\boldsymbol{x}))}{\partial y(\boldsymbol{x}; \gamma)}(c_f + c_v(\gamma)) > 0, \\ \frac{\partial^2 C(\boldsymbol{x}, t)}{(\partial y(\boldsymbol{x}; \gamma))^2} =& \frac{\partial^2 D(\boldsymbol{x}, t)}{(\partial y(\boldsymbol{x}; \gamma))^2}(c_f + c_v(\gamma)) < 0, \quad \text{and} \\ \frac{\partial^2 C(\boldsymbol{x}, t^*(\boldsymbol{x}))}{(\partial y(\boldsymbol{x}; \gamma))^2} =& \frac{\partial^2 D(\boldsymbol{x}, t^*(\boldsymbol{x}))}{(\partial y(\boldsymbol{x}; \gamma))^2}(c_f + c_v(\gamma)) < 0, \end{aligned}$$

and the results follow.

To prove the results for $TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})$, we derive, for $k' = 1, \cdots, \overline{\eta}$:

$$\begin{aligned}
\frac{\partial TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})}{\partial x_i^{k'}} =& \frac{\partial C(\boldsymbol{x}^k, t^k)}{\partial x_i^{k'}} + \frac{\partial p(\boldsymbol{x}^{12\cdots\overline{\eta}})}{\partial x_i^{k'}} c_{SC} > 0,\ i \in \Omega, \\
\frac{\partial TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} =& \frac{\partial C(\boldsymbol{x}^k, t^k)}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} + \frac{\partial p(\boldsymbol{x}^{12\cdots\overline{\eta}})}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} c_{SC} > 0, \\
\frac{\partial TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} =& \frac{\partial C(\boldsymbol{x}^{k'}, t^*(\boldsymbol{x}^{k'}))}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} + \frac{\partial p(\boldsymbol{x}^{12\cdots\overline{\eta}})}{\partial y(\boldsymbol{x}^{k'};\gamma^{k'})} c_{SC} > 0, \\
\frac{\partial^2 TC((\boldsymbol{x}^k, t^k)_{k=1,\cdots,\overline{\eta}})}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} =& \frac{\partial^2 C(\boldsymbol{x}^{k'}, t^{k'})}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} + \frac{\partial^2 p(\boldsymbol{x}^{12\cdots\overline{\eta}})}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} c_{SC} < 0, \quad \text{and} \\
\frac{\partial^2 TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} =& \frac{\partial^2 C(\boldsymbol{x}^{k'}, t^*(\boldsymbol{x}^{k'}))}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} + \frac{\partial^2 p(\boldsymbol{x}^{12\cdots\overline{\eta}})}{(\partial y(\boldsymbol{x}^{k'};\gamma^{k'}))^2} c_{SC} < 0,
\end{aligned}$$

and the results follow.

4. $Pr\left(FN(\boldsymbol{x})\right) = \left(1 - \sum_{i\in\Omega} x_i q_i\right)^2 P_N\left(2\right) = \left(1 - y(\boldsymbol{x})\right)^2 P_N\left(2\right)$. Then, for any $\boldsymbol{x} \neq \mathbf{1}_m$, we have that $y(\boldsymbol{x}) < 1$, and hence, $\frac{\partial Pr(FN(\boldsymbol{x}))}{\partial y(\boldsymbol{x})} = -2\left(1 - y(\boldsymbol{x})\right) P_N\left(2\right) < 0$, and $\frac{\partial^2 Pr(FN(\boldsymbol{x}))}{(\partial y(\boldsymbol{x}))^2} = 2P_N\left(2\right) > 0$, completing the proof. □

E.2 Proofs of the Analytical Results

Proof of Lemma 1. By Lemma E.1, $C(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is strictly concave increasing in conditional panel coverage $y(\boldsymbol{x};\gamma)$, i.e., considering all $\boldsymbol{x} : z(\boldsymbol{x}) = \gamma$. Hence, we have that for any set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, there exists an ordered partition that minimizes $\sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k > \mathbf{0}\}} \times C(\boldsymbol{x}^k, t^k)$ [19]. Further, since $c_{SC} \times p(\boldsymbol{x}^{12\cdots\overline{\eta}})$ is a constant for any set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, i.e., independent of the set partition, by Remark 1 it follows that there exists an ordered partition of set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ that minimizes $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}}) = c_{SC} \times p(\boldsymbol{x}^{12\cdots\overline{\eta}}) + \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k > \mathbf{0}\}} \times C(\boldsymbol{x}^k, t^k)$. This completes the proof. □

Proof of Corollary 1. By Lemma 1, $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,\cdots,\overline{\eta}})$ is minimized by an ordered partition of some set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$. Then, the problem of finding the ordered partition that minimizes the $TC(.)$ function for a given variant set reduces to the shortest path problem described in the corollary [19, 21], and the complexity result follows from [11, 19, 21, 29, 41]. □

Proof of Lemma 2. Consider the variant partition problem with $\overline{\eta} = 2$ for some variant set $S(\boldsymbol{x}^{12})$ that has l variants, $l = 2, \cdots, \min\{2\overline{z}, m\}$ (ordered following a non-increasing order of variant frequency). By Lemma 1, the $TC(.)$ function for the given variant set is minimized by either an ordered 2-partition or the 1-partition. Then, among all 2-partitions with panel sizes $z(\boldsymbol{x}^1) = \gamma$ and $z(\boldsymbol{x}^2) = l - \gamma$, $\gamma \in Z^+ : \left\lceil \frac{l}{2} \right\rceil \leq \gamma \leq \min\{l, \overline{z}\}$, the $TC(.)$ function is minimized by one of the following ordered 2-partitions: (1) $S(\boldsymbol{x}^1) = \{1, \cdots, \gamma\}$ and $S(\boldsymbol{x}^2) = \{\gamma+1, \cdots, l\}$; or (2) $S(\boldsymbol{x}^1) = \{l-\gamma+1, \cdots, l\}$ and $S(\boldsymbol{x}^2) = \{1, \cdots, l-\gamma\}$. Then, the result follows because the **2-Partition Algorithm**, given below, considers all feasible ordered 2-partitions to determine an optimal 2-partition, and compares it with the 1-partition. Observe that considering only those panel sizes, $z(\boldsymbol{x}^k), k = 1, 2 : z(\boldsymbol{x}^1) \geq z(\boldsymbol{x}^2)$ (implied by the range on γ above) is symmetry-breaking, and is without loss of optimality.

2-Partition Algorithm (for variant set $S(\boldsymbol{x}^{12}) = \{1, \cdots, l\}$)

Initialization: $\boldsymbol{x}^{k*} = \mathbf{0}$, $k = 1, 2$, $TC = \infty$
for $\gamma \in [1, \min\{l, \overline{z}\}]$ **do** (*Generation of all optimal* 1- *and* 2-*partitions*)
  Set $S(\boldsymbol{x}^1) = \{1, \cdots, \gamma\}$, i.e., $x_i^1 = 1$, for $i = 1, \cdots, \gamma$, and $x_i^1 = 0$ otherwise.
  Set $S(\boldsymbol{x}^2) = \{\gamma + 1, \cdots, l\}$, i.e., $x_i^2 = 1$, for $i = \gamma + 1, \cdots, l$, and $x_i^2 = 0$ otherwise.
  Determine optimal pool sizes, $t^*(\boldsymbol{x}^k)$, $k = 1, 2$ (via Property C.1)
  Compute the expected total cost, $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,2})$
  if $TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,2}) < TC$ **then**
    $TC = TC((\boldsymbol{x}^k, t^*(\boldsymbol{x}^k))_{k=1,2})$, $\boldsymbol{x}^{k*} = \boldsymbol{x}^k, k = 1, 2$
  end if
end for
Output: Variant partition, pool sizes, expected testing cost: $\boldsymbol{x}^{k*}$, $t^*(\boldsymbol{x}^{k*}), k = 1, 2$, TC

Regarding the computational complexity of the algorithm, for any variant set $S(\boldsymbol{x}^{12})$ having l variants, the algorithm performs at most one operation for each $\gamma = \{1, \cdots, l\}$ (some panel sizes may not be feasible due to the panel size limit, $\overline{z}$, hence requiring no operation); and the algorithm considers at most l different values of γ, leading to a complexity of $O(l)$. □

Proof of Theorem 1. In the following, we use the term "expected cost" of a given variant set to refer to its expected total cost at the optimal variant partition and pool size vector. Without loss of generality, all variant sets are arranged following a non-increasing order of variant frequency. Also recall that budget breakpoint $B^l, l = 1 \cdots, m$, corresponds to the expected cost for ordered variant set $S(l)$ (Definition 3). To simplify the subsequent notation, we refer to an arbitrary variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ as set S, with $\eta(S)$ non-empty panels in the optimal variant partition.

Consider any budget $B < B^l$ for some $l = 1, \cdots, m$. The proof is two-fold. Considering all ordered and non-ordered variant sets, we first show that there does not exist any feasible variant set $S \subset \Omega$ with size $z(S) < l$ such that $Pr(FN(S)) \leq Pr(FN(S(l)))$; then we show that there also does not exist any feasible variant set $S \subseteq \Omega$ with size $z(S) \geq l$ such that $Pr(FN(S)) < Pr(FN(S(l)))$.

First consider any variant set $S \subset \Omega : z(S) < l$, i.e., any ordered or non-ordered set with at most $l - 1$ variants, for which there is a feasible solution at budget B. Because $S(l)$ is an ordered variant set, we must have that $y(S) < y(S(l))$ (Definition 2); and because $Pr(FN(S))$ is strictly decreasing in $y(S)$ (Lemma E.1), it follows that $Pr(FN(S)) > Pr(FN(S(l)))$.

We prove the second part by contradiction. Consider any variant set $S \subseteq \Omega : z(S) \geq l$, for which there is a feasible solution at budget B with $Pr(FN(S)) = Pr(FN(S(l)))$. In the following, we show that the solution corresponding to set S must have an expected cost $TC(S) \geq B^l$, hence set S cannot be feasible at budget $B < B^l$. We do this by studying the optimal ordered variant partitions (Lemma 1) for variant sets S and $S(l)$, respectively denoted by the collection of subsets, $\tilde{S}^k, k = 1, \cdots, \eta(S)$, and $S^{k*}(l), k = 1, \cdots, \eta(S(l))$, the total cost of which are respectively denoted by $TC(S)$ and $TC(S(l))$; and comparing them with each other, and with a set of "dummy" solutions that are obtained by perturbing the variant frequency vector.

To this end, denote the variants in set S as $\{1, \cdots, z(S)\}$, with optimal ordered partition, $\tilde{S}^1 = \{1, \cdots, z(\tilde{S}^1)\}$, $\tilde{S}^k = \{1 + \sum_{r=1}^{k-1} z(\tilde{S}^r), \cdots, \sum_{r=1}^{k} z(\tilde{S}^r)\}$, $k = 2, \cdots, \eta(S)$, and $\tilde{S}^k = \emptyset$, $k = \eta(S) + 1, \cdots, \overline{\eta}$, i.e., the subsets are arranged such that subset 1 has the highest frequency variants, subset 2 has the second most highest frequencies, and so on. Because $z(S) \geq l$ by definition, there must exist some integer $k' \leq \eta(S) : \sum_{k=1}^{k'-1} z(\tilde{S}^k) < l \leq \sum_{k=1}^{k'} z(\tilde{S}^k)$. Next consider ordered variant set $S(l)$, for which we construct a set of dummy partitions in a sequential manner, and use the collection of subsets, $S^k(l), k = 1, \cdots, \eta(S)$, to refer to the "current" (most recent) dummy partition. We start with the first dummy partition, $S^k(l), k = 1, \cdots, \eta(S) : y(S^1(l)) = \sum_{i=1}^{z(\tilde{S}^1)} q_i$, $y(S^k(l)) = \sum_{i=z(\cup_{r=1}^{k-1}\tilde{S}^r)+1}^{z(\cup_{r=1}^{k}\tilde{S}^r)} q_i$, for $k = 2, \cdots, k' - 1$, $y(S^{k'}(l)) = \sum_{i=1}^{l} q_i - \sum_{r=1}^{k'-1} y(S^r(l))$, and $y(S^k(l)) = 0$, $k = k' + 1, \cdots, \eta(S)$. Observe that by construction, $\sum_{k=1}^{\eta(S(l))} y(S^{k*}(l)) = \sum_{k=1}^{\eta(S)} y(S^k(l)) = \sum_{k=1}^{\eta(S)} y(\tilde{S}^k)$, hence the FN probabilities of all three solutions are equal (Lemma E.1).

In the remainder of the proof, we compare the expected cost of the optimal partition of set S, i.e., $TC(S)$, with that of each dummy partition of set $S(l)$. We do this by creating the set of dummy partitions in such a way that they all incur the same FN probability as $Pr(FN(S))$ (which equals $Pr(FN(S(l)))$ by construction of set S), and showing that each dummy partition results in a lower expected cost than $TC(S)$. Specifically, each new dummy partition is obtained by modifying the coverage of the two specific subsets (panels) in the current dummy partition by some constant such that the coverage of one subset increases, while the coverage of the other subset decreases by this constant: These two subsets respectively correspond to the subset with the lowest index $k : y(S^k(l)) < y(S^{k*}(l))$, and the subset with the highest index among the non-empty subsets, based on the current dummy partition (starting with the optimal partition of set S). Then we show that the expected cost of each new dummy partition is lower than that of the previous dummy partition. Finally, we show that the optimal partition of set $S(l)$ satisfies, $TC((S^{k*}(l))_{k=1,\cdots,\overline{\eta}}) \leq TC((S^k(l))_{k=1,\cdots,\overline{\eta}})$, that is, it cost-dominates all the dummy partitions.

To show that $TC((S^k(l))_{k=1,\cdots,\overline{\eta}}) \leq TC(S)$, recall that by construction, $y(S^k(l)) \geq y(\tilde{S}^k)$, $k = 1, \cdots, k'-1$, and $y(S^k(l)) < y(\tilde{S}^k)$, $k = k'+1, \cdots, \eta(S)$. Let $\epsilon_k \equiv |y(S^k(l)) - y(\tilde{S}^k)|$, $k = 1, \cdots, \eta(S)$. $\tilde{S}^1 \cup \tilde{S}^{\eta(S)}$ denotes the set of all variants in panels 1 and $\eta(S)$ of set S. By Lemma 1, the optimal partition of set $S(\boldsymbol{x}^1(S)+\boldsymbol{x}^{\eta(S)}(S))$ must be ordered. Let $(\tilde{S}^1, \tilde{S}^{\eta(S)})$ and $(\tilde{S}'^1, \tilde{S}'^{\eta(S)})$ denote two possible ordered partitions of set $\tilde{S}^1 \cup \tilde{S}^{\eta(S)}$. Since $(\tilde{S}^1, \tilde{S}^{\eta(S)})$ is the optimal partition for set S, the following must be true:

$$D(\tilde{S}^1)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) + D(\tilde{S}^{\eta(S)})\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right) \leq D(\tilde{S}'^1)\left(c_f + c_v\left(z(\tilde{S}'^1)\right)\right) + D(\tilde{S}'^{\eta(S)})\left(c_f + c_v\left(z(\tilde{S}'^{\eta(S)})\right)\right).$$

In the following, we write $D\left(y(S)\right)$, to show its dependence on $y(S)$ (Property C.1, Eqs. (1)-(2)). By Lemma E.1, we also have that for any set S, $D\left(S\right)$ is concave increasing in $y(S)$. By definition of $\tilde{S}^1, \tilde{S}'^1, \tilde{S}^{\eta(S)}$, and $\tilde{S}'^{\eta(S)}$, we have that $y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)}) = y(\tilde{S}^1) - y(\tilde{S}'^1) \geq 0$. Then, due to the strict concavity of $D(y(S))$ in $y(S)$, the following result must hold for any positive $\epsilon \leq \min\{\epsilon_1, \epsilon_{\eta(S)}\}$, because $y(\tilde{S}^1) - y(\tilde{S}'^1) = y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)})$:

$$\begin{aligned}
&\frac{D\left(y(\tilde{S}^1)+\epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) - D\left(y(\tilde{S}^1)\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right)}{\epsilon} \\
&< \frac{D\left(y(\tilde{S}^1)\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) - D\left(y(\tilde{S}'^1)\right)\left(c_f + c_v\left(z(\tilde{S}'^1)\right)\right)}{y(\tilde{S}^1) - y(\tilde{S}'^1)} \\
&\leq \frac{D\left(y(\tilde{S}'^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}'^{\eta(S)})\right)\right) - D\left(y(\tilde{S}^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right)}{y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)})} \\
&< \frac{D\left(y(\tilde{S}^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right) - D\left(y(\tilde{S}^{\eta(S)})-\epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right)}{\epsilon}.
\end{aligned}$$

Then, by setting $\epsilon = \min\{\epsilon_1, \epsilon_{\eta(S)}\}$, we reach a new dummy partition with the same or lower expected cost than the previous study solution. There are two possible cases:

Case 1: $\epsilon = \epsilon_1$

Then, in the new dummy partition, $y'(\tilde{S}^1) = y(S^1(l))$, $y'(\tilde{S}^{\eta(S)}) = y(\tilde{S}^{\eta(S)}) - \epsilon_1$, and $y'(\tilde{S}^k) = y(\tilde{S}^k)$, $k = 2, \cdots, \eta(S)-1$; and the following must hold for any $0 < \epsilon \leq \min\{\epsilon_2, \epsilon_{\eta(S)} - \epsilon_1\}$, because $y(\tilde{S}^2) - y(\tilde{S}'^2) = y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)})$:

$$\begin{aligned}
&\frac{D\left(y(\tilde{S}^2)+\epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^2)\right)\right) - D\left(y(\tilde{S}^2)\right)\left(c_f + c_v\left(z(\tilde{S}^2)\right)\right)}{\epsilon} \\
&< \frac{D\left(y(\tilde{S}^2)\right)\left(c_f + c_v\left(z(\tilde{S}^2)\right)\right) - D\left(y(\tilde{S}'^2)\right)\left(c_f + c_v\left(z(\tilde{S}'^2)\right)\right)}{y(\tilde{S}^2) - y(\tilde{S}'^2)}
\end{aligned}$$

$$\leq \frac{D\left(y(\tilde{S}'^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}'^{\eta(S)})\right)\right) - D\left(y(\tilde{S}^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right)}{y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)})}$$
$$< \frac{D\left(y(\tilde{S}^{\eta(S)})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right) - D\left(y(S^{\eta(S)}) - \epsilon_1\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right)}{\epsilon_1}$$
$$< \frac{D\left(y(\tilde{S}^{\eta(S)}) - \epsilon_1\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right) - D\left(y(\tilde{S}^{\eta(S)}) - \epsilon_1 - \epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)})\right)\right)}{\epsilon}.$$

Hence, by setting $\epsilon = \min\{\epsilon_2, \epsilon_{\eta(S)} - \epsilon_1\}$, we have obtained another dummy partition with the same or lower expected cost.

Case 2: $\epsilon = \epsilon_{\eta(S)}$

Then, in the new dummy partition, $y'(\tilde{S}^1) = y(\tilde{S}^1) + \epsilon_{\eta(S)}$, $y'(\tilde{S}^{\eta(S)}) = y(\tilde{S}^{\eta(S)}(l)) = 0$, and $y'(\tilde{S}^k) = y(\tilde{S}^k)$, $k = 2, \cdots, \eta(S) - 1$. Similarly, the following must hold for any $0 < \epsilon \leq \min\{\epsilon_1 - \epsilon_{\eta(S)}, \epsilon_{\eta(S)-1}\}$, because $y(\tilde{S}^1) - y(\tilde{S}'^1) = y(\tilde{S}'^{\eta(S)}) - y(\tilde{S}^{\eta(S)})$:

$$\frac{D\left(y(\tilde{S}^1) + \epsilon_{\eta(S)} + \epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) - D\left(y(\tilde{S}^1) + \epsilon_{\eta(S)}\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right)}{\epsilon}$$
$$< \frac{D\left(y(\tilde{S}^1) + \epsilon_{\eta(S)}\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) - D\left(y(\tilde{S}^1)\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right)}{\epsilon_{\eta(S)}}$$
$$< \frac{D\left(y(\tilde{S}^1)\right)\left(c_f + c_v\left(z(\tilde{S}^1)\right)\right) - D\left(y(\tilde{S}'^1)\right)\left(c_f + c_v\left(z(\tilde{S}'^1)\right)\right)}{y(\tilde{S}^1) - y(\tilde{S}'^1)}$$
$$\leq \frac{D\left(y(\tilde{S}'^{\eta(S)-1})\right)\left(c_f + c_v\left(z(\tilde{S}'^{\eta(S)-1})\right)\right) - D\left(y(\tilde{S}^{\eta(S)-1})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)-1})\right)\right)}{y(\tilde{S}'^{\eta(S)-1}) - y(\tilde{S}^{\eta(S)-1})}$$
$$< \frac{D\left(y(\tilde{S}^{\eta(S)-1})\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)-1})\right)\right) - D\left(y(\tilde{S}^{\eta(S)-1}) - \epsilon\right)\left(c_f + c_v\left(z(\tilde{S}^{\eta(S)-1})\right)\right)}{\epsilon}.$$

Thus, by repeating the same argument for each $\tilde{S}^k$, $k = 1, \cdots, \eta(S)$, in both cases we obtain the dummy partition $(S^k(l))_{k=1,\cdots,\overline{\eta}}$, which leads to a lower or equal expected cost.

Finally, we show that $(S^k(l))_{k=1,\cdots,\overline{\eta}}$ has a higher expected cost than $(S^{k*}(l))_{k=1,\cdots,\overline{\eta}}$. Observe that panels $k' + 1, \cdots, \eta(S)$ have zero coverage, hence zero cost. Then:

$$\sum_{k=1}^{\eta(S)} D\left(S^k(l)\right)\left(c_f + c_v\left(z(\tilde{S}^k)\right)\right) \geq \sum_{k=1}^{k'} D\left(S^k(l)\right)\left(c_f + c_v\left(z(\tilde{S}^k)\right)\right).$$

We also have that panel k' has $z(S^{k'}(l)) \leq z(\tilde{S}^{k'})$ variants. Then, we have that:

$$\sum_{k=1}^{k'} D\left(S^k(l)\right)\left(c_f + c_v\left(z(\tilde{S}^k)\right)\right) \geq \sum_{k=1}^{k'-1} D\left(S^k(l)\right)\left(c_f + c_v\left(z(\tilde{S}^k)\right)\right) + D\left(S^{k'}(l)\right)\left(c_f + c_v\left(z(S^{k'}(l))\right)\right)$$
$$= \sum_{k=1}^{k'} D\left(S^k(l)\right)\left(c_f + c_v\left(z(S^k(l))\right)\right)$$
$$\geq \sum_{k=1}^{\eta(l)} D\left(S^{k*}(l)\right)\left(c_f + c_v\left(z(S^{k*}(l))\right)\right),$$

where the last inequality follows by optimality of partition $(S^{k*}(l)_{k=1,\cdots,\overline{\eta}})$ for set $S(l)$. Hence, we must have

that $B^l = TC(S(l)) \leq TC(S)$, completing the proof. □

Proof of Corollary 2.

1-2 The results follow directly from Theorem 1. For part 2, observe that if $Pr(FN(S)) < Pr(FN(S(l))$ for some variant set $S \subseteq \Omega$, then we must have that $z(S) > l$ (Lemma E.1) because set Ω has variants with distinct frequencies, as assumed in this part.

3. The result follows from Theorem 1, because the optimal variant set for B^m remains feasible for all budgets $B \geq B^m$. Then, because $Pr(FN(\boldsymbol{x}))$ is strictly decreasing in $y(\boldsymbol{x})$ (Lemma E.1), the optimal variant set must be $S(m) = \{1, \cdots, m\} = \Omega$, i.e., the optimal variant set for B^m. □

Proof of Corollary 3. Solving the shortest path problem for set $S(m)$, i.e., from vertex 1 to vertex $m+1$, generates the shortest path from vertex 1 to every other vertex $l = 2, \cdots, m+1$ [10, 21, 29]. Then, to generate the entire family of optimal breakpoint designs and budget breakpoints, it is sufficient to solve the shortest path problem only once, for variant set $S(m) = \Omega$. The computational complexity then follows from Corollary 1. □

Proof of Theorem 2. Let $\boldsymbol{x}^{12\cdots\overline{\eta}*}$ denote the optimal variant set at budget B, where $B^l < B < B^{l+1}$ for some $l = 1, \cdots, m-1$. Hence, by Definitions 2-3, variant set $S(l)$, equivalently variant vector $\boldsymbol{x}^{12\cdots\overline{\eta}*}(l) = \mathbf{1}_l$, incurs an expected total cost (at the optimal partition and pool size vector) of B^l, and is feasible at budget B, whereas variant set $S(l+1)$, equivalently variant vector $\boldsymbol{x}^{12\cdots\overline{\eta}*}(l+1) = \mathbf{1}_{l+1}$, with an expected total cost of B^{l+1}, is not. Then, by Theorem 1, variant sets $S(l)$ and $S(l+1)$ respectively provide a lower bound and an upper bound on the optimal objective function value at budget B, that is, $Pr(FN(\mathbf{1}_{l+1})) < Pr(FN(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \leq Pr(FN(\mathbf{1}_l))$, and the result follows. □

Proof of Lemma 3. Consider any $B^l < B \leq B^{l+1} < B' < B^{l+2}$, $l = 1, \cdots, m-2$. By Definition 2, for ordered variant set $S(l)$, we have that $\boldsymbol{x}^{12\cdots\overline{\eta}*}(l) = \mathbf{1}_l, \forall l \in Z^+$. We can write:

$$\begin{aligned}\frac{\varepsilon_{UB}(B)}{q_{l+1}} =& \frac{Pr(FN(\mathbf{1}_l)) - Pr(FN(\mathbf{1}_{l+1}))}{q_{l+1}} \\ >& \frac{Pr(FN(\mathbf{1}_{l+1})) - Pr(FN(\mathbf{1}_{l+2}))}{q_{l+2}} \\ \geq& \frac{Pr(FN(\mathbf{1}_{l+1})) - Pr(FN(\mathbf{1}_{l+2}))}{q_{l+1}} = \frac{\varepsilon_{UB}(B')}{q_{l+1}},\end{aligned}$$

where the first inequality follows by the strict convexity of $Pr(FN(\boldsymbol{x}))$ in $y(\boldsymbol{x})$ (Lemma E.1) and the second inequality follows because $q_{l+1} \geq q_{l+2}$ by construction of set Ω.
Similarly, for $B^l < B < B' < B^{l+1}$, we have that:

$$\frac{\varepsilon_{UB}(B)}{q_{l+1}} = \frac{Pr(FN(\mathbf{1}_l)) - Pr(FN(\mathbf{1}_{l+1}))}{q_{l+1}} = \frac{\varepsilon_{UB}(B')}{q_{l+1}}.$$

In either case $\varepsilon_{UB}(B) \geq \varepsilon_{UB}(B')$, and the result follows. □

Proof of Theorem 3. Given any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$ with size $z(\boldsymbol{x}^{12\cdots\overline{\eta}}) \geq 2$, consider the optimal η-panel design for some $\eta = 2, \cdots, \min\{\overline{\eta}, z(\boldsymbol{x}^{12\cdots\overline{\eta}})\}$ (i.e., with η non-empty panels) with partition $(\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}$: $\sum_{k=1}^{\eta} \boldsymbol{x}^{k*} = \boldsymbol{x}^{12\cdots\overline{\eta}}$, and pool sizes $(t^*(\boldsymbol{x}^{k*}))_{k=1,\cdots,\overline{\eta}}$; and the optimal 1-panel design with $\boldsymbol{x}^{12\cdots\overline{\eta}*}$ and pool size $t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})$. In the following, we refer to these optimal designs as the η-panel and 1-panel, respectively.

First note that $D(\boldsymbol{x}, t^*(\boldsymbol{x})) \geq 0$, and $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is strictly concave increasing in $y(\boldsymbol{x})$ (Lemma E.1), and hence $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is sub-additive in $y(\boldsymbol{x})$, which implies:

$$D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right) \leq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right). \tag{18}$$

1. We prove this result in two steps. First we show that for any variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}})$, if there exists a fixed cost $c_f \geq 0$ at which the η-panel cost-dominates the 1-panel, then the η-panel will continue to

cost-dominate for all lower values of c_f; and similarly, we show that if there exists a $c_f \geq 0$ at which the 1-panel cost-dominates the η-panel, then the 1-panel will continue to cost-dominate for all higher values of c_f.

First assume $\exists c_f \geq 0$ at which the η-panel cost-dominates the 1-panel, that is:

$$\begin{aligned} & TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \geq TC((\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*}))_{k=1,\cdots,\overline{\eta}}) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right). \end{aligned} \tag{19}$$

Now assume that c_f decreases by ϵ, for any $0 \leq \epsilon \leq c_f$, and let $(\boldsymbol{x}'^{k*})_{k=1,\cdots,\overline{\eta}}$ denote the new optimal η-panel partition at $c_f - \epsilon$. Because $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is sub-additive in $y(\boldsymbol{x})$, Eqs. (18) and (19) lead to the following:

$$\begin{aligned} & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) - D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\epsilon \\ & \quad \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right) - \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\epsilon \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f - \epsilon + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f - \epsilon + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right) \\ & \quad \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}'^{k*}, t^*(\boldsymbol{x}'^{k*})\right)\left(c_f - \epsilon + c_v\left(z\left(\boldsymbol{x}'^{k*}\right)\right)\right), \end{aligned} \tag{20}$$

where the last inequality follows by optimality of $(\boldsymbol{x}'^{k*})_{k=1,\cdots,\overline{\eta}}$ for the η-panel design when the fixed cost is $c_f - \epsilon$.

Next assume $\exists c_f \geq 0$ at which the 1-panel cost-dominates the η-panel, that is:

$$\begin{aligned} & TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \leq TC((\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*}))_{k=1,\cdots,\overline{\eta}}) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \leq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right). \end{aligned} \tag{21}$$

We prove this result by contradiction. Assume c_f increases by ϵ, and assume that the new optimal η-panel design, denoted by $(\boldsymbol{x}''^{k*})_{k=1,\cdots,\overline{\eta}}$, incurs a lower expected cost than the 1-panel design, that is:

$$\begin{aligned} & TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) > TC((\boldsymbol{x}''^{k*}, t^*(\boldsymbol{x}''^{k*}))_{k=1,\cdots,\overline{\eta}}) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right)\right)\left(c_f + \epsilon + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) > \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}''^{k*}, t^*\left(\boldsymbol{x}''^{k*}\right)\right)\left(c_f + \epsilon + c_v\left(z\left(\boldsymbol{x}''^{k*}\right)\right)\right). \end{aligned}$$

Then, by Eq. (20), if $c_f + \epsilon$ decreases by ϵ, then the new optimal η-panel design will continue to cost-dominate the 1-panel design, that is, we must have that:

$$D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) > \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right),$$

reaching a contradiction with Eq. (21). Hence it must be true that:

$$TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \leq TC((\boldsymbol{x}''^{k*}, t^*(\boldsymbol{x}''^{k*}))_{k=1,\cdots,\overline{\eta}}),$$

and the result follows.

2. We prove this result in two steps. First we show that if there exists a convex non-decreasing variable cost function $c'_v(.) \geq 0$ for which the η-panel cost-dominates the 1-panel, then the η-panel will continue to cost-dominate for all cost functions $c_v(z(\boldsymbol{x})) = \epsilon \times c'_v(z(\boldsymbol{x}))$, with $\epsilon \geq 1$; and similarly, we show that if there exists a variable cost function $c'_v(.) \geq 0$ for which the 1-panel cost-dominates the η-panel, then the 1-panel will continue to cost-dominate for all cost functions $c_v(z(\boldsymbol{x})) = \epsilon \times c'_v(z(\boldsymbol{x}))$, with $0 < \epsilon \leq 1$, thus establishing the existence of a threshold, $\overline{\epsilon}(S(\boldsymbol{x}^{12\cdots\overline{\eta}}), \eta)$.

   Recall that $(\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}$ denotes the optimal η-panel partition for the original variable cost function $c'_v(.)$. Let $(\boldsymbol{x}'^{k*})_{k=1,\cdots,\overline{\eta}}$, and $(\boldsymbol{x}''^{k*})_{k=1,\cdots,\overline{\eta}}$ denote the optimal η-panel partition for variable cost function $\epsilon \times c'_v(z(\boldsymbol{x}))$, when $\epsilon \geq 1$ and $0 < \epsilon \leq 1$, respectively.

   Now assume that the η-panel cost-dominates the 1-panel at variable cost function $c'_v(.)$, that is:

$$\begin{aligned} & TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \geq TC((\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*}))_{k=1,\cdots,\overline{\eta}}) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right). \end{aligned} \tag{22}$$

   Assume that $c'_v(.)$ is replaced by $c_v(.) = \epsilon \times c'_v(.)$, for some $\epsilon \geq 1$. Because of the sub-additivity of $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ in $y(\boldsymbol{x})$, Eqs. (18) and (22) lead to the following:

$$\begin{aligned} & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)(1-\epsilon)c_f \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)(1-\epsilon)c_f, \text{ and} \\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\epsilon\left(c_f + c'_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\epsilon\left(c_f + c'_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right). \end{aligned} \tag{23}$$

   Therefore, by Eqs. (18) and (23), we have that:

$$\begin{aligned} & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\epsilon\left(c_f + c'_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) + D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)(1-\epsilon)c_f \\ & \quad \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\epsilon\left(c_f + c'_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right) + \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)(1-\epsilon)c_f \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + \epsilon c'_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + \epsilon c'_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right) \\ & \quad \geq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}'^{k*}, t^*(\boldsymbol{x}'^{k*})\right)\left(c_f + \epsilon c'_v\left(z\left(\boldsymbol{x}'^{k*}\right)\right)\right), \end{aligned} \tag{24}$$

   where the last inequality follows by the optimality of $(\boldsymbol{x}'^{k*})_{k=1,\cdots,\overline{\eta}}$ for the η-panel design for variable cost function $\epsilon \times c'_v(.)$.

   Next assume the 1-panel cost-dominates the η-panel at variable cost function $c'_v(.)$, that is:

$$\begin{aligned} & TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) \leq TC((\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*}))_{k=1,\cdots,\overline{\eta}}) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f + c'_v(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right) \leq \sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f + c'_v\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right). \end{aligned} \tag{25}$$

   We prove this result by contradiction. Assume that function $c'_v(.)$ is replaced by $c_v(.) = \epsilon \times c'_v(.)$, for some $0 < \epsilon \leq 1$, and assume that the new optimal η-panel design, $(\boldsymbol{x}''^{k*})_{k=1,\cdots,\overline{\eta}}$, incurs a lower expected cost than the 1-panel design, that is:

$$TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})) > TC((\boldsymbol{x}''^{k*}, t^*(\boldsymbol{x}''^{k*}))_{k=1,\cdots,\overline{\eta}})$$

$$\Leftrightarrow D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f+\epsilon c_v'(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right)>\sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}''^{k*}, t^*(\boldsymbol{x}''^{k*})\right)\left(c_f+\epsilon c_v'\left(z\left(\boldsymbol{x}''^{k*}\right)\right)\right).$$

Then, by Eq. (24), we have that if $\epsilon \times c_v'(.)$ $(0<\epsilon\leq 1)$ is replaced by $c_v'(.)$, then the new optimal η-panel design will continue to cost-dominate the 1-panel design, that is, we must have that:

$$D\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*})\right)\left(c_f+c_v'(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}*}\right))\right)>\sum_{k=1}^{\overline{\eta}} D\left(\boldsymbol{x}^{k*}, t^*(\boldsymbol{x}^{k*})\right)\left(c_f+c_v'\left(z\left(\boldsymbol{x}^{k*}\right)\right)\right),$$

reaching a contradiction with Eq. (25). Hence, it must be true that:

$$TC(\boldsymbol{x}^{12\cdots\overline{\eta}*}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}*}))\leq TC((\boldsymbol{x}''^{k*}, t^*(\boldsymbol{x}''^{k*}))_{k=1,\cdots,\overline{\eta}}),$$

and the result follows.

□

Proof of Theorem 4. Consider any budget B, and let $(\boldsymbol{x}^{k*})_{k=1,\cdots,\overline{\eta}}$ denote the corresponding optimal η-panel design, $\eta=1,\cdots,\overline{\eta}$.

1. The proof follows by induction. First we show that *any* 2-panel cost-dominates the 1-panel for *any* variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}): z(\boldsymbol{x}^{12\cdots\overline{\eta}})\geq 2$, i.e., with at least 2 variants. We then use this result to show that if there exists an $\eta+1$-panel design that cost-dominates the optimal η-panel design for a given variant set, then it must be true that the optimal $\eta+1$-panel design FN-dominates the η-panel design at budget B, and we repeat this argument to compare an $\eta+2$-panel with the optimal $\eta+1$-panel, and so on, until adding one more panel is no longer feasible at budget B.

   For the first part of the proof, consider some variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}): z(\boldsymbol{x}^{12\cdots\overline{\eta}})\geq 2$, for which the 1-panel design is feasible at budget B, and consider *any* 2-panel design of this variant set, which we denote by $((\boldsymbol{x}^k)_{k=1,2})$. $D\left(\boldsymbol{x}, t^*(\boldsymbol{x})\right)$ is strictly increasing in in $y(\boldsymbol{x})$ (Lemma E.1). Then, by definition of a 2-panel design, we have that $\boldsymbol{x}^k>\mathbf{0}, k=1,2$, hence it follows that:

$$D\left(\boldsymbol{x}^{12\cdots\overline{\eta}}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}})\right)>D\left(\boldsymbol{x}^k, t^*(\boldsymbol{x}^k)\right), k=1,2. \tag{26}$$

   Further, because $c_f+c_v\left(z(\boldsymbol{x})\right)$ is super-additive in $z(\boldsymbol{x})$, as assumed in part 1 of the theorem, we have that:

$$c_f+c_v\left(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}}\right)\right)\geq c_f+c_v\left(z\left(\boldsymbol{x}^1\right)\right)+c_f+c_v\left(z\left(\boldsymbol{x}^2\right)\right). \tag{27}$$

   Therefore, from Eqs. (26) and (27), we have that:

$$\begin{aligned}D\left(\boldsymbol{x}^{12\cdots\overline{\eta}}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}})\right)\left(c_f+c_v\left(z\left(\boldsymbol{x}^{12\cdots\overline{\eta}}\right)\right)\right)\geq& D\left(\boldsymbol{x}^{12\cdots\overline{\eta}}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}})\right)\left(c_f+c_v\left(z\left(\boldsymbol{x}^1\right)\right)+c_f+c_v\left(z\left(\boldsymbol{x}^2\right)\right)\right)\\ >& D\left(\boldsymbol{x}^1, t^*(\boldsymbol{x}^1)\right)\left(c_f+c_v\left(z\left(\boldsymbol{x}^1\right)\right)\right)+D\left(\boldsymbol{x}^2, t^*(\boldsymbol{x}^2)\right)\left(c_f+c_v\left(z\left(\boldsymbol{x}^2\right)\right)\right),\end{aligned}$$

   that is, any 2-panel design reduces the expected cost over the 1-panel design, that is, it cost-dominates, at any given variant set.

   For the second part of the proof, consider any $\eta=3,\cdots,\overline{\eta}$, and let $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))$ denote the optimal variant set for the η-panel design, with optimal partition $(\boldsymbol{x}^{k*}(\eta)_{k=1,\cdots,\eta})$. Assume, without loss of generality, that $z(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))\geq\eta+1$ (otherwise the result trivially follows with $\overline{k}(B,\eta)=\eta$). Then, there must exist at least one panel with at least 2 variants in the optimal η-panel design, and we partition any such panel (i.e., with at least 2 variants) into 2 panels, converting the given solution into an $\eta+1$-panel design for variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta))$. Because any 2-panel design reduces the expected cost over the 1-panel design for any variant set (shown in the first part of the proof), the new $\eta+1$-panel, which we denote by $(\boldsymbol{x}^k(\eta+1)_{k=1,\cdots,\eta+1})$, remains budget feasible. Then, we have

that:

$$\sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\mathbf{0}\}} \times D\left(\boldsymbol{x}^{k*}(\eta), t^*(\boldsymbol{x}^{k*}(\eta))\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}(\eta)\right)\right)\right) > \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\mathbf{0}\}} \times D\left(\boldsymbol{x}^{k}(\eta+1), t^*(\boldsymbol{x}^{k}(\eta+1))\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k}(\eta+1)\right)\right)\right)$$

Thus, an optimal variant set of an η-panel design can be partitioned into $\eta+1$ panels to obtain a feasible $\eta+1$-panel design, and the result, that the optimal $\eta+1$-panel design FN-dominates the optimal η-panel design, follows.
Similarly, we can show that the optimal $\eta+2$-panel design FN-dominates the optimal $\eta+1$-panel design, which in turn dominates the optimal η-panel design, and so on, that is, the optimal $\eta+k$-panel design FN-dominates the optimal $\eta+k-1$-panel design, as well as all designs with fewer panels. Therefore, $\exists \overline{k}(B,\eta) : z(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta)) \leq \overline{k}(B,\eta) \leq \overline{\eta}$, such that the $\eta+k$-panel design dominates the η-panel design, $\forall k = 0, \cdots, \overline{k}(B,\eta)$.

2. The proof follows by induction. We first show that the 1-panel cost-dominates *any* 2-panel design for *any* variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) : z(\boldsymbol{x}^{12\cdots\overline{\eta}}) \geq 2$, i.e., with at least 2 variants. We then use this result to show that the optimal η-panel design cost dominates *any* $\eta+1$-panel design for *any* variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) : z(\boldsymbol{x}^{12\cdots\overline{\eta}}) \geq \eta+1$. Then, it must be true that the optimal η-panel design FN-dominates the optimal $\eta+1$-panel design at budget B, and we repeat this argument to compare an $\eta+2$-panel with the optimal $\eta+1$-panel, and so on, until adding one more panel is no longer feasible at budget B.

For the first part of the proof, consider some variant set $S(\boldsymbol{x}^{12\cdots\overline{\eta}}) : z(\boldsymbol{x}^{12\cdots\overline{\eta}}) \geq 2$, for which there is a 2-panel design that is feasible at budget B, which we denote by $((\boldsymbol{x}^k)_{k=1,2}) : \boldsymbol{x}^1 + \boldsymbol{x}^2 = \boldsymbol{x}^{12\cdots\overline{\eta}}$. Because $D(\boldsymbol{x}, t^*(\boldsymbol{x})) \geq 0$, and $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is strictly concave increasing in $y(\boldsymbol{x})$ (Lemma E.1), $D(\boldsymbol{x}, t^*(\boldsymbol{x}))$ is sub-additive in $y(\boldsymbol{x})$. Therefore, we can write:

$$\begin{aligned} & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}})\right) \leq D\left(\boldsymbol{x}^1, t^*(\boldsymbol{x}^1)\right) + D\left(\boldsymbol{x}^2, t^*(\boldsymbol{x}^2)\right) \\ \Leftrightarrow\ & D\left(\boldsymbol{x}^{12\cdots\overline{\eta}}, t^*(\boldsymbol{x}^{12\cdots\overline{\eta}})\right)(c_f + c) \leq D\left(\boldsymbol{x}^1, t^*(\boldsymbol{x}^1)\right)(c_f + c) + D\left(\boldsymbol{x}^2, t^*(\boldsymbol{x}^2)\right)(c_f + c). \end{aligned}$$

Thus, the 1-panel design results in a lower or equal expected cost over any 2-panel design, that is, it cost-dominates, for any variant set.

For the second part of the proof, consider any $\eta = 2, \cdots, \overline{\eta}$, and let $S(\boldsymbol{x}^{12\cdots\overline{\eta}*}(\eta+1))$ denote the optimal set of variants, with an optimal$\eta+1$-panel, $(\boldsymbol{x}^{k*}(\eta+1)_{k=1,\cdots,\eta+1})$. Because the 1-panel reduces the expected cost over any 2-panel for any variant set (shown in the first part of the proof), combining any 2 panels in the $\eta+1$-panel, thus creating an η-panel denoted by $(\boldsymbol{x}^k(\eta)_{k=1,\cdots,\eta})$, reduces the expected cost over the optimal $\eta+1$-panel, that is:

$$\sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\mathbf{0}\}} \times D\left(\boldsymbol{x}^{k}(\eta), t^*(\boldsymbol{x}^{k}(\eta))\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k}(\eta)\right)\right)\right) \leq \sum_{k=1}^{\overline{\eta}} I_{\{\boldsymbol{x}^k>\mathbf{0}\}} \times D\left(\boldsymbol{x}^{k*}(\eta+1), t^*(\boldsymbol{x}^{k*}(\eta+1))\right)\left(c_f + c_v\left(z\left(\boldsymbol{x}^{k*}(\eta+1)\right)\right)\right)$$

Thus, the optimal variant set of an $\eta+1$-panel design can be transformed into a feasible η-panel with lower or equal expected cost, and the result, that the optimal η-panel design FN-dominates the optimal $\eta+1$-panel design, follows. Similarly, we can show that the optimal $\eta+1$-panel FN-dominates the optimal $\eta+2$-panel, and so on, that is, the optimal η-panel FN-dominates the optimal $\eta+1$-panel, as well as all designs with higher panels, completing the proof. □

Proof of Corollary 4. Both results follow directly from Theorem 4. □