Efficient Prevalence Estimation for Emerging and Seasonal Diseases Under Limited Resources

Ngoc Thu Nguyen

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

Doctor of Philosophy
in
Industrial and Systems Engineering

Ebru K. Bish
Douglas R. Bish
Ran Jin
Weijun Xie

May 1, 2019
Blacksburg, Virginia

Keywords: Prevalence estimation, Testing pool design, Limited resources, Emerging and/or seasonal diseases, Robust optimization.

Efficient Prevalence Estimation for Emerging and Seasonal Diseases Under Limited Resources

Ngoc Thu Nguyen

(ACADEMIC ABSTRACT)

Estimating the prevalence rate of a disease is crucial for controlling its spread, and for planning of healthcare services. Due to limited testing budgets and resources, prevalence estimation typically entails pooled, or group, testing where specimens (e.g., blood, urine, tissue swabs) from a number of subjects are combined into a testing pool, which is then tested via a single test. Testing outcomes from multiple pools are analyzed so as to assess the prevalence of the disease. The accuracy of prevalence estimation relies on the testing pool design, i.e., the number of pools to test and the pool sizes (the number of specimens to combine in a pool). Determining an optimal pool design for prevalence estimation can be challenging, as it requires prior information on the current status of the disease, which can be highly unreliable, or simply unavailable, especially for emerging and/or seasonal diseases. We develop and study frameworks for prevalence estimation, under highly unreliable prior information on the disease and limited testing budgets. Embedded into each estimation framework is an optimization model that determines the optimal testing pool design, considering the trade-off between testing cost and estimation accuracy. We establish important structural properties of optimal testing pool designs in various settings, and develop efficient and exact algorithms. Our numerous case studies, ranging from prevalence estimation of the human immunodeficiency virus (HIV) in various parts of Africa, to prevalence estimation of diseases in plants and insects, including the Tomato Spotted Wilt virus in thrips and West Nile virus in mosquitoes, indicate that the proposed estimation methods substantially outperform current approaches developed in the literature, and produce robust testing pool designs that can hedge against the uncertainty in model inputs. Our research findings indicate that the proposed prevalence estimation frameworks are capable of producing accurate prevalence estimates, and are highly desirable, especially for emerging and/or seasonal diseases under limited testing budgets.

Efficient Prevalence Estimation for Emerging and Seasonal Diseases Under Limited Resources

Ngoc Thu Nguyen

(GENERAL AUDIENCE ABSTRACT)

Accurately estimating the proportion of a population that has a disease, i.e., the disease prevalence rate, is crucial for controlling its spread, and for planning of healthcare services, such as disease prevention, screening, and treatment. Due to limited testing budgets and resources, prevalence estimation typically entails pooled, or group, testing where biological specimens (e.g., blood, urine, tissue swabs) from a number of subjects are combined into a testing pool, which is then tested via a single test. Testing results from the testing pools are analyzed so as to assess the prevalence of the disease. The accuracy of prevalence estimation relies on the testing pool design, i.e., the number of pools to test and the pool sizes (the number of specimens to combine in a pool). Determining an optimal pool design for prevalence estimation, e.g., the pool design that minimizes the estimation error, can be challenging, as it requires information on the current status of the disease prior to testing, which can be highly unreliable, or simply unavailable, especially for emerging and/or seasonal diseases. Examples of such diseases include, but are not limited to, Zika virus, West Nile virus, and Lyme disease. We develop and study frameworks for prevalence estimation, under highly unreliable prior information on the disease and limited testing budgets. Embedded into each estimation framework is an optimization model that determines the optimal testing pool design, considering the trade-off between testing cost and estimation accuracy. We establish important structural properties of optimal testing pool designs in various settings, and develop efficient and exact optimization algorithms. Our numerous case studies, ranging from prevalence estimation of the human immunodeficiency virus (HIV) in various parts of Africa, to prevalence estimation of diseases in plants and insects, including the Tomato Spotted Wilt virus in thrips and West Nile virus in mosquitoes, indicate that the proposed estimation methods substantially outperform current approaches developed in the literature, and produce robust testing pool designs that can hedge against the uncertainty in model input parameters. Our research findings indicate that the proposed prevalence estimation frameworks are capable of producing accurate prevalence estimates, and are highly desirable, especially for emerging and/or seasonal diseases under limited testing budgets.

Acknowledgements

I have had some of best time of my life at Virginia Tech, and everything that I have accomplished thus far would not have been possible without the help, support, and guidance of many incredible individuals.

First, I would like to thank my Ph.D. advisors, Dr. Ebru Bish and Dr. Douglas Bish, for their never-ending support and faith in me, not only during my doctoral study, but also during my time as an undergraduate student. I am forever thankful for their kindness, patience, enthusiasm, and insights that have inspired me to take on a career in Operations Research, and have guided me throughout my study; and for their work ethics, dedication, and professionalism that have improved me both professionally and personally.

I would also like to thank my committee members, Dr. Ran Jin and Dr. Weijun Xie, for their insightful guidance and suggestions that have improved the quality of my work, and for all of their advices and endless support for my professional goals.

As I navigate through my graduate study and professional life, I am deeply grateful to the professors and mentors who have made incredibly positive impacts on my academic career: Dr. Joel Nachlas, Dr. Maury Nussbaum, and Dr. Hrayer Aprahamian. Thank you for inspiring me to take on a career in research, and for always guiding me towards the right directions.

I am also thankful to all the friends I have made at Virginia Tech, and to my friends from Vietnam. Thanks to each and everyone of you, my time at Virginia Tech, and my life abroad in general, has been an amazing journey.

Last, but certainly not least, I am truly and deeply grateful to my big extended family for their unconditional, never-ending, and overflowing love, support, and faith in me no matter what endeavor I am up to: my parents, Chuc and Thu; my sisters, Thuy and Van; my brother, Tuan; my brothers- and sister-in-law, Nam, Cuong, and Hai; my favorite cousin, Trang; my nephew, Michou; my niece, Soc; and my dear grandparents. Thank you for always loving, supporting, and believing in me, at my best or worst. None of this would have been possible without all of you by my side.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter is organized as follows. In Section 1.1, we describe the context and motivation for our research problem and provide a brief overview of current practices for prevalence estimation of diseases. Then, in Section 1.2, we outline our approaches to testing pool design optimization for prevalence estimation of emerging and/or seasonal diseases.

1.1 Motivation

Emerging and/or seasonal diseases are becoming more and more common, and are causing great harm worldwide; examples include Zika virus, West Nile virus, Lyme disease, and Babesiosis. These diseases pose significant challenges to public health and healthcare system management due to their highly stochastic nature; and their outbreaks can be extremely costly to the society, leading to poor health outcomes and economic losses. Consider, for example, the Zika virus outbreak in 2016, affecting Central America, the Caribbean, and the northern part of South America [87], costing these regions around $3.5 billion in disease treatment costs and short-term economic losses alone [107]. Yet another costly consequence of emerging and seasonal diseases is the rising costs of blood transfusion. As there are more transfusion-transmittable diseases for which donated blood is screened, the cost of blood transfusion has increased substantially, from around $1,100 in 2012 to around $3,600 in 2017 [61, 62]. Therefore, an efficient and effective surveillance method for emerging and seasonal diseases is of utmost importance to predict, control, and mitigate their outbreaks [32, 35].

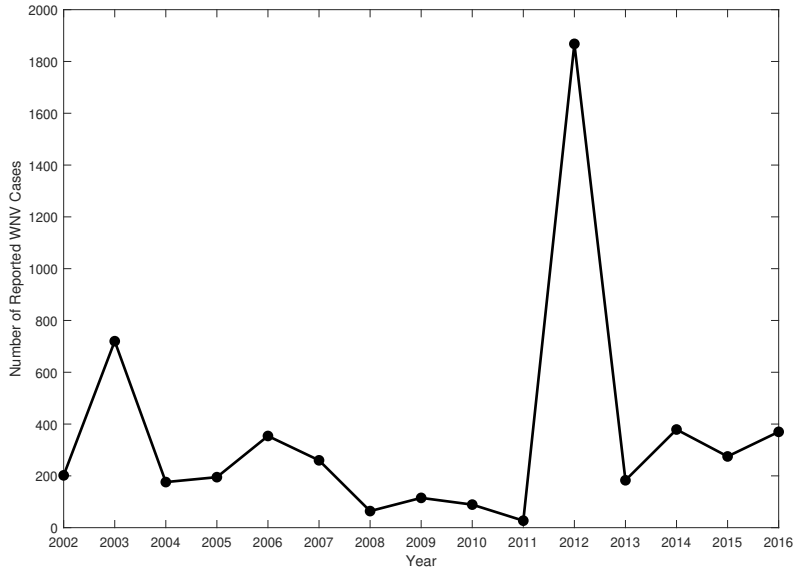
An important input for establishing an efficient and effective surveillance method is the estimated prevalence rate of the disease in question. Prevalence estimation typically involves large-scale testing of subjects (humans, insects, plants) via in-vitro laboratory tests performed on biological specimens, such as blood,

urine, or tissue swabs, collected from the subjects, in order to measure a disease-specific bio-marker, which serves as an indicator for the presence of the virus or bacteria causing the disease. However, the testing efforts are typically constrained by limited testing budgets and resources, and, as a result, individual testing of each subject is highly inefficient, or simply infeasible, for large populations, because disease prevalence rates are often very low [99]. Therefore, prevalence estimation typically entails *pooled (group) testing* of specimens from multiple subjects, which refers to the practice of combining specimens from multiple subjects in a testing pool, and measuring the pool's concentration (load) of a bio-marker. Thus, one test is used on the multiple specimens in the pool. The inference on the unknown prevalence rate is then made based on an analysis of the pooled test outcomes. Hence, an important decision in prevalence estimation is the testing pool design, i.e., how many testing pools to use, and how many specimens to combine in each pool (pool size).

The testing pool design problem requires prior information on the disease, e.g., an initial estimate of the current prevalence rate of the disease, the distribution of the bio-marker load in disease-positive subjects. However, for an emerging or a highly seasonal disease, prior information regarding its characteristics is often highly unreliable, or even unavailable. As an example, consider the prevalence rates of West Nile virus in Texas between 2002 and 2016, shown in Figure 1.1. Given the substantial fluctuations in the disease prevalence rate from year to year, simply using the disease prevalence from a previous year as an input for testing pool design can lead to highly inefficient testing pools and, hence, to inaccurate prevalence estimates. Consequently, determining optimal testing designs for prevalence estimation of emerging and/or seasonal diseases poses significant challenges to public health policy-makers and practitioners.

Given the challenges of the testing pool design problem for prevalence estimation, various methods for prevalence rate estimation have been developed in the statistics literature, in an effort to overcome the impact of inefficient pool designs. On the other hand, the number of studies that focus on the pool design aspect of prevalence estimation is very limited. Further, these studies on pool design are typically restrictive, in that they assume a fixed number of pools, or tests, they do not explicitly account for limited testing budgets, and while they require an initial estimate of the unknown prevalence rate, they do not offer a mechanism to update this estimate as testing proceeds, or a mechanism to hedge against the uncertainty on the initial prevalence rate estimate. As a result, the resulting pool designs can lead to inaccurate prevalence estimates, especially for emerging and/or seasonal diseases, for which limited, and highly unreliable, information is available prior to testing. Motivated by these gaps in the knowledge-base, our research goal is to develop robust and efficient testing pool designs for prevalence estimation, with a special focus on emerging and seasonal diseases under limited testing resources.

Figure 1.1: Number of reported cases of West Nile virus (WNV) in Texas from 2002 to 2016 (Data Source: [24])



1.2 Research Overview

We develop and study various novel frameworks for optimal testing pool design for prevalence estimation, while relaxing the restrictive assumptions often used in the existing literature. As discussed above, the use of an exogenously fixed number of pools is a very common practice in prevalence estimation, and is often assumed in current pool design models. We relax this assumption in all testing pool design models that we develop, in order to explicitly account for, and to efficiently allocate, the available testing budget. We also relax other restrictive assumptions in our models to provide guidelines on best practices for optimal pool design for prevalence estimation.

In particular, in Chapter 2, we study a sequential and adaptive prevalence estimation procedure that utilizes continuous test outcomes (i.e., the bio-marker concentration), as opposed to the commonly used binary test outcomes (i.e., the test outcome is positive if the bio-marker concentration is above a pre-set threshold, and negative otherwise). Relaxing the assumption of binary test outcomes allows us to model and account for testing errors and the dilution effect of pooling (i.e., the potential reduction in test sensitivity as pool size increases), so as to accurately estimate an unknown prevalence rate. Specifically, we propose a two-stage sequential estimation procedure in which the pool design is optimized in every stage, under testing budget constraints, based on the most recent estimate of the unknown prevalence rate; this estimate is revised after the first stage of testing, to take into account the new information obtained via the first stage

of testing. As a result, the prevalence estimate derived from this sequential estimation procedure is highly accurate, in comparison to estimates derived from other commonly used single-stage estimation procedures with binary test outcomes.

We complement the findings and analysis of Chapter 2 with a novel methodology for estimating the sensitivity of pooled testing, based on an integration of a viral load progression model with a probit model to consider the dilution effect of pooling, presented in Chapter 3. Our viral load progression model accounts for viral load progression throughout the lifetime of the disease, as opposed to existing models that only consider the window period of the disease. Our case study, of testing pool design for prevalence estimation of the human immunodeficiency virus (HIV) infection in various parts of Africa, indicates that the proposed methodology generates highly accurate sensitivity values for pooled testing, and, as a result, can lead to efficient pool designs for both prevalence estimation and subject classification.

In Chapter 4, we establish key structural properties of optimal testing pool designs for prevalence estimation of diseases under the commonly used assumption of binary test outcomes. More importantly, we relax the assumption of an exogenously fixed number of testing pools, and develop a joint optimization model that determines both the pool size and the number of pools, while explicitly accounting for the testing budget. Existing studies on pool design for prevalence estimation are mostly numerical in nature, and do not offer any general insight and guidelines into optimal testing pool design in different settings. As opposed to this, we establish several analytical properties of the asymptotic variance function and optimal pool designs, which allow us to provide guidelines for public health practitioners, further contributing to the existing literature.

In Chapter 5, we develop and study robust pool design optimization models that hedge against the uncertainty in the initial prevalence estimate of the disease. Robust pool design optimization is essential, because the testing pool design problem for prevalence estimation requires, as an input, an initial estimate of the unknown prevalence rate, creating significant challenges, especially for emerging and/or highly seasonal diseases, due to a lack of reliable information about the current status of such diseases. While one can account for such uncertainty via a multi-stage sequential estimation procedure studied in Chapter 2, such procedures may not be desirable for large-scale testing of subjects, due to operational challenges and complexity, including the need for additional on-site data collection, and additional operational decisions, e.g., how to split up the available testing budget among different testing stages. Therefore, in Chapter 5, we apply robust optimization methodologies and extend upon the pool design optimization models in Chapter 4, in order to hedge against the uncertainty in the initial prevalence estimate of the disease, even within a single-stage estimation framework. Specifically, we consider a mini-max model and a regret-based model, both of which require minimal information on the disease prior to testing, i.e., the support of the disease prevalence rate. We study robust pool designs derived from the proposed models, and compare them with

pool designs derived from the deterministic model presented in Chapter 4, in both single-stage and sequential estimation frameworks. Our analysis underscores the value of robust optimization in pool design for prevalence estimation, and leads to key insights and guidelines for practitioners.

We conclude this dissertation in Chapter 6 with a summary of research findings. To facilitate the presentation, mathematical derivations, analytical proofs, and some tables are relegated to the Appendix.

Chapter 2

Sequential Prevalence Estimation with Pooling and Continuous Test Outcomes

2.1 Introduction

Surveillance, or prevalence estimation, is an essential component for assessing the current status or dynamics of an infection, a disease, or a genetic disorder. In the following, we use the term "infection" to refer to the binary characteristic, the prevalence rate of which we would like to estimate using in vitro laboratory testing via bio-specimens (e.g., blood, urine, tissue swabs) collected from subjects in a certain population. As a recent example that highlights the need for efficient and effective surveillance methods, consider the recent outbreak of the Zika virus infection in central America, the Caribbean, and the northern part of south America [87], affecting 45 countries and territories across the Americas [78]. Given its association with the congenital syndrome, Guillain-Barre syndrome, and other neurological disorders, accurately estimating the prevalence rate of the Zika virus infection becomes crucial in planning for healthcare services. In addition to emerging infections, prevalence estimation plays an important role in controlling the spread of existing infections, an example of which is the human immunodeficiency virus (HIV) infection. The requested federal funding for HIV prevention programs in fiscal year 2017 totaled $27.5 billion in the United States [97]. With such substantial funding at stake, HIV regional prevalence rate estimates play an important role in the allocation of funds to the different regions [19]. Accurate prevalence rate estimation is also important for

controlling plant, insect, and animal diseases [12, 56, 81].

As surveillance studies are often constrained by testing budgets, however, individual testing is rendered inefficient, or simply infeasible (e.g., [99,100]). Therefore, ever since its introduction by Dorfman in 1943 [38], pooled, or group, testing is considered to be a highly efficient and effective approach to both classification (i.e., identification of all infected subjects) and estimation problems. Pooled testing for the purpose of classification is often followed by individual testing to identify the infected subjects. On the other hand, in the estimation problem, which is the focus of this paper, the identification component is often not necessary, as the ultimate goal is to derive an accurate estimate of the infection prevalence rate (e.g., [50,56,69,70,72]).

In pooled testing, individual bio-specimens are pooled together, and the pool concentration of a bio-marker, which serves as an indicator for the presence of the virus causing the infection, is measured. The bio-marker may, for example, correspond to the antibody or antigen level, which can be measured by serologic tests, or genetic material from the virus, such as viral RNA or DNA, which can be measured using virologic testing technology. While these bio-marker measurements are continuous in nature, for both classification and estimation purposes they are typically compared to pre-set thresholds to produce a binary test outcome: "positive" if the pooled test's reading exceeds the threshold, indicating the presence of at least one infected specimen in the pool, and "negative" otherwise. Several aspects of the estimation problem under binary test outcomes, and for a fixed number of pools, i.e., with no explicit testing budget limitation, have been studied in the literature. Several studies investigate the characteristics of the maximum likelihood estimator (MLE) of the prevalence rate in various settings, including multi-pool configurations, i.e., pools with different sizes [27], and under relaxation of the commonly used binomial assumption [26], while others develop various approaches for bias reduction in the MLE (e.g., [54]). Furthermore, since pooling design may impact the estimation efficiency significantly (see, e.g., [25, 92, 96]), many studies focus on the pool size determination problem, under perfect tests (e.g., [57,58,89]), and imperfect tests (e.g., [45,69,99,100]). All these studies use the MLE and assume a homogeneous population. While some recent studies account for the heterogeneity of the population through Bayesian analyses [40, 81, 94], or through regression analyses, which allow individual covariate information to be utilized in the estimation [28, 41, 55, 102, 110], these studies do not consider an optimal pooling design. The aforementioned studies on the pool size determination problem use binary pooled test outcomes and assume that pooling does not alter the test's sensitivity (true positive probability) and specificity (true negative probability). These assumptions can lead to inaccurate prevalence rate estimates: It is well-known that the bio-marker concentration of infected specimens may be "diluted" by the uninfected specimens in the pool, to the point that the pool's reading may fall below the test's positivity threshold (the "dilution effect" of pooling), with the binary test outcome becoming a false negative (e.g., [59, 91]). Furthermore, valuable information on the pool's bio-marker concentration is lost when binary test outcomes

are used in the estimation (e.g., [111]). Therefore, a more accurate approach to inferring about the unknown prevalence rate based on pooled testing is to directly utilize the continuous reading of the test, along with a model that explicitly considers the dilution effect of pooling, as we do in this paper.

In particular, Zenios and Wein [111] develop an estimation procedure that uses continuous test outcomes from pooled testing and considers the dilution effect of pooling, to estimate the HIV prevalence rate for a homogeneous population via the Enzyme Linked Immunosorbent Assay (ELISA). They also determine the optimal pool size to minimize the cost per unit information when the total number of specimens to be tested is fixed *a priori*. While their pool size optimization model requires an initial estimate of the unknown prevalence rate, it does not offer a mechanism to update this estimate as testing proceeds, i.e., it is a single-stage estimation procedure. McMahan, Tebbs and Bilder [72] extend this single-stage estimation procedure to a heterogeneous population; while the continuous test outcome distribution is accounted for in their derivation of the MLE of the prevalence rate, the prevalence rate inference is ultimately made based on binary test outcomes, i.e., by converting the continuous test outcomes to binary outcomes via a pre-set threshold, and without consideration of the optimal pool size.

An alternative approach is to use a sequential and adaptive estimation procedure, which allows the prevalence rate estimate to be revised as testing proceeds, so that the remaining tests can be conducted with a more effective pooling design that is based on a more accurate prevalence rate estimate. This is the approach we take in this paper. Of particular relevance to our model are the works by Hughes-Oliver and Swallow [58], and Hughes-Oliver and Rosenberger [57], both of which utilize binary test outcomes and are based on the perfect test assumption, i.e., there is no testing error, hence no dilution. Various sequential estimation-classification procedures have also been proposed for the ultimate goal of classification, i.e., the first stage typically involves the estimation of the unknown prevalence rate, while the second stage involves the identification of the infected specimens in the pool; e.g., [15, 46, 51, 95, 103]; thus, these works are not within the scope of this paper.

Specifically, we study a sequential and adaptive estimation procedure in which the pooling design in each stage is optimized with respect to a testing budget constraint and is based on the most recent estimate of the unknown prevalence rate, which changes as testing progresses. Moreover, the estimation is based on continuous test outcomes, and explicitly considers the dilution effect of pooling. From this perspective, our estimation procedure can be seen as an integration of the methodologies proposed in Hughes-Oliver and Swallow [58] and Zenios and Wein [111], with the following enhancements: Our model allows for multiple pooling configurations in each stage, enhancing the flexibility of testing, and our pooling optimization model *simultaneously* determines the number of pools *and* the pool sizes given a testing budget and based on the most recent prevalence rate estimate.

The remainder of this paper is organized as follows. In Section 2.2, we provide details of the proposed estimation procedure; in Section 2.3, we demonstrate its effectiveness through two case studies that respectively focus on estimating the prevalence rate for HIV via pooled HIV Nucleic Acid Amplification Test (NAT), Ultrio Plus Assay; and for the Tomato Spotted Wilt Virus (TSWV) in thrips using the pooled ELISA test. Finally, we conclude in Section 2.4 with a discussion of our findings and suggestions for future research. To facilitate the presentation, some tables and mathematical derivations are relegated to the Appendix.

2.2 The Proposed Sequential and Adaptive Estimation Procedure

Pooling design, which involves determining a pool size (m) and the number of pools (n), is an important decision in pooled testing (e.g., [25,71,85,92,96]). Determining an optimal pooling design requires an initial estimate of the unknown prevalence rate p, which we denote by p_0; and inaccurate choices of p_0 can lead to sub-optimal pooling designs, which, in turn, can lower the estimation efficiency of the MLE, resulting in a higher mean squared error (MSE) and/or a higher bias. Under pooled testing with binary outcomes, researchers show that the MLE of the prevalence rate is robust to pooling design when the initial estimate, p_0, is close to, or greater than, p; however, this is not necessarily the case when p_0 is an underestimate of p [58, 92, 98]. Meanwhile, it is shown that using continuous test outcomes in pooled testing can improve the efficiency of estimation, e.g., [104, 111]. However, when using continuous test outcomes, the optimal pooling design as well as the resulting MLE of the prevalence rate depend not only on p_0, but also on the distribution of the bio-marker concentration in infected subjects, which is rarely known with certainty; and this dependence has not been well-studied in the literature.

Motivated by these observations, we propose a sequential and adaptive estimation procedure that utilizes continuous test outcomes, and study its efficiency and robustness, with respect to deviations from the initially assumed prevalence rate estimate and bio-marker distribution. Specifically, our approach expands upon that in [58] by accounting for the testing error (i.e., false positive and false negative test outcomes are possible), continuous test outcomes, and the dilution effect of pooling; and expands upon that in [111] by offering a mechanism to update the estimate of the unknown prevalence rate for use in the pooling optimization model, i.e., it is sequential and adaptive; further, the pooling optimization model *jointly* optimizes for both pool sizes and the number of pools under a testing budget constraint. However, unlike [111], we consider the measurement error of the employed testing kit to be negligible, but we still account for the testing error, i.e., false positives and false negatives, by incorporating, into our estimation model, the dilution effect of pooling and the "noise" on the pool's reading contributed by the uninfected specimens in the pool. We also briefly discuss the impact of the measurement error on our results. Our sequential estimation procedure can

be easily extended to a multi-stage estimation procedure, with three or more stages, in the same manner as the procedure proposed by Hughes-Oliver and Swallow [58].

In this section, we outline the various components of the proposed estimation procedure. In particular, Section 2.2.1 provides an overview of our estimation procedure. Then, Sections 2.2.2 and 2.2.3 respectively detail the probabilistic model used to represent the continuous test outcomes and the dilution effect of pooling, and the construction of the MLE based on continuous test outcomes. Then Section 2.2.4 presents the pooling design optimization model, which is embedded into the sequential estimation procedure.

Throughout, we denote vectors in boldface, random variables in upper-case letters and their realization in lower-case letters, and use $F_Y(\cdot)$ and $f_Y(\cdot)$ to respectively denote the cumulative distribution function (CDF) and probability density function (pdf) of a random variable Y; see Appendix A.1 for a summary of the notation.

The goal is to obtain an accurate estimate of the unknown prevalence rate, p, which we assume to be homogeneous across the population, using a total testing budget, B, via sequential and adaptive testing. Towards this end, in each stage of the testing procedure, the decision-maker determines the pooling design in that stage, based on the most recent estimate of the prevalence rate, i.e., the decision-maker determines both pool sizes, $\mathbf{m} = \{m_1, m_2, \cdots, m_C\}$, and the number of pools of each size to test, $\mathbf{n} = \{n_1, n_2, \cdots, n_C\}$, for a given $C \in \mathbb{N}$, where C is the number of pooling configurations. Hence, the decision variable in each stage is the pooling design matrix,

$$\mathbf{D} = \begin{bmatrix} m_1 & m_2 & \cdots & m_C \\ n_1 & n_2 & \cdots & n_C \end{bmatrix}.$$

Let us denote the optimal pooling design in stage s, $s = 1, 2$, by $\mathbf{D}_\mathbf{s}^*$, which is the solution to an optimization problem that minimizes the MSE of the prevalence rate estimate (see Section 2.2.4).

2.2.1 Outline of the Sequential and Adaptive Estimation Procedure

The sequential estimation procedure (**SE**) is outlined below.

Sequential and Adaptive Estimation Procedure with Continuous Test Outcomes (SE):

For a given λ, $\lambda \in (0, 1]$, and a total testing budget, B:

Stage 1:

1. Given p_0, an initial estimate on p, determine the optimal pooling design for stage 1 that is feasible

with respect to a testing budget of $B^{(1)} = \lambda B$, given by:

$$\mathbf{D_1^*}(\lambda, p_0) = \begin{bmatrix} m_{11}^* & m_{12}^* & \cdots & m_{1C}^* \\ n_{11}^* & n_{12}^* & \cdots & n_{1C}^* \end{bmatrix}.$$

2. Obtain the set of continuous test outcomes corresponding to the pooling design, $\mathbf{D_1^*}(\lambda, p_0)$, and compute the MLE of p in stage 1, $\hat{p}_{MLE}^{(1)}$.

Stage 2:

1. Given $\hat{p}_{MLE}^{(1)}$, the (random) outcome of stage 1, determine the optimal pooling design for stage 2 that is feasible with respect to a testing budget of $B^{(2)} = (1-\lambda)B$, given by[1]:

$$\mathbf{D_2^*}((1-\lambda), \hat{p}_{MLE}^{(1)}) = \begin{bmatrix} m_{21}^* & m_{22}^* & \cdots & m_{2C}^* \\ n_{21}^* & n_{22}^* & \cdots & n_{2C}^* \end{bmatrix}.$$

2. Obtain the set of continuous test outcomes corresponding to the pooling design $\mathbf{D_2^*}\big((1-\lambda), \hat{p}_{MLE}^{(1)}\big)$, and use both stage 1 and stage 2 testing outcomes to compute $\hat{p}_{MLE}^{(2)}$, i.e., the final estimate of p.

In our case studies in Section 2.3,we investigate the effectiveness of single- versus dual-configuration pooling designs, and of single-stage versus sequential estimation procedures under different values of the budget allocation factor, λ, the testing budget, B, and the initial estimate of p at the beginning of stage 1, $p_0^{(1)}$; and under different bio-marker distributions.

2.2.2 Modeling the Pool's Continuous Test Outcome

To model the continuous test outcome of a pool of size m, we define the following random variables:

Y^+: Bio-marker concentration of a random infected subject, with pdf $f_{Y^+}(\cdot)$

Y^-: Noise level (i.e., contribution to a pool's concentration) coming from a random uninfected subject, with pdf $f_{Y^-}(\cdot)$

Let Y_i denote the bio-marker concentration of subject i, $i = 1, \cdots, m$, whose specimen is included in a pool of size m, that is, for all $i = 1, \cdots, m$,

$$Y_i = \begin{cases} Y_i^+, & \text{with probability } p \\ Y_i^-, & \text{with probability } 1-p \end{cases},$$

[1]Any unused budget in stage 1 (due to integrality constraints on $\mathbf{m}$ and $\mathbf{n}$), is added to $B^{(2)}$.

and we assume that $Y_i, i = 1, \cdots, m$, are independent, that is, the infection status, hence, bio-marker concentration, of different subjects are independent.

Similar to [111] and [104], we represent the dilution effect by modeling the pool's reading as the *average* bio-marker concentration of the pool, given by $Y^{(m)} = \frac{\sum_{i=1}^{m} Y_i}{m}$, with pdf:

$$f_{Y^{(m)}}(y;p) = \sum_{k=0}^{m} Pr[W(m;p) = k]\ f_{Y^{(m;k)}}(y) = \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{(m-k)}\ f_{Y^{(m;k)}}(y),$$

where $W(m;p)$ denotes the number of infected specimens in a pool of size m, i.e., $W(m;p) \sim$ Binomial (m,p), and $Y^{(m;k)}$ is the conditional average bio-marker concentration of a pool of size m, given k infected specimens in the pool, i.e., $Y^{(m;k)} = \frac{1}{m}\Big(\sum_{i=1}^{k} Y_i^+ + \sum_{j=1}^{m-k} Y_j^-\Big)$, for $k \leq m$, $k \in \mathbb{N}$. Let $S^{(m,k)}$ denote the sum of concentrations of all specimens in a pool of size m, with k infected specimens, i.e., $S^{(m,k)} = mY^{(m,k)}$. We have the following:

$$f_{S^{(m;k)}} = f_{Y^+}^{[k*]} * f_{Y^-}^{[(m-k)*]},\ k = 0, 1, 2, \cdots, m, \tag{2.1}$$

where $f_Y^{[n*]}$ denotes the n-fold convolution of f_Y, $\forall n \in \mathbb{N}$. Thus, the density function $f_{Y^{(m;k)}}$ is given by:

$$f_{Y^{(m;k)}}(y^{(m,k)}) = m \cdot f_{S^{(m;k)}}(m \cdot y^{(m,k)}),\ k = 0, 1, 2, \cdots, m. \tag{2.2}$$

2.2.3 Constructing the MLE under Continuous Test Outcomes

Given a pooling design $\mathbf{D}$, the observed test outcome (i.e., the vector of average reading of each pool) is denoted by $\mathbf{y}(\mathbf{m},\mathbf{n}) = \{\mathbf{y}^{(m_1,n_1)}, \cdots, \mathbf{y}^{(m_C,n_C)}\}$, with each $\mathbf{y}^{(m_i,n_i)} = \big(y_j^{(m_i)}\big)_{j=1,\cdots,n_i}$, $i = 1, \cdots, C$, denoting a vector of n_i test outcomes, each corresponding to a pool of size m_i. Then, the likelihood function is given by:

$$L\big(p; \mathbf{y}(\mathbf{m},\mathbf{n})\big) = \prod_{i=1}^{C}\prod_{j=1}^{n_i}\Big(\sum_{k=0}^{m_i}\binom{m_i}{k} p^k (1-p)^{(m_i-k)} f_{Y^{(m_i;k)}}(y_j^{(m_i)})\Big).$$

Then, extending upon the MLE expression in [111], the MLE for p, corresponding to a pooling design $\mathbf{D}$ and test outcome vector $\mathbf{y}(\mathbf{m},\mathbf{n})$, follows:

$$\hat{p}_{MLE} = \frac{1}{(\sum_{i=1}^{C} n_i m_i)} \sum_{i=1}^{C}\sum_{j=1}^{n_i}\sum_{k=0}^{m_i} k\tau^{(m_i)}(k; y_j^{(m_i)}, \hat{p}_{MLE}), \tag{2.3}$$

where $\tau^{(m)}(k;y,p) = Pr\Big(W(m;p) = k \mid Y^{(m)} = y\Big)$, i.e., the conditional probability of having k infected specimens in a pool of size m, given the pool's reading of y and a prevalence rate of p [111]:

$$\tau^{(m)}(k;y,p) = \frac{\binom{m}{k}p^k(1-p)^{(m-k)}f_{Y^{(m;k)}}(y)}{\sum_{j=0}^{m}\binom{m}{j}p^j(1-p)^{(m-j)}f_{Y^{(m;j)}}(y)}, \qquad \text{for } k \leq m, k \in \mathbb{N}. \tag{2.4}$$

From Eqn.(2.3), the MLEs in stages 1 and 2 of the sequential estimation procedure, **SE**, follow:

$$\begin{aligned}
\hat{p}^{(1)}_{MLE} &= \frac{1}{(\sum_{i=1}^{C} n^*_{1i}m^*_{1i})}\sum_{i=1}^{C}\sum_{j=1}^{n^*_{1i}}\sum_{k=0}^{m^*_{1i}} k\tau^{(m^*_{1i})}(k;y_j^{(m^*_{1i})},\hat{p}^{(1)}_{MLE}), \text{ and} \\
\hat{p}^{(2)}_{MLE}(\hat{p}^{(1)}_{MLE}) &= \frac{1}{(\sum_{s=1}^{2}\sum_{i=1}^{C} n^*_{si}m^*_{si})}\sum_{s=1}^{2}\sum_{i=1}^{C}\sum_{j=1}^{n^*_{si}}\sum_{k=0}^{m^*_{si}} k\tau^{(m^*_{si})}(k;y_j^{(m^*_{si})},\hat{p}^{(2)}_{MLE}),
\end{aligned} \tag{2.5}$$

where (m^*_{si}, n^*_{si}), $s = 1, 2, i = 1, \cdots, C$, is the solution to the pooling design optimization model (see Section 2.2.4). Thus, $\hat{p}^{(2)}_{MLE}$ is the final output of **SE**, i.e., the estimate of p, and the only purpose of deriving $\hat{p}^{(1)}_{MLE}$ is to use it as an input for the pooling optimization model in stage 2. We then solve for $\hat{p}^{(1)}_{MLE}$ and $\hat{p}^{(2)}_{MLE}$ numerically, using a tolerance level of 10^{-7}; see Appendix A.2.

The above analysis considers that the test's outcome for a pool may include some noise originating from uninfected specimens in the pool due to various reasons (e.g., health conditions or medications unrelated to the infection in question). If there is a significant amount of measurement error in the test's outcome that is independent of the number of uninfected specimens in the pool, then one can explicitly model this measurement error. If this is the case, then, similar to [111], let $X^{(m)}$ denote the *measured* bio-marker concentration of a pool of size m, with an *actual* (and *unobservable*) bio-marker concentration of $Y^{(m)}$; and Eqs. (2.4) and (2.5) continue to hold, with $f_{Y^{(m;k)}}(\cdot)$ replaced by $f_{X^{(m;k)}}(\cdot)$, and with the pdf of the random variable $X^{(m;k)}$ expressed as a function of the pdf of the random variable $Y^{(m;k)}$; see [111] for details, and see Appendix A.8 for the potential impact of the test's measurement error on estimation efficiency in our setting.

2.2.4 Pooling Design Optimization

In this section, we present the optimization model used in **SE** to determine the optimal pooling design, (m^*_{si}, n^*_{si}), $s = 1, 2, i = 1, \cdots, C$, that minimizes the MSE of the estimator in stage s, where $p_0^{(1)} = p_0$ and $p_0^{(2)} = \hat{p}^{(1)}_{MLE}$:

Pooling Design Optimization Problem for SE (stage s, s = 1, 2) :

$$\begin{aligned} \min_{m_{si}, n_{si}, i=1,\cdots,C} \quad & MSE(\hat{p}^{(s)}_{MLE}; p^{(s)}_0) \\ \text{subject to:} \quad & c_f(\sum_{i=1}^{C} n_{si}) + c_v(\sum_{i=1}^{C} n_{si} m_{si}) \leq B^{(s)} \\ & m_{si} \leq M, \quad m_{si}, n_{si} \in \mathbb{N}, \quad i = 1, 2, \cdots, C \end{aligned} \tag{2.6}$$

where c_f is the testing cost per pool, c_v is the collection cost per specimen, $B^{(s)}$ is the testing budget available in stage s, and M is a technological upper bound (if any) on pool sizes. In **SE**, we split the total testing budget, B, between the two testing stages using some multiplier $\lambda \in (0, 1)$, so that the budget for stage 1 is $B^{(1)} = \lambda B$, and that for stage 2 is $B^{(2)} = (1 - \lambda)B$; see Section 2.2.1.

We next derive an expression for the MSE in our setting, i.e., for imperfect tests and considering continuous test outcomes and the dilution effect of pooling.

MSE Derivation:

We first consider the case where a single-pool configuration, $\mathbf{D} = \{(m, n)\}$, is utilized in stage s, $s = 1, 2$:

$$MSE\big(\hat{p}^{(s)}_{MLE}; p^{(s)}_0\big) = Var(\hat{p}^{(s)}_{MLE}; p^{(s)}_0) + Bias^2(\hat{p}^{(s)}_{MLE}; p^{(s)}_0), \tag{2.7}$$

where:

$$Var(\hat{p}^{(s)}_{MLE}; p^{(s)}_0) = Var\left(\frac{1}{nm}\sum_{j=1}^{n}\sum_{k=0}^{m} k\tau^{(m)}(k; Y^{(m)}_j, \hat{p}^{(s)}_{MLE})\right) = \Big(\frac{1}{nm^2}\Big) Var\left(E[W(m; p^{(s)}_0) \mid Y^{(m)}]\right), \tag{2.8}$$

$$Bias^2(\hat{p}^{(s)}_{MLE}; p^{(s)}_0) = \Big(E[\hat{p}^{(s)}_{MLE}; p^{(s)}_0] - p_0\Big)^2 = \left(\frac{1}{m} E\Big(E[W(m; p^{(s)}_0) \mid Y^{(m)}]\Big) - p^{(s)}_0\right)^2. \tag{2.9}$$

Substituting Eqs. (2.8) and (2.9) into Eqn. (2.7), we obtain:

$$MSE(\hat{p}^{(s)}_{MLE}; p^{(s)}_0) = \Big(\frac{1}{nm^2}\Big) Var\left(E[W(m; p^{(s)}_0) \mid Y^{(m)}]\right) + \left(\frac{1}{m} E\Big(E[W(m; p^{(s)}_0) \mid Y^{(m)}]\Big) - p^{(s)}_0\right)^2, \tag{2.10}$$

where the first two moments of the random variable $E[W(m;p_0^{(s)}) \mid Y^{(m)}]$, $s=1,2$, are given by:

$$\begin{aligned} E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big] &= E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}] \mid W(m;p_0^{(s)})=0\Big] Pr[W(m;p_0^{(s)})=0] \\ &\quad + E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}] \mid W(m;p_0^{(s)}) \geq 1\Big] Pr[W(m;p_0^{(s)}) \geq 1] \\ &= E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}] \mid W(m;p_0^{(s)}) \geq 1\Big] \big(1-(1-p_0^{(s)})^m\big), \\ E\Big[\Big(E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big)^2\Big] &= E\Big[\Big(E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big)^2 \mid W(m;p_0^{(s)}) \geq 1\Big] \big(1-(1-p_0^{(s)})^m\big). \end{aligned} \tag{2.11}$$

Substituting Eqn. (2.11) into Eqn. (2.10), we obtain, $s=1,2$:

$$\begin{aligned} MSE(\hat{p}_{MLE}^{(s)};p_0^{(s)}) = &\Big(\frac{1}{nm^2}\Big)\left\{E\Big[\Big(E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big)^2\Big] - \left(E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big]\right)^2\right\} \\ &+ \left\{\Big(\frac{1}{m}\Big)E\Big[E[W(m;p_0^{(s)}) \mid Y^{(m)}]\Big] - p_0^{(s)}\right\}^2. \end{aligned} \tag{2.12}$$

For $m=1$, Eqn. (2.12) reduces to: $MSE(\hat{p}_{MLE}^{(s)};p_0^{(s)}) = \frac{1}{n}\, p_0^{(s)}(1-p_0^{(s)})$, $s=1,2$.

Similarly, for the multiple pooling configuration case, i.e., $C \geq 2$, the MSE in each estimation stage s, $s=1,2$, given a starting estimate $p_0^{(s)}$, can be expressed as:

$$\begin{aligned} MSE(\hat{p}_{MLE}^{(s)};p_0^{(s)}) = &\left(\frac{1}{(\sum_{i=1}^{C} n_{si}m_{si})^2}\right)\left\{\sum_{i=1}^{C} n_{si} Var\Big(E[W(m_{si};p_0^{(s)}) \mid Y^{(m_{si})}]\Big)\right\} \\ &+ \left\{\left(\frac{1}{\sum_{i=1}^{C} n_{si}m_{si}}\right)\sum_{i=1}^{C} n_{si}\ E\Big(E[W(m_{si};p_0^{(s)}) \mid Y^{(m_{si})}]\Big) - p_0^{(s)}\right\}^2. \end{aligned} \tag{2.13}$$

The computation of the $MSE(\hat{p}_{MLE}^{(s)};p_0^{(s)})$ is also important for approximating a confidence interval for the unknown prevalence rate, p. For example, for **SE**, $\hat{p}_{MLE}^{(2)}$ is the final estimate of p. Then, from [58, 98], $\big(\hat{p}_{MLE}^{(2)} - p\big) \sim$ Normal $\Big(Bias(\hat{p}_{MLE}^{(2)};p), Var(\hat{p}_{MLE}^{(2)};p)\Big)$, and the confidence interval of p, with a confidence level of $100(1-\alpha)\%$, can be approximated by:

$$\left[\hat{p}_{MLE}^{(2)} - Bias(\hat{p}_{MLE}^{(2)};p)\right] \pm z_{1-\alpha/2}\sqrt{Var(\hat{p}_{MLE}^{(2)};p)},$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. We propose an approximation method for $Bias(\hat{p}_{MLE}^{(2)};p)$ and $Var(\hat{p}_{MLE}^{(2)};p)$ based on the continuous test outcomes and $\hat{p}_{MLE}^{(2)}$ in Appendix A.10.

Observe that the expressions of MSE in Eqn. (2.12) and (2.13) rely on the first two moments of the random variable, $E[W(m;p_0^{(s)}) \mid Y^{(m)}]$, i.e., the conditional expectation of the number of infected specimens

in a random pool of size m, given an initial estimate of $p_0^{(s)}$ on p, and the pool's reading $Y^{(m)}$, which is unknown, and, thus, uncertain at the time the MSE is computed for use in the pooling design optimization model. Further, $\hat{p}_{MLE}^{(s)}$, for $s = 1, 2$, is numerically computed, by obtaining a solution to Eqn. (2.5) via an iterative algorithm (see Appendix A.2); hence, the distribution of the random variable $E[W(m; p_0^{(s)}) \mid Y^{(m)}]$ cannot be expressed in closed-form. As a result, deriving an exact expression for $MSE(\hat{p}_{MLE}^{(s)}; p_0^{(s)})$ is analytically challenging. Therefore, in the case studies of Section 2.3, we simulate the distribution of the random variable $\left[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\right]$, and approximate its first two moments, which are then used in Eqs. (2.12) and (2.13) to approximate $MSE(\hat{p}_{MLE}^{(s)}; p_0^{(s)})$, for $m \geq 2, m \in \mathbb{N}$.

2.3 Case Studies

In this section, we present and discuss two case studies: the HIV prevalence rate estimation via the HIV NAT Ultrio Plus Assay, which measures the viral RNA concentration, i.e., the viral load, and the TSWV prevalence rate estimation for thrips via the ELISA test, which measures the antibody concentration. Our objectives are: **(1)** to compare the effectiveness of: **(i)** utilizing continuous versus binary test outcomes (the latter with and without the numerical corrections proposed in the literature, [54]), **(ii)** single-stage versus sequential estimation (i.e., **SE**) procedures, and **(iii)** single- versus dual-configuration pooling designs, for estimating an unknown prevalence rate; **(2)** to study the robustness of the sequential procedure, **SE**, to deviations from the initial prevalence rate estimate, $p_0^{(1)}$, and the assumed distribution and parameters of bio-marker levels, Y^+, and of noise terms, Y^-; and **(3)** to understand the impact of the testing budget, B, and budget allocation factor, λ, on the efficiency of the estimation.

This section is organized as follows. Section 2.3.1 provides an overview of the numerical study and the simulation model. Then, Sections 2.3.2 and 2.3.3 respectively discuss the findings from the HIV case study and the TSWV case study.

2.3.1 Description of the Numerical Study

Both case studies are conducted via Monte–Carlo simulations, considering a wide range of parameter values. An input to each simulation replication is the optimal pooling design in stage 1, $\mathbf{D}_\mathbf{1}^* = (m_{1i}^*, n_{1i}^*)_{i=1,\cdots,C}$, which is a function of the initial prevalence rate estimate, $p_0^{(1)}$, and the assumed distributions and parameters of Y^+ and Y^-. In what follows, we first describe the implementation of the pooling design optimization model, and then detail the simulation procedures.

Implementation of the Pooling Design Optimization Model:

We solve the pooling design optimization problem, given in Eqn. (2.6), via an exhaustive search procedure

that considers the entire feasible region, comprised of *all* budget-feasible combinations of (m_{1i}, n_{1i}), $i = 1, \cdots, C$, i.e., $c_f \sum_{i=1}^{C} n_{1i} + c_v \sum_{i=1}^{C} n_{1i} m_{1i} \leq B^{(1)}$, such that n_{1i} (number of pools) is between 1 and 1,000, and m_{1i} (pool size) is between 1 and M, i.e., the maximum acceptable pool size for the utilized testing kit, $\forall i = 1, \cdots, C$. We consider that $M = 48$ in case study 1, and $M = 50$ in case study 2. The exhaustive search procedure is implemented in MATLAB, which simply computes the MSE value for each budget-feasible combination of (m_{1i}, n_{1i}) and returns the combination (m_{1i}^*, n_{1i}^*), $i = 1, \cdots, C$, that yields the lowest value of MSE (i.e., the optimal pooling design). Since we enumerate and examine all budget-feasible combinations of (m_{1i}, n_{1i}), the optimality of the resulting pooling design, (m_{1i}^*, n_{1i}^*), is guaranteed. Both the MATLAB code and equivalent R code are included in Web Supplementary Materials.

Observe that the objective of the pooling design optimization problem is to minimize the corresponding MSE, $MSE(\hat{p}_{MLE}^{(s)}; p_0^{(s)})$, $s = 1, 2$, which relies on the first two moments of the random variable $\big[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\big]$ (see Eqs. (2.12) and (2.13)). As discussed in Section 2.2.3, we do not have closed-form expressions for the first two moments of this random variable. Therefore, in both case studies, we first simulate the distribution of $\big[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\big]$ for various values of m and $p_0^{(s)}$ (this distribution does not depend on the number of pools, n). In particular, for each value of m and $p_0^{(s)}$, we perform a Monte–Carlo simulation to generate 100,000 pools, each of size m, where each specimen is infected with probability $p_0^{(s)}$. For each infected specimen, we generate its corresponding biomarker concentration from the distribution of Y^+, and for each uninfected specimen, we generate its noise from the distribution of Y^-. Then, we determine the pool's reading, $Y^{(m)}$, and compute:

$$\big[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\big] = \sum_{k=1}^{m} k Pr[W(m; p_0^{(s)}) = k \mid Y^{(m)}, W(m; p_0^{(s)}) \geq 1],$$

where $Pr[W(m; p_0^{(s)}) = k \mid Y^{(m)}, W(m; p_0^{(s)}) \geq 1] = \frac{\binom{m}{k} p^k (1-p)^{(m-k)} f_{Y^{(m;k)}}(y)}{\sum_{j=1}^{m} \binom{m}{j} p^j (1-p)^{(m-j)} f_{Y^{(m;j)}}(y)}$ for $1 \leq k \leq m, k \in \mathbb{N}$. We then use the 'allfitdist' function in MATLAB for each pair of m and $p_0^{(s)}$ to fit a distribution to random variable $\big[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\big]$. In both case studies, the fitted distribution of random variable $\big[E[W(m; p_0^{(s)}) \mid Y^{(m)}] \mid W(m; p_0^{(s)}) \geq 1\big]$ turns out to be log-normal $(\mu(m, p_0^{(s)}), \sigma^2(m, p_0^{(s)}))$, where both parameters are functions of the pool size, m, and the initial estimate $p_0^{(s)}$, $s = 1, 2$, leading to the following approximations (see Eqn. (2.11)):

For $m \geq 2, m \in \mathbb{N}$,

$$\begin{aligned} E\Big[E[W(m; p_0^{(s)}) \mid Y^{(m)}]\Big] &\approx \exp\Big[\mu(m, p_0^{(s)}) + \frac{\sigma^2(m, p_0^{(s)})}{2}\Big]\big(1 - (1 - p_0^{(s)})^m\big), \\ E\Big[\Big(E[W(m; p_0^{(s)}) \mid Y^{(m)}]\Big)^2\Big] &\approx \exp\Big[2\mu(m, p_0^{(s)}) + 2\sigma^2(m, p_0^{(s)})\Big]\big(1 - (1 - p_0^{(s)})^m\big). \end{aligned} \tag{2.14}$$

The functional forms of the parameters, $\mu(m, p_0^{(s)})$ and $\sigma^2(m, p_0^{(s)})$, are determined via regression analyses. Specifically, when we fit a regression function to the data generated by the Monte–Carlo simulation, we compute the resulting values of $\sigma^2(m, p_0)$ using the fitted function so as to ensure that these values are positive for all values of m (i.e., if this is not the case, then we adjust the regression function such that the computed $\sigma^2(m, p_0)$ value is always positive). Although this process may lead to lower adjusted R^2 values, for all values of m considered in case studies 1 and 2, the adjusted R^2 value of the fitted function for $\sigma^2(m, p_0)$ is above 0.90; see Appendices E and F for case studies 1 and 2, respectively.

Using the approximations in Eqn. (2.14), the approximation for $MSE(\hat{p}_{MLE}^{(s)}; p_0^{(s)})$, $s = 1, 2$, for a single-configuration pool follows (see Eqn. (2.12)):

$$MSE(\hat{p}_{MLE}^{(s)}; p_0^{(s)}) \approx \left(\frac{1}{nm^2}\right)\left(1-(1-p_0^{(s)})^m\right)\Big(\exp\left[2\mu(m,p_0^{(s)})+2\sigma^2(m,p_0^{(s)})\right]-\exp\left[2\mu(m,p_0^{(s)})+\sigma^2(m,p_0^{(s)})\right]\left(1-(1-p_0^{(s)})^m\right)\Big)$$
$$+\left(\frac{1}{m}\left(1-(1-p_0^{(s)})^m\right)\exp\left[\mu(m,p_0^{(s)})+\frac{\sigma^2(m,p_0^{(s)})}{2}\right]-p_0^{(s)}\right)^2, \qquad \text{for } m \geq 2, m \in \mathbb{N}. \tag{2.15}$$

We tabulate the MSE values approximated via Eqn. (2.15) for each combination of $(m, p_0^{(s)})$, and utilize these values in the pooling design optimization model. The approximation for the multiple-pool configuration case follows similarly. Our numerical study indicates that this approximation is fairly accurate; see Appendix A.9 for details.

For the two infections that are considered in the case studies of this section, practitioners interested in utilizing the proposed estimation procedure can simply utilize the MSE function approximation in Eqn. (2.15), along with the coefficients given in Tables A.3 – A.6, to determine the optimal pooling design. For other infections, practitioners will need to derive the corresponding MSE function based, for example, on simulation studies as we do here.

The Simulation Procedure:

We next describe the simulation procedure. Prior to the simulation of each scenario, we solve the pooling design optimization model, detailed above, to determine the optimal pooling design in stage 1, $\mathbf{D_1^*} = (m_{1i}^*, n_{1i}^*)_{i=1,\cdots,C}$, based on $p_0^{(1)}$, and the assumed distributions and parameters of Y^+ and Y^-. Then, in each simulation replication, we assign an infection status to each specimen using a Bernoulli random number generator with the "true" prevalence rate of p. For each infected specimen, we generate its bio-marker concentration from the "true" distribution of Y^+, and for each uninfected specimen, we generate its noise from the "true" distribution of Y^- (the true distributions of Y^+ and Y^- correspond to the assumed distributions unless otherwise stated; only when we study the robustness to deviations from the assumed distributions, the true distributions differ from the assumed distributions, and are unknown to the experimenter). The specimens are then assigned to pools randomly following the optimal pooling design, $\mathbf{D_1^*}$. Each pool's reading

is computed by averaging out the bio-marker concentrations and noise levels in the pool, and the corresponding MLE, $\hat{p}^{(1)}_{MLE}$, is computed and used to determine the optimal pooling design in stage 2, $\mathbf{D}^*_{\mathbf{2}}$, and the simulation is repeated in the same manner to obtain the final MLE of p, $\hat{p}^{(2)}_{MLE}$, which we simply denote by $\hat{p}_{MLE}$. When $\lambda = 1$, **SE** terminates at the end of the first stage (i.e., it is a single-stage procedure), with an output of $\hat{p}_{MLE}$.

In the single-stage procedure that utilizes binary test outcomes, we compare the average bio-marker concentration of each pool to a pre-set threshold: If the average bio-marker concentration exceeds the threshold, then the pool outcome is "positive;" otherwise, it is "negative." Based on the number of positive pools, the estimator, $\hat{p}^{(B)}_{MLE}$, in the binary test outcome case, is then computed as follows [111]:

$$\hat{p}^{(B)}_{MLE} = 1 - \left(\frac{Se(m,Th) - \frac{n^+}{n}}{Se(m,Th) + Sp(m,Th) - 1} \right)^{\frac{1}{m}},$$

where n^+, $Se(m,Th)$, and $Sp(m,Th)$ respectively denote the number of positive pools, and test sensitivity and specificity given a threshold, Th. Consequently, for the simulation of the binary test outcome case, we only need to generate the realizations, n^+, for each simulation replication.

For the single-stage procedure that utilizes binary test outcomes, we also compute and report the *corrected MLE*, based on the numerical *horizontal* and *vertical* correction procedures in [54]. In general, these corrections aim to reduce the bias of the MLE obtained from binary outcomes of pooled testing; see Appendix A.3 for details.

For each estimation procedure that we consider, we compute and report the following performance measures based on 20,000 simulation replications:

(i) the average value of $\hat{p}_{MLE}$ observed over 20,000 replications, denoted by *MLE* in the tables - this serves as an estimate of $E[\hat{p}_{MLE}; p_0]$;

(ii) the average value of $(\hat{p}_{MLE} - p)^2$ observed over 20,000 replications, denoted by *MSE* in the tables - this serves as an estimate of $MSE(\hat{p}_{MLE}; p)$;

(iii) the relative bias ($rBias(\%)$) of the estimated $E[\hat{p}_{MLE}; p_0]$ (denoted by MLE, see performance measure (i)):

$$rBias(\%) = 100 \left| \frac{MLE - p}{p} \right|.$$

We also compute and report the sample standard deviation for $\hat{p}_{MLE}$ and $(\hat{p}_{MLE} - p)^2$; see Section 2.3 and Appendix A.1 for mathematical expressions.

2.3.2 Case Study 1 – HIV Prevalence Rate Estimation

In the first case study, we investigate the efficiency and robustness of the sequential estimation procedure for estimating the prevalence rate of HIV in various parts of Africa.

To construct the viral load distribution of an HIV-infected individual at a random post-infection time, we use published efficacy data for the HIV NAT Ultrio Plus Assay [91,105,106] and the HIV viral load progression model developed in [4, 74], and assume that the time, from exposure, when the infected individual takes the test follows a uniform distribution [2]; see Appendix A.4 for details.

Based on a set of randomly generated viral load realizations from the HIV viral load distribution (Appendix A.4), we approximate Y^+ by $Gamma(\alpha^+, \beta)$, with shape parameter $\alpha^+ = 0.4195$ and inverse scale parameter $\beta = 3.8040 \times 10^{-7}$ (with Akaike Information Criterion (AIC) score[3] of 2.9×10^6). (This corresponds to the "assumed" distribution that the decision-maker uses for pooling design and estimation; we examine the robustness of the estimation procedures to deviations from the assumed distribution of Y^+ in Section 2.3.2). We model the noise level, Y^-, coming from an uninfected individual, as $Gamma(\alpha^-, \beta)$, with shape parameter $\alpha^- = 10^{-12}$, to represent the case where the mean noise is close to zero. Then, $Y^{(m;k)} \sim Gamma(\alpha_k, m\beta)$, with $\alpha_k = k\alpha^+ + (m-k)\alpha^-$, for $k \leq m, k, m \in \mathbb{N}$; see Eqn. (2.2).

We consider a testing cost per pool, c_f, of \$31.5 [63] and a collection cost per specimen, c_v, of \$8 [33]. For the other parameters, we consider a range of values through a number of *scenarios*, where each scenario is characterized by the triplet, $(p, p_0^{(1)}, B)$, i.e., the true (and unknown) HIV prevalence rate, the initial prevalence rate estimate, and the total testing budget. Specifically, we consider, $p \in \{0.022, 0.044, 0.071\}$, which respectively represent the HIV prevalence rate in Western and Central Africa [8], Sub-Saharan Africa [108], and East and Southern Africa [7]; $p_0^{(1)} \in \left\{\frac{p}{2}, \frac{3}{2}p\right\}$, which respectively correspond to scenarios that initially underestimate and overestimate the actual rate p, and three levels of the testing budget, $B \in \{\$3,345; \$4,460; \$5,575\}$. In the pooling design optimization model, we consider pool sizes of $m \in [1, 48]$, which is consistent with the literature on the HIV NAT Ultrio Plus Assay (e.g., [82, 86]).

The remainder of this section is organized as follows. Section 2.3.2 studies the impact of utilizing binary test outcomes versus continuous test outcomes. Section 2.3.2 compares the efficiency of the sequential estimation procedure, **SE**, and the single-stage procedure, considering various values of the budget allocation factor, λ, which determines the split of the testing budget between the two stages of **SE**. Finally, Section 2.3.2 studies the robustness of **SE** and of the single-stage procedure (both under continuous test outcomes)

[2] We consider a uniform distribution between 0 and 100 days post-infection to account for the different phases of the HIV infection [106].

[3] AIC score is a relative measure of model fit to a given set of data [2]. The AIC score rewards goodness of fit based on the likelihood function and penalizes over-fitting based on the number of parameters being fit. The best-fitting model has the lowest AIC score in comparison to other models.

to deviations from the assumed distribution of the bio-marker concentration (viral load), i.e., distribution of Y^+.

The optimal pooling designs for each budget level, B, and initial estimate, $p_0^{(1)}$, are summarized in Table A.7; see Appendix A.7.

Estimation Efficiency for Binary versus Continuous Test Outcomes

We first examine the impact of utilizing continuous test outcomes over binary test outcomes on the MLE of the prevalence rate and the efficiency of the estimation, measured in terms of MSE and the relative bias (rBias). It is well-known that the MLE obtained in pooled testing via binary test outcomes is biased [58,100], and several correction procedures have been developed in the literature to account for the bias. In this section, we consider a single-stage estimation procedure: for the binary outcome case, we report the MLE obtained directly from the binary test outcomes, as well as the corrected MLEs under the numerical vertical and horizontal correction procedures proposed by Hepworth and Watson [54], and compare them with those obtained under continuous test outcomes; see Tables 2.1 and 2.2 for the corresponding performance measures respectively corresponding to scenarios with a true prevalence rate, p, of 0.071 and 0.044.

Table 2.1: Case Study 1: Performance measures for the single-stage estimation procedure with binary and continuous test outcomes, $p = 0.0710$. MLE and MSE are reported in the form: sample average (± sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Meas.	Binary (No Correction)	Binary (Vertical Correction)	Binary (Horizontal Correction)	Continuous
		MLE	0.0896 (±0.0335)	0.0913 (±0.0490)	0.0570 (±0.0216)	0.0725 (±0.0152)
	0.0355	MSE ($\times 10^4$)	14.7 (±24.0)	28.1 (±54.5)	6.63 (±5.12)	2.32 (±3.83)
		rBias(%)	26.20	28.58	19.62	2.15
$5,575						
		MLE	0.0740 (±0.0165)	0.0726 (±0.0160)	0.0500 (±0.0108)	0.0716 (±0.0136)
	0.1065	MSE ($\times 10^4$)	2.83 (±4.80)	2.58 (±4.14)	5.55 (±4.32)	1.86 (±2.78)
		rBias(%)	4.28	2.31	29.6	0.87
		MLE	0.0837(±0.0332)	0.0791 (±0.0413)	0.0538 (±0.0204)	0.0726 (±0.0728)
	0.0355	MSE ($\times 10^4$)	12.6 (±31.3)	17.7 (±60.4)	7.09 (±7.40)	2.70 (±2.76)
		rBias(%)	17.91	11.34	24.20	2.18
$4,460						
		MLE	0.0742 (±0.0182)	0.0726 (±0.0176)	0.0502 (±0.0119)	0.0718 (±0.0152)
	0.1065	MSE ($\times 10^4$)	3.42 (±5.75)	3.12 (±5.01)	5.74 (±4.80)	2.31 (±3.46)
		rBias(%)	4.45	2.31	29.26	1.11
		MLE	0.0827 (±0.0345)	0.0775 (±0.0408)	0.0599 (±0.0289)	0.0730 (±0.0188)
	0.0355	MSE ($\times 10^4$)	13.3 (±32.6)	17.0 (±59.2)	9.57 (±28.3)	3.56 (±6.00)
		rBias(%)	16.53	9.22	19.62	2.80
$3,345						
		MLE	0.0744 (±0.0214)	0.0722 (±0.0202)	0.0559 (±0.0155)	0.0719 (±0.0177)
	0.1065	MSE ($\times 10^4$)	4.68 (±8.02)	4.11 (±6.48)	4.70 (±4.90)	3.13 (±4.67)
		rBias(%)	4.77	1.72	21.29	1.32

Table 2.2: Case Study 1: Performance measures for the the single-stage estimation procedure with binary and continuous test outcomes, $p = 0.0440$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Meas.	Binary (No Correction)	Binary (Vertical Correction)	Binary (Horizontal Correction)	Continuous
		MLE	0.0467 ($\pm$0.0174)	0.0431 ($\pm$0.0176)	0.0347 ($\pm$0.0133)	0.0450 ($\pm$0.0109)
	0.0110	MSE ($\times 10^4$)	3.08 ($\pm$10.10)	3.10 ($\pm$17.2)	2.64 ($\pm$8.77)	1.20 ($\pm$1.98)
		rBias(%)	6.12	1.95	21.15	2.20
$5,575						
		MLE	0.0445 ($\pm$0.0119)	0.0436 ($\pm$0.0115)	0.0343 ($\pm$0.0087)	0.0446 ($\pm$0.0103)
	0.0660	MSE ($\times 10^4$)	1.42 ($\pm$2.38)	1.33 ($\pm$2.10)	1.70 ($\pm$1.73)	1.06 ($\pm$1.59)
		rBias(%)	1.18	0.93	22.10	1.29
		MLE	0.0477($\pm$0.0205)	0.0434 ($\pm$0.0221)	0.0349 ($\pm$0.0160)	0.0453 ($\pm$0.0124)
	0.0110	MSE ($\times 10^4$)	4.33 ($\pm$14.30)	4.88 ($\pm$27.0)	3.39 ($\pm$12.8)	1.54 ($\pm$2.53)
		rBias(%)	8.45	1.30	20.76	2.94
$4,460						
		MLE	0.0446 ($\pm$0.0131)	0.0436 ($\pm$0.0127)	0.0343 ($\pm$0.0096)	0.0445 ($\pm$0.0114)
	0.0660	MSE ($\times 10^4$)	1.73 ($\pm$2.85)	1.62 ($\pm$2.56)	1.87 ($\pm$1.95)	1.31 ($\pm$1.99)
		rBias(%)	1.27	0.89	22.10	1.19
		MLE	0.0498 ($\pm$0.0236)	0.0471 ($\pm$0.0286)	0.0356 ($\pm$0.0171)	0.0459 ($\pm$0.0145)
	0.0110	MSE ($\times 10^4$)	5.89 ($\pm$11.6)	8.28 ($\pm$19.6)	3.64 ($\pm$4.16)	2.14 ($\pm$3.72)
		rBias(%)	13.18	6.98	19.08	4.38
$3,345						
		MLE	0.0449 ($\pm$0.0156)	0.0434 ($\pm$0.0147)	0.0342 ($\pm$0.0112)	0.0447 ($\pm$0.0135)
	0.0660	MSE ($\times 10^4$)	2.45 ($\pm$4.77)	2.22 ($\pm$3.77)	4.70 ($\pm$2.48)	1.81 ($\pm$2.87)
		rBias(%)	1.99	1.46	22.32	1.68

Tables 2.1 and 2.2 indicate that using continuous outcomes improves the estimation efficiency, in terms of both bias and MSE, especially when the unknown p is large (e.g., scenarios with $p = 0.071$ in Table 2.2). However, when p is small (e.g., scenarios with $p = 0.022$; see Tables A.8 and A.9 in Appendix A.7), utilizing continuous test outcomes introduces more bias into MLE, especially in scenarios where the initial estimate is an overestimate, i.e., $p_0^{(1)} > p$. In the case of binary outcomes, while the correction procedures reduce the bias in some cases (e.g., see vertical correction in scenarios with $p = 0.022$), they have a tendency to over-adjust the prevalence rate estimate in the presence of testing errors (see horizontal correction in all scenarios). This effect is especially amplified when p is large, as in the case of $p = 0.071$. These findings support the use of continuous test outcomes, especially for estimating the prevalence of emerging infections, for which initial prevalence estimates may be highly unreliable. Furthermore, when utilizing binary test outcomes, an appropriate testing threshold needs to be specified, and there are often no clear guidelines on how this should be done.

The estimation efficiency of the single-stage procedure that utilizes continuous test outcomes can be further enhanced by relaxing the assumption of a single-configuration pooling design, i.e., $C = 1$; see the optimization model in Section 2.2.4. This relaxation expands the feasible region of the optimization problem

(2.6), and may lead to a better pooling design that further reduces the MSE of the prevalence rate MLE. Our numerical studies indicate that the estimation efficiency of single- and dual-configuration pooling designs is quite similar for scenarios where $p = 0.046$; however, the dual-configuration design improves both the estimation efficiency and robustness over the single-configuration design in scenarios where p is larger, i.e., $p = 0.071$, and when there is uncertainty regarding the distributions of Y^+ and Y^-. Consequently, in Sections 2.3.2 and 2.3.2, we utilize a dual-configuration pooling design in **SE**, i.e., $C = 2$, together with continuous test outcomes.

Estimation Efficiency for Single-stage versus Sequential Estimation Procedures

Next, we compare the estimation efficiency of the proposed sequential procedure, **SE**, and the single-stage procedure, considering continuous test outcomes. Tables A.10 and A.11 (see Appendix A.7) report the estimation efficiency, in terms of MSE and relative bias, of the single-stage procedure and **SE** (with a budget allocation factor of $\lambda \in \{0.25, 0.5\}$) for various scenarios. As these results indicate, **SE** performs especially well when $p_0^{(1)}$ is an underestimate of p, i.e., $p_0^{(1)} < p$, yielding both lower bias and lower MSE in comparison to the single-stage procedure. Furthermore, even in cases where $p_0^{(1)}$ is an overestimate of p, **SE**, with an appropriate choice of λ, is at least as efficient as the single-stage procedure, and in some cases, it produces a significantly lower bias; see, e.g., scenarios with $p = 0.071$ and $p = 0.044$. The choice of λ is especially important when the budget is tight and $p_0^{(1)}$ is an underestimate of p: When p is large, using a large value of λ, i.e., $\lambda = 0.5$, results in both higher MSE and higher bias, due to a higher weight placed on the initial estimate of p. When p is small, however, using a smaller value of λ, i.e., $\lambda = 0.25$, may lead to an insufficient number of pools in the first stage due to the tight budget, which results in an MLE that can potentially be lower than the true prevalence rate at the end of the first stage. These findings highlight the importance of setting the budget allocation parameter, λ, appropriately, depending on the setting.

Robustness: Effect of Incorrect Specification of the Bio-marker Concentration Distribution

Utilizing continuous test outcomes in an estimation procedure requires the experimenter to assume a distribution for bio-marker concentration in infected individuals (Y^+), e.g., the HIV viral load distribution in this case study. In Section 2.3.2, we approximate the HIV viral load distribution via a Gamma distribution with parameters α^+ and β; see Appendix A.4. Therefore, we now study the robustness of the sequential and single-stage procedures to departures from the assumed viral load distribution.

Specifically, while implementing the estimation procedures, the experimenter continues to estimate that Y^+ is Gamma (α^+, β) (see Section 2.3.2); this assumption is used for both designing the testing pools and for estimating p. However, in the Monte–Carlo simulation, the "true" viral loads of infected individuals

are generated from a different distribution. In particular, for the true distribution, we generate a random time t, i.e., the time post-infection when the infected individual is tested, from a Uniform distribution with support in $(0, 100)$ (see footnote [2]), and compute the viral load using the analytical viral load model given in Appendix A.4. Thus, we compare the two settings: (i) when the true distribution of Y^+ corresponds to its estimated distribution (i.e., Gamma (α^+, β) (labeled as "correct distribution of Y^+"), and (ii) when the true distribution of Y^+ deviates from its estimated distribution (labeled as "incorrect distribution of Y^+").

From the analysis in Section 2.3.2, scenarios where $p_0^{(1)}$ is an underestimate of p result in the least efficient MLE, with high MSE and bias. Therefore, in this section, we focus on scenarios where $p_0^{(1)} < p$, specifically, $p_0^{(1)} = \frac{p}{2}$, for a budget of $\$4,460$, and for $p \in \{0.022, 0.044, 0.071\}$. The results are provided in Table 2.3.

Table 2.3: Case Study 1: Performance measures for the single-stage estimation procedure and **SE** with continuous outcomes, with correct and incorrect distributions for Y^+, $B = \$4,460$. MLE and MSE are reported in the form: sample average (± sample standard deviation).

	p	$p_0^{(1)}$	Perf. Measures	Single-stage	**SE**	
					$\lambda = 0.25$	$\lambda = 0.5$
Correct Distribution of Y^+	0.0220	0.0110	MLE	0.0232 (±0.0086)	0.0218(±0.0093)	0.0228 (±0.0081)
			MSE ($\times 10^4$)	0.76 (±1.43)	0.86 (±1.39)	0.66 (±1.11)
			rBias (%)	5.43	0.95	3.42
	0.0440	0.0220	MLE	0.0453 (±0.0124)	0.0447(±0.0121)	0.0447 (±0.0118)
			MSE ($\times 10^4$)	1.54 (±2.53)	1.47 (±2.48)	1.40 (±2.26)
			rBias (%)	2.94	1.58	1.70
	0.0710	0.0355	MLE	0.0726 (±0.0728)	0.0719(±0.0156)	0.0722 (±0.0160)
			MSE ($\times 10^4$)	2.70 (±2.76)	2.45 (±3.78)	2.59 (±4.17)
			rBias (%)	2.18	1.31	1.71
Incorrect Distribution of Y^+	0.0220	0.0110	MLE	0.0265 (±0.0107)	0.0217(±0.0100)	0.0227 (±0.0081)
			MSE ($\times 10^4$)	1.35 (±2.88)	0.99 (±1.72)	0.67 (±1.08)
			rBias (%)	20.51	1.55	3.06
	0.0440	0.0220	MLE	0.0532 (±0.0164)	0.0446(±0.0123)	0.0447 (±0.0121)
			MSE ($\times 10^4$)	3.53 (±6.45)	1.51 (±2.67)	1.47 (±2.53)
			rBias (%)	20.84	1.45	1.68
	0.0710	0.0335	MLE	0.0876 (±0.0235)	0.0716(±0.0158)	0.0716 (±0.0158)
			MSE ($\times 10^4$)	8.25 (±13.2)	2.51 (±3.81)	2.49 (±3.53)
			rBias (%)	23.33	0.84	0.91

Table 2.3 demonstrates that using an incorrect viral load distribution can greatly deteriorate the performance of the single-stage estimation procedure, significantly increasing the bias and MSE; this impact is particularly significant when p is large, i.e., $p = 0.071$, mainly due to a higher probability of having a pool

with infected specimens, whose viral load distribution is incorrectly specified. On the other hand, **SE** is robust to incorrect choices of the viral load distribution: even though the assumed viral load distribution is not updated at the end of the first stage of **SE**, the pooling design is re-optimized based on a revised, and likely more accurate, estimate of p. This re-optimization of the pooling design seems sufficient when p is relatively small, and only the assumed distribution of Y^+ is inaccurate. In this case, using a smaller value of λ in **SE**, i.e., $\lambda = 0.25$, is especially beneficial, as this places less weight on the first stage, in which a poor initial estimate may be used, and allows for more flexibility in the second stage, which is based on a revised pooling design. However, when p is relatively large, and the assumed distributions of both Y^+ and Y^- are inaccurate, re-optimizing the pooling design based on a revised the estimate of p does not appear to be sufficient for **SE** to be robust to these deviations, as demonstrated in Section 2.3.3. Therefore, in these cases, the distributions of Y^+ and Y^-, as well as the estimate of p, need to be adjusted at the end of stage 1 to improve the robustness of **SE**. We suggest this as an important future research direction (Section 2.4.3).

2.3.3 Case Study 2 – Tomato Spotted Wilt Virus Prevalence Estimation in Plants

In our second case study, we apply our methodology to the surveillance of plant diseases. In particular, we consider the problem of estimating the prevalence rate of the Tomato Spotted Wilt Virus (TSWV) in thrips via the Double Sandwich ELISA test. Studies have found a significant correlation between the number of western flower thrips and TSWV disease incidence [30]. Thus, the detection of TSWV in thrips is important in predicting disease outbreak [29].

The ELISA test, which measures the antibody concentration, is often used to detect infections in plants [88], including the TSWV in thrips [29]. Thus, the bio-marker in this setting corresponds to the TSWV-antibody concentration in a thrip, which is measured in terms of the "absorbance readings" (i.e., the A_{405nm} value) by the ELISA test. In particular, we utilize the absorbance readings data for *Frankliniella occidentalis*, published in [29], to construct the bio-marker distribution for TSWV-infected thrips, as well as the noise distribution coming from uninfected thrips.

TSWV Absorbance Readings Distribution

Based on data on mean and standard deviation of absorbance readings in *Frankliniella occidentalis*, summarized in Table 2.4, we construct the absorbance readings distribution for TSWV-infected thrips as log-normal ($\mu_+ = -1.495, \sigma^2{}_+ = 0.4204$), and the noise coming from uninfected thrips as log-normal ($\mu_- = -4.410, \sigma^2{}_- = 0.2826$); both set of parameters correspond to the average of the data, reported

in [29] and summarized in Table 2.4.

Table 2.4: Case Study 2: Number of *Frankliniella occidentalis* adult thrips testing positive for TSWV by Double Sandwich ELISA and absorbance readings (A_{405nm}) of ELISA-positive thrips [29]

Test Number	Plant Status	Infected thrips/total tested	A_{405nm} value (mean ± SD)
3	Infected	10/49	0.23 ± 0.12
3	Healthy	0/32	0.002 ± 0.004
5	Infected	64/104	0.28 ± 0.22
5	Healthy	0/114	0.02 ± 0.01
6	Infected	112/197	0.32 ± 0.26
6	Healthy	0/74	0.02 ± 0.01

We consider a true prevalence rate, p, of 0.12, which corresponds to the prevalence rate of TSWV in adult thrips reported in [29]. We consider a testing cost per pool of \$1.35, a collection cost per specimen of \$0.04 [111], and a testing budget of \$52.5, which is sufficient for testing of 30 pools, each with pool size 10. Similar to case study 1, we consider a range of budget allocation factors, i.e., $\lambda \in \{0.25, 0.5\}$. Further, to model the scenarios of underestimation and overestimation, we consider ${p_0}^{(1)} \in \left\{\frac{p}{3}, \frac{p}{2}, \frac{3p}{2}, \frac{5p}{3}\right\}$. Thus, in this case study, a scenario is defined by the parameter, $p_0^{(1)}$.

In order to study the robustness of each estimation procedure to deviations from the assumed distributions and parameters of the absorbance readings for infected thrips (Y^+) and of the noise coming from uninfected thrips (Y^-), we consider that the "true" distributions and parameters are as follows (each case is implemented independently of other cases): (1) the true distribution of Y^+ is log-normal, but with parameters ($\mu_+ = -1.4428, {\sigma^2}_+ = 0.4869$), corresponding to the *weighted* average of the data in [29], see Table 2.4; (2) the true distribution of Y^+ is a multi-modal distribution, which is a combination of three log-normal distributions with respective parameters $({\mu_+}, {\sigma^2}_+) = (-1.5901, 0.2408), (-1.5134, 0.4808), (-1.3929, 0.5069)$, corresponding to each set of mean and variance pairs reported in Table 2.4, rows 1, 3, and 5; (3) the true distributions of Y^+ and Y^- are both log-normal, but with respective parameters ($\mu_+ = -1.4428, {\sigma^2}_+ = 0.4869$) and ($\mu_- = -4.1741, {\sigma^2}_- = 0.2435$), corresponding to the *weighted* average of the data in Table 2.4; and (4) the true distributions of Y^+ and Y^- are both multi-modal distributions, with the multi-modal distribution of Y^+ modeled as in (2). Similarly, the multi-modal distribution of Y^- is a combination of two log-normal distributions with respective parameters $({\mu_-}, {\sigma^2}_-) = (-7.0193, 1.6094), (-4.0236, 0.2231)$, corresponding to each set of mean and variance pairs reported in Table 2.4, rows 2, 4, and 6 (the mean and variance values reported in rows 4 and 6 are exactly the same).

Numerical Results

The optimal pooling design for each combination of p_0 and λ is reported in Table A.12; see Appendix A.7. Table 5.2 reports the performance measures for each scenario under the assumption of correct distributions of Y^+ and Y^-. Table 5.2 indicates that when the distributions of Y^+ and Y^- are accurately specified, both **SE**

Table 2.5: Case Study 2: Performance measures of the single-stage estimation procedure and **SE** with continuous test outcomes, $p = 0.12$, $B = \$52.5$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
			$\lambda = 0.25$	$\lambda = 0.5$
0.04	MLE	0.1199 ($\pm$0.0141)	0.1204($\pm$0.0155)	0.1202 ($\pm$0.0150)
	MSE ($\times 10^4$)	2.00 ($\pm$2.86)	2.41 ($\pm$3.48)	2.24 ($\pm$3.11)
	rBias (%)	0.05	0.37	0.16
0.06	MLE	0.1199 ($\pm$0.0142)	0.1205($\pm$0.0156)	0.1204 ($\pm$0.0151)
	MSE ($\times 10^4$)	2.03 ($\pm$2.88)	2.43 ($\pm$3.48)	2.27 ($\pm$3.26)
	rBias (%)	0.06	0.43	0.33
0.18	MLE	0.1199 ($\pm$0.0170)	0.1205($\pm$0.0160)	0.1204 ($\pm$0.0165)
	MSE ($\times 10^4$)	2.89 ($\pm$4.10)	2.56 ($\pm$3.76)	2.71 ($\pm$3.79)
	rBias (%)	0.05	0.40	0.29
0.20	MLE	0.1199 ($\pm$0.0182)	0.1205($\pm$0.0160)	0.1204 ($\pm$0.0169)
	MSE ($\times 10^4$)	3.30 ($\pm$4.71)	2.57 ($\pm$3.74)	2.85 ($\pm$3.79)
	rBias (%)	0.10	0.40	0.35

and the single-stage estimation procedure perform well in terms of MSE and relative bias. Similarly, Table 2.6 reports the performance measures of **SE** and the single-stage estimation procedure under incorrectly specified parameters and distributions for Y^+ or Y^-; see Table A.13 in Appendix A.7 for the corresponding performance measures under incorrectly specified parameters and distributions of Y^+ only. The results in Tables A.11 and A.13 indicate, not surprisingly, that the single-stage procedure and **SE** perform substantially worse when the parameters and distributions of both Y^+ and Y^- deviate from their assumed distributions, as opposed to the case where only the distribution of Y^+ deviates. However, we also see that **SE** continues to be robust in the presence of uncertainty on the distributions of Y^+ and Y^-, yielding much lower MSE and relative bias than the single-stage procedure. In this setting, at the end of the first stage of **SE**, it is highly desirable to revise not only the estimate of p, but also the distributions and parameters of Y^+ and Y^-; this can be conducted, for example, by adapting the estimation procedure to the Bayesian framework, i.e., utilizing the Bayes estimator of p, in lieu of its MLE, and we discuss this as a future research direction (Section 2.4.3). Thus, another advantage of a sequential procedure, such as **SE**, is that it allows for such revision, while a single-stage procedure does not.

Table 2.6: Case Study 2: Performance measures for the single-stage estimation procedure and **SE** with continuous test outcomes, with incorrect parameters for $Y+$ and Y^-, and incorrect distributions for Y^+ and Y^-, $p = 0.12$, $B = \$52.5$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

	$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
				$\lambda = 0.25$	$\lambda = 0.5$
Incorrect Parameters - Y^+ and Y^-	0.18	MLE	0.1508 ($\pm$0.0188)	0.1377 ($\pm$0.0361)	0.1413 ($\pm$0.0265)
		MSE ($\times 10^3$)	1.30 ($\pm$1.28)	1.62 ($\pm$2.90)	1.16 ($\pm$1.81)
		rBias (%)	25.66	14.74	17.78
	0.20	MLE	0.1509 ($\pm$0.0197)	0.1374($\pm$0.0386)	0.1407 ($\pm$0.0255)
		MSE ($\times 10^3$)	1.35 ($\pm$1.34)	1.79 ($\pm$3.21)	1.07 ($\pm$1.71)
		rBias (%)	25.78	14.54	17.21
Incorrect Distributions - Y^+ and Y^-	0.18	MLE	0.0316 ($\pm$0.0185)	0.0692($\pm$0.0492)	0.0845 ($\pm$0.0497)
		MSE ($\times 10^3$)	8.15 ($\pm$3.07)	5.01 ($\pm$3.96)	3.73 ($\pm$3.10)
		rBias (%)	73.64	42.36	29.56
	0.20	MLE	0.0335 ($\pm$0.0182)	0.0699($\pm$0.0496)	0.0818 ($\pm$0.0479)
		MSE ($\times 10^3$)	7.82 ($\pm$2.98)	4.98 ($\pm$4.00)	3.75 ($\pm$3.07)
		rBias (%)	72.11	41.79	31.82

2.4 Discussion

We study the effectiveness of a sequential and adaptive estimation procedure, **SE**, to estimate an unknown prevalence rate using pooled testing. **SE** utilizes continuous test outcomes, while accounting for the dilution effect of pooling and testing errors, and allows the prevalence rate estimate to be revised as testing proceeds, so that the remaining tests can be conducted with a more effective pooling design that is based on a more accurate prevalence rate estimate. The pooling design optimization model, embedded into **SE**, simultaneously optimizes for both pool sizes and number of testing pools under a testing budget constraint so as to minimize the MSE of the prevalence rate MLE.

We conclude our study by summarizing our findings and suggestions for future research. In Sections 2.4.1 and 2.4.2, we discuss the effectiveness of utilizing continuous outcomes, and of the proposed sequential estimation procedure in estimating an unknown prevalence rate in the presence of unreliable and limited information on the infection's characteristics, including the bio-marker dynamics in infected subjects. We then provide insights into the choice of the budget allocation factor, λ, and discuss the trade-offs involved. Finally, in Section 2.4.3, we suggest potential directions for future research regarding the prevalence rate estimation problem with pooled testing.

2.4.1 On the Use of Continuous Test Outcomes

As discussed in our case studies, continuous test outcomes should be utilized, especially for surveillance of new or emerging infections. As Tables 2.1 and 2.2 demonstrate, using continuous test outcomes provides a more robust MLE for the prevalence rate under various assumptions on $p_0^{(1)}$, and various values of p and the testing budget, B. Therefore, utilizing continuous test outcomes is especially beneficial in studying emerging infections, as information about those infections is likely to be highly unreliable.

We next discuss the benefit of utilizing the proposed sequential estimation procedure, in conjunction with continuous test outcomes, to further enhance the estimation efficiency and robustness of the MLE.

2.4.2 On the Proposed Sequential Estimation Procedure

Our numerical study indicates that the proposed sequential and adaptive estimation procedure is especially beneficial in cases where $p_0^{(1)}$ is a poor estimate of p, or the assumed bio-marker distribution or parameters are inaccurate. This finding is of particular importance for surveillance studies of emerging infections, where an initial estimate of p may have significant discrepancy in comparison to the true prevalence rate, p, and/or the bio-marker dynamics may be highly uncertain. In particular, the proposed **SE** with $0 < \lambda < 1$ leads to a lower variability in the estimation efficiency, in terms of both MSE and the relative bias, under different values of $p_0^{(1)}$, i.e., this is independent of whether $p_0^{(1)}$ is an underestimate or an overestimate of p. Furthermore, as shown in Sections 2.3.2 and 2.3.3, the use of **SE** also reduces the negative impact, on estimation efficiency, of incorrectly specifying the distributions and parameters of Y^+, i.e., the bio-marker concentration of an infected subject, and Y^-, i.e., the noise coming from an uninfected subject. In particular, when only the distribution of Y^+ is incorrectly specified, **SE**, with $\lambda = 0.25$, remains robust, yielding comparable estimation efficiency to that under a correctly specified distribution of Y^+ (see Section 2.3.2). On the other hand, when the distributions of both Y^+ and Y^- are incorrectly specified, estimation efficiency of both **SE** and the single-stage procedure is negatively impacted. However, it should be noted that **SE** performs significantly better, in terms of MSE and the relative bias, even in this setting. Therefore, when limited information is available about the current status of, or bio-marker dynamics related to, an infection, **SE** is an attractive estimation strategy, as **SE** mitigates the initial bias, whether optimistic or pessimistic, under uncertainty.

Among the variations of **SE**, the choice of λ, i.e., the budget allocation factor that governs the splitting of the total testing budget B into a stage 1 budget, $B^{(1)} = \lambda B$, and a stage 2 budget, $B^{(2)} = (1 - \lambda)B$, proves to be crucial for the quality of the prevalence estimate. This is because λ represents the trade-off between exploration and exploitation in prevalence rate estimation; a large value of λ places more weight on the stage 1 testing of **SE** (exploration), while a small value of λ places more weight on the stage 2

testing of **SE** (exploitation). Furthermore, the choice of λ is also highly dependent on the characteristics of the infection of interest and the available testing budget. More specifically, the choice of λ is especially important when the budget is tight, as a large value of λ may lead to high bias and MSE, while a small value may lead to an inadequate number of testing pools, resulting in a firsts-stage MSE value that is lower than the actual prevalence rate. Therefore, in order to have an optimal allocation of the testing budget, λ needs to be selected with careful consideration of the testing budget and the characteristics of the infection.

2.4.3 Future Research Directions

The determination of an optimal value of λ, i.e., the budget allocation factor, is one of the potential directions for future research regarding the prevalence rate estimation problem with pooled testing. Another important future research direction is to incorporate Bayesian and regression analyses into the estimation procedure to respectively improve the quality and robustness of the estimate, especially when both the assumed bio-marker and noise distributions are not reliable, and to account for the potential heterogeneity in the population. Additionally, it is important to study strategies that integrate the estimation and classification components, for example, individual follow-up testing for positive-testing pools, performed in practice for diseases such as the HIV, can be incorporated into the estimation procedure, or the proposed estimation procedure can be incorporated into a classification scheme so as to improve the classification accuracy.

Chapter 3

A Methodology for Deriving the Sensitivity of Pooled Testing, based on Viral Load Progression and Pooling Dilution

3.1 Background

Pooled testing, in which biological specimens (e.g., blood, urine, tissue swabs) from multiple subjects are combined into a testing pool and tested via a single test, can substantially improve the efficiency of public health screening and population-level surveillance of diseases; and enables the use of expensive testing technologies, such as the nucleic acid amplification testing (NAT) technology [4]. Ever since its introduction in the 1940's [38], pooled testing has been commonly used for both surveillance and screening purposes, including donated blood screening for transfusion-transmittable infections, e.g., the human immunodeficiency virus (HIV), or regional HIV surveillance [3, 23].

In general, biomarker tests have less than perfect *sensitivity* (true positive probability), mainly because the progression of the load (concentration) of a disease-related biomarker (e.g., the HIV viral RNA, measured by the NAT) in the infected host follows various phases post-exposure, with varying growth rates, e.g., pre ramp-up phase, ramp-up phase with accelerating growth rates, and post ramp-up phase, with the biomarker load tending to a plateau or significantly diminishing due to a resolved infection (e.g., [42, 80]). A majority

of false negative testing errors occur during those earlier phases of the infection (pre ramp-up and early ramp-up phases), also known as the *window period*, the length of which depends on the specific infection and the biomarker being measured by the test (e.g., [105]). Pooled testing may further reduce the test's sensitivity due to *pooling dilution*, where the biomarker load of an infected specimen is diluted by infection-free specimens in the pool so that the infected specimen may no longer be detectable by the pooled test. Pooled testing sensitivity at various pool sizes is an essential input to key decisions in surveillance and screening efforts, including testing pool design. However, clinical data on test sensitivity values for different pool sizes are limited, and the extant literature that analytically derives the sensitivity of a pooled test does so under restrictive assumptions, including that the test is perfectly reliable outside of the window period, i.e., all infected specimens that are outside of the window period are detected with probability 1 regardless of the pool size (e.g., [13,105,106]). There are commonly adopted mathematical models of viral load progression in infected subjects, but these models consider only the window period (e.g., [17]).

Therefore, our objective in this paper is to develop a generic methodology for analytically deriving the sensitivity of pooled testing at various pool sizes, based on models that account for viral load progression and pooling dilution; and by relaxing various restrictive assumptions adopted in the literature. In particular, our methodology integrates the following components within a probabilistic framework: (1) the "doubling time" model [17], which we expand to model the host's viral load progression throughout the infection's life-time; and (2) the probit function [105], which we expand to model pooling dilution to consider the number of infected specimens in a pool. The proposed methodology derives the *conditional test sensitivity*, conditioned on the number of infected specimens in a pool; and uses the law of total probability to derive overall (unconditional) test sensitivity values for a wide range of pool sizes. We validate this methodology via published test sensitivity data and show that it is highly accurate. This methodology utilizes higher dimensional integrals, which may be computationally expensive for large pools. As a result, we also propose an easy-to-compute, and a highly accurate, approximation function that is based on establishing a functional relationship between the sensitivity of pooled testing and the number of infected specimens in a pool. Our methodology can be used to provide important inputs for surveillance and screening activities, including testing pool design, which has received considerable attention in the literature, (e.g., [68, 69, 75, 102, 104, 111, 112]). Further, our methodology is generic, and can be calibrated for various infections; and we demonstrate its application for the HIV and HIV ULTRIO Plus NAT Assay. For this purpose, we calibrate model parameters using published test efficacy data for the HIV ULTRIO Plus Assay [80, 91], and clinical data on viral RNA load progression in HIV-infected patients [17,105]; and use this methodology to derive and validate the sensitivity of the HIV ULTRIO Plus Assay for various pool sizes. We also demonstrate the value of this methodology through optimal testing pool design for HIV prevalence estimation in Sub-Saharan Africa. This

case study indicates that the optimal testing pool design is highly efficient, and outperforms a benchmark pool design.

The remainder of the paper is organized as follows. In Section 3.2, we present our methodology for deriving the sensitivity of pooled testing. In Section 3.3, we apply this methodology to testing pool design for prevalence estimation of HIV in Sub-Saharan Africa. Finally, in Section 3.4, we summarize our findings and provide a discussion.

3.2 Methods

Our methodology is based on the integration of viral load progression and pooling dilution models. In Section 3.2.1, we discuss each component, and their calibration, in detail. Then in Section 3.2.2, we provide an easy-to-compute approximation function for pooled sensitivity estimation. A summary of all the notation used in our study is provided in the Appendix.

3.2.1 Pooled Sensitivity Estimation Methodology

Viral Load Progression Model

We first describe the viral load progression model, which expands the widely adopted doubling time model proposed by Busch et al. (2000) [17]. The original doubling time model [17] considers viral load progression only through the window period of an infection, and we expand it to model the infection's life-time. This is needed to relax a common assumption used in test sensitivity calculations, that all infected specimens outside of their window period are detected with probability 1, regardless of the pool size (e.g., [105, 106]). According to numerous studies, the viral load in infected subjects progresses through various phases of growth rates post-exposure: pre ramp-up phase, ramp-up phase with accelerating growth rates, and post ramp-up phase during which growth rate slows down, eventually reaching a plateau or resolution of the infection (e.g., [14, 17, 42, 47, 80, 105, 106]). To model this phenomenon, we let t_w, t_p, and t_s respectively denote the time at which the window period ends, viral load peaks, and viral load reaches steady state; and let $VL(t)$ denote the infected subject's viral load at time t post-exposure. Based on clinical data for HIV and hepatitis B and C infections [14,17,42,47], we model the infected subject's viral load beyond the window period and up to the steady state as follows:

For $t_w \leq t \leq t_s$:

$$VL(t) = VL(t_w) + \frac{C_w}{t} \exp\left(-\frac{(ln(t - t_w) - a)^2}{b}\right) ,$$

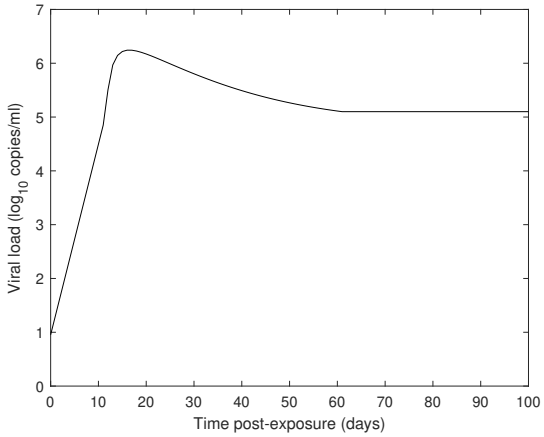
where C_w, a, and b are infection-specific calibration parameters. In this study, we assume that the viral load

reaches steady state at time t_s, beyond which it remains constant at a level of $VL(t_s)$ (i.e., $VL(t) = VL(t_s)$, $\forall t > t_s$); this assumption can be easily relaxed. We note that the steady state viral load, denoted by $VL(t_s)$, can equal zero for acute infections, or can remain at some positive level for chronic infections. Consequently, the complete viral load model follows:

$$VL(t) = \begin{cases} C_0 2^{t/\lambda}, & \text{if } t \leq t_w \\ VL(t_w) + \frac{C_w}{t} \exp\left(-\frac{(ln(t-t_w)-a)^2}{b}\right), & \text{if } t_w < t \leq t_s \\ VL(t_s), & \text{if } t > t_s, \end{cases} \cdot \tag{3.1}$$

where the window period component, $C_0 2^{t/\lambda}$, is the doubling time model in Busch et al. [17], with infection-specific calibration parameters C_0 and λ, where λ represents the viral load doubling time within the window period. For demonstration, Figure 3.1 plots the base 10 logarithm of the HIV viral RNA load, obtained by Eqn. (3.1), versus post-exposure time in HIV-infected subjects, calibrated as discussed in Section 3.2.1.

Figure 3.1: HIV viral RNA load progression spanning the infection's life-time, covering the window period, peak viremia phase, and chronic phase (based on the data in Table 3.1).



Pooling Dilution Model

Pooled testing may reduce the test's sensitivity due to pooling dilution, that is, the biomarker load of an infected specimen is diluted by infection-free specimens in the pool so that the infected specimen may no longer be detectable by the pooled test [91]. In this section, we model the test sensitivity considering pooling dilution. For this purpose, we first describe the probit function, proposed in the literature to model pooling dilution, and discuss how it is expanded to consider the number of infected specimens in a pool.

Towards this end, let $\tau \in \Re^+$ denote the life-time of a certain infection, and $n \in Z^+$ denote the testing

pool size. We also let $T^+(n)$ denote the event that the test outcome is positive for a pool of size n, indicating the presence of at least one infected specimen in the pool; and let $N_I(n)$ denote the number of infected specimens in the pool, which is a random variable with possible values $\{1, \cdots, n\}$. Therefore, the test sensitivity for pool size n ($Sens(n)$), that is, the probability that the test outcome is positive given that the pool contains at least one infected specimen, follows:

$$Sens(n) = P(T^+(n); N_I(n) \geq 1), \tag{3.2}$$

where the ";" notation denotes probabilistic conditioning. Weusten et al. (2002, 2011) [105, 106] propose the following probit model to derive the sensitivity of pooled testing, under the assumptions that the test is perfectly reliable outside of the infection's window period (i.e., $\tau = t_w$), and the pool contains at most one infected specimen, regardless of the pool size (i.e., $N_I(n) = 1$ with probability 1):

$$Sens(n) = \frac{1}{t_w} \int_0^{t_w} \Phi\left(z \frac{log\left(\frac{\chi C_0 2^{t/\lambda}}{n x_{50}}\right)}{log(x_{95}/x_{50})} \right) dt, \quad \text{(from [106])} \tag{3.3}$$

where following [106], $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution; z is a constant such that $\Phi(z) = 0.95$, i.e., $z = 1.6449$; χ is the number of nucleic acid copies per viral particle, and x_{50} and x_{95} respectively denote the viral load measurement at which the probability of a pool testing positive is 50% and 95% [106].

We expand the probit model in Eqn. (3.3) to consider the performance of a pooled test during the infection's life-time, and to account for the possibility of multiple infected specimens in a testing pool. In particular, we first derive the test's *conditional sensitivity* for a pool size of n, given that the pool contains i infected specimens, denoted by $Sens(n; i) = P(T^+(n); N_I(n) = i)$, $\forall i \in \{1, \cdots, n\}, n \in \mathbb{Z}^+$:

$$Sens(n; i) = P(T^+(n); N_I(n) = i) = \frac{1}{\tau^i} \underbrace{\int_0^{\tau} \int_0^{\tau} \cdots \int_0^{\tau}}_{i\text{-fold}} \Phi\left(z \frac{log\left(\frac{\chi \sum_{j=1}^{i} (VL(t_j)}{n x_{50}}\right)}{log(x_{95}/x_{50})} \right) dt_1 dt_2 \cdots dt_i, \tag{3.4}$$

where t_j denotes the (random) post-exposure time for infected specimen $j, j = 1 \cdots, i$, in the pool, and $VL(t_j)$ can be derived from Eqn. (3.1). Observe that the probit model in Eqn. (3.3) follows as a special case of Eqn. (3.4), with $\tau = t_w$ and $N_I(n) = 1$. Then, using the common binomial model for the number of infected specimens in a pool, and the law of total probability, the overall sensitivity of the pooled test, for pool size of n and infection prevalence rate of p, follows:

$$Sens(n) = \sum_{i=1}^{n} Sens(n;i)P(N_I(n) = i) = \sum_{i=1}^{n} Sens(n;i)\binom{n}{i}p^i(1-p)^{n-i}. \tag{3.5}$$

On the other hand, the test's *specificity* (true negative probability), given by:

$$Spec = 1 - P(T^+(n); N_I(n) = 0),\ \forall n \in \mathbb{Z}^+,$$

is independent of the pool size, because in the absence of infected specimens in the pool ($N_I(n) = 0$), pooling dilution does not apply.

In summary, the proposed *sensitivity estimation model* in Eqs. (3.4)–(3.5) can be used in conjunction with Eqn. (3.1) to determine the sensitivity of pooled testing for any pool size.

Calibration and Validation

Calibration: We calibrate the sensitivity estimation model based on Stramer et al. (2013) [91], which provides the test sensitivity of an infected window period blood specimen diluted 16-fold (i.e., tested within a pool of size 16) as 88%. Therefore, C_0 is calibrated such that Eqn. (3.3) equals 0.88 with $n = 16$. Further, according to various studies, the HIV viral RNA load in blood peaks typically around day 17, with an average load of 6.8 $\log_{10}$ copies/ml, and reaches steady state around day 61, with an average load of 5.1 $\log_{10}$ copies/ml; the HIV doubling time (λ) is 0.85 days, and the number of nucleic acid copies per viral particle (χ) for HIV is 2 [42,80,106]. Therefore, we calibrate the remaining parameters of our model, namely C_w, a, and b, in Eqn. (3.1), based on these values; see Table 3.1 for the clinical data used and the calibrated parameter values. We note that this calibration is for demonstration purposes, and our model parameters can be calibrated for any given set of data.

Validation: We validate our sensitivity estimation model using the overall (life-time) efficacy data for the HIV ULTRIO Plus NAT Assay, in terms of the 95% confidence interval (CI), published by the Food and Drug Administration (FDA); see Table 3.1. We use our model (Eqs. (3.1), (3.4), and (3.5)), with calibrated parameters reported in Table 3.1, to derive the conditional sensitivity values for the HIV ULTRIO Plus Assay for various pool sizes; see Table 3.2, which reports the derived conditional test sensitivity values as a function of the pool size and the number of infected specimens in a pool. According to Table 3.2, both $Sens(n = 1; N_I(1) = 1) = 99.98\%$ and $Sens(n = 16; N_I(16) = 1) = 99.26\%$ values are contained within the 95% confidence intervals reported by the FDA (see Table 3.1).

As discussed above, the overall test sensitivity at any prevalence rate, p, can then be derived from the conditional sensitivity values in Table 3.2 via the law of total probability; see Eqn. (3.5). As expected,

Table 3.1: Calibration and validation data for the HIV and HIV ULTRIO Plus NAT Assay.

Calibration Data	
HIV Viral RNA Load Data	
t_w	11 days [80]
t_p	17 days [80]
t_s	61 days [80]
$VL(t_p)$	6.8 $\log_{10}$ copies/ml [80]
$VL(t_s)$	5.1 $\log_{10}$ copies/ml [80]
λ	0.85 days [42]
χ	2 copies/particle [106]
Test Sensitivity Data	
$P(T^+(16); N_I(16) = 1, \tau = t_w)$	0.88 [91]
Calibrated Model Parameters	
C_0	9.000
C_w	1.096×10^8
a	1.980
b	1.730
Validation Data	
$Sens(n = 1; N_I(1) = 1)$	(99.7%-100%) [43]
$Sens(n = 16; N_I(16) = 1)$	(98.2%-99.5%)) [43]

Table 3.2: Derived conditional sensitivity values for the HIV ULTRIO Plus Assay (in %) as a function of the pool size and the number of infected specimens in a pool ($Sens(n; i)$) (reported in 9 decimal point accuracy)

Number of infected specimens (i) \ Pool size (n)	1	2	3	4	5	6	7	8
1	99.98	99.93	99.88	99.82	99.76	99.70	99.65	99.59
2		99.9998	99.9995	99.9992	99.9988	99.9983	99.9979	99.9974
3			100.00000	99.99999	99.99999	99.99999	99.99998	99.99998
4				99.9999999	99.9999999	99.9999999	99.9999998	99.9999997
6					99.999999999	99.999999998	99.999999997	99.999999996
7						100.000000000	100.000000000	100.000000000
8							100.000000000	100.000000000

Number of Infected Specimens (i) \ Pool size (n)	9	10	11	12	13	14	15	16
1	99.54	99.50	99.45	99.41	99.37	99.33	99.29	99.26
2	99.9969	99.9963	99.9958	99.9953	99.9948	99.9942	99.9937	99.9932
3	99.99997	99.99996	99.99996	99.99995	99.99994	99.99994	99.99993	99.99992
4	99.9999997	99.9999996	99.9999995	99.9999994	99.9999993	99.9999991	99.9999990	99.9999989
5	99.999999995	99.999999994	99.999999992	99.999999991	99.999999989	99.999999987	99.999999985	99.999999983
6	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
7	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
8	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
9	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
10		100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
11			100.000000000	100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
12				100.000000000	100.000000000	100.000000000	100.000000000	100.000000000
13					100.000000000	100.000000000	100.000000000	100.000000000
14						100.000000000	100.000000000	100.000000000
15							100.000000000	100.000000000
16								100.000000000

conditional test sensitivity decreases with pool size, and increases with the number of infected specimens in a pool. Moreover, we observe that test sensitivity rapidly approaches 1 as $N_I(n)$, the number of infected

specimens in a pool of size n, increases, and for $N_I(n) \geq 4$, test sensitivity becomes almost perfect.

We also derive $P(T^+(n); N_I(n) = 0) = 0.07\% = 1 - Spec$, i.e., $Spec = 99.93\%$, which is also consistent with the efficacy data for the HIV ULTRIO Plus NAT Assay, published by the FDA [43].

3.2.2 An Approximation for Sensitivity Estimation

Our model in Section 3.2.1 derives the conditional test sensitivity values, $Sens(n; i)$, and uses the law of total probability, along with higher dimensional integrals (up to pool size), to derive the overall (unconditional) test sensitivity values for a wide range of pool sizes. Thus, it can be computationally expensive, especially for large pool sizes. Therefore, in this section, we provide an approximation function for computing the pooled test sensitivity,which does not require higher dimensional integrals. We do this by fitting a function to the sensitivity data derived in Table 3.2 via linear regression so as to minimize the mean squared error (MSE) of the proposed approximation.

Consider the following functional form for conditional test sensitivity for pool size n, given i infected specimens in a pool:

$$\widetilde{Sens}(n; i) = 1 - \beta\alpha^{\left(\frac{i}{n^\gamma}\right)}, \quad i \in \{0, 1, \cdots, n\}, n \in \mathbb{Z}^+, \tag{3.6}$$

where α, β, and γ are calibration parameters. In particular, by definition of pooling dilution, the probability of detection reduces with pool size, implying that $\gamma \geq 0$ and $\alpha \in [0, 1]$; and $P(T^+(n); N_I(n) = 0) = 1 - Spec$ (see Section 3.2.1), implying that $\beta = Spec$. The remaining parameters (i.e., α and γ) are derived so as to minimize the MSE between the fitted function and the data in Table 3.2, that is:

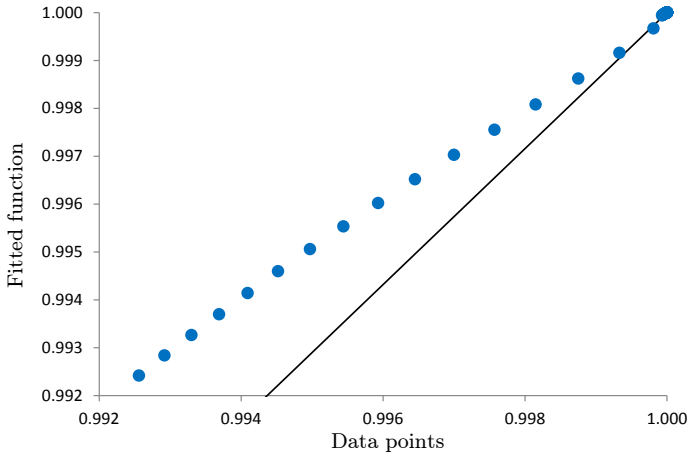
$$(\alpha^*, \gamma^*) = \underset{\alpha,\gamma}{\arg\min} \left(\sum_{n=1}^{16} \sum_{i=0}^{n} \left[Sens(n; i) - \widetilde{Sens}(n; i) \right]^2 \right),$$

This minimization problem is a non-convex optimization problem, which we solve numerically in Python for the HIV ULTRIO Plus Assay, obtaining $(\alpha^* = 0.00033, \gamma^* = 0.179)$. The goodness of fit, measured by the coefficient of determination (i.e., R^2), is equal to 0.9995, suggesting that the fit is highly accurate; see Figure 3.2 for the fitted model versus the data points in Table 3.2.

3.3 Results

We apply our sensitivity estimation models (both exact and approximation models, respectively detailed in Sections 3.2.1 and 3.2.2) to determine an optimal testing pool design for HIV prevalence estimation in Sub-Saharan Africa. Specifically, we use the methodologies proposed in Sections 3.2.1 and 3.2.2, along with

Figure 3.2: Fitted function versus the data points in Table 3.2



the calibrated parameters in Section 3.2.1, to derive sensitivity estimates for the HIV ULTRIO Plus Assay for various pool sizes; and use these sensitivity values as inputs to a testing pool design optimization model, studied in the literature [69, 76, 100, 111].

3.3.1 Testing Pool Design Optimization

The optimization model determines an optimal testing pool design for prevalence estimation, in terms of the number of testing pools to be utilized, s, and the size of each testing pool, n, under a testing budget constraint, so as to minimize the *asymptotic variance* of the *maximum likelihood estimator* (MLE) of the unknown prevalence rate [76]:

$$\begin{aligned} \underset{n,s}{\text{minimize}} \quad & \sigma^2(n, s; p_0) \\ \text{subject to} \quad & c_f \; s + c_v \; sn \leq B \\ & n \leq \overline{N} \\ & n, s \in \mathbb{Z}^+, \end{aligned} \tag{3.7}$$

where $\sigma^2(n, s; p_0)$ denotes the asymptotic variance of the MLE for a pool design (n, s), given an initial estimate of the unknown prevalence rate p, which we denote by p_0. The testing cost consists of a fixed testing cost per pool (e.g., cost of the testing kit), denoted by c_f, and a collection cost per specimen (e.g., cost of drawing blood), denoted by c_v. The tester has a total testing budget of B for prevalence estimation. Additionally, the maximum pool size that can be used may be restricted due, for example, to technological constraints, regulations, or other considerations, and we denote the maximum allowable pool size by $\overline{N}$.

The asymptotic variance is a commonly used criterion for optimal testing design in prevalence estimation and for evaluation of estimators in statistical inference, and is also related to the Fisher's information (e.g., [58, 69, 98, 100, 102, 111, 112]).

In pooled testing, only one test is used on each pool, and the test provides a binary outcome, with a positive outcome indicating the presence of at least one infected specimen in the pool; and a negative outcome indicating that all specimens in the pool are infection-free. Using the test outcomes, the tester derives the MLE of the unknown prevalence rate ($\hat{p}$). In particular, for a given testing design, (n, s), let $S_I(s)$ denote the number of positive-testing pools among s pools, which is a random variable prior to testing. Then, after the testing is conducted and a realization of $S_I(s) = k$ is observed, the MLE of the prevalence rate corresponds to the value of p that maximizes the following likelihood function:

$$\begin{aligned} L(p; S_I(s) = k) = &\binom{s}{k} \Big[Sens(n;p) - (1-p)^n (Sens(n;p) + Spec - 1)\Big]^k \\ &\times \Big[1 - Sens(n;p) + (1-p)^n (Sens(n;p) + Spec - 1)\Big]^{s-k} \\ \Rightarrow \hat{p} \equiv &\underset{p \in (0,1)}{\operatorname{argmax}} \Big\{ L(p; S_I(s) = k) \Big\}. \end{aligned} \tag{3.8}$$

The asymptotic variance function, $\sigma^2(n, s; p)$, for a pool design of (n, s), and with respect to the unknown prevalence rate, p, is then given by (e.g., [69]):

$$\sigma^2(n, s; p) = \frac{\{Sens(n;p) - (1-p)^n (Sens(n;p) + Spec - 1)\} \{1 - Sens(n;p) + (1-p)^n (Sens(n;p) + Spec - 1)\}}{sn^2 (1-p)^{2(n-1)} (Sens(n;p) + Spec - 1)^2}. \tag{3.9}$$

3.3.2 Study Design and Data

Our goal is in this section is to demonstrate the value of the sensitivity estimation methodologies developed in this paper through a numerical study. We do this by designing an optimal testing pool, based on the sensitivity estimates derived for the HIV ULTRIO Plus Assay for various pool sizes using the methodologies described in Section 3.2; and comparing the efficiency of the *optimal testing design* with a *benchmark design* that does not consider pooling dilution (hence does not need to use our methodology for sensitivity estimation at various pool sizes). As discussed above, we consider pool design for prevalence estimation of HIV in Sub-Saharan Africa using the HIV ULTRIO Plus Assay.

Model parameters are as follows. We assume that the actual prevalence rate is $p = 0.044$ [108], which is unknown to the tester; this prevalence rate is representative of the HIV prevalence rate in Sub-Saharan Africa. In the absence of this information, the tester determines an initial estimate of $p_0 = 0.022$, i.e.,

we consider the case of undershooting. Based on published data, we consider a fixed testing cost per pool of \$31.5 [63], a collection cost per specimen of \$8 [33], and a total testing budget of \$5,575 [75], which corresponds to a testing budget of 50 pools, each of size 10. Finally, we consider a maximum allowable pool size, of $\overline{N} = 48$ [86]. These parameter values are for demonstration purposes, and one can conduct similar analyses with different parameter values.

As sensitivity inputs, we utilize the sensitivity values in Table 3.2, which are derived by the sensitivity estimation model in Section 3.2.1, in conjunction with the calibration parameters in Section 3.2.1. The sensitivity values in Table 3.2 correspond to pool sizes of $n = \{1, 2, \cdots, 16\}$. As discussed above, the sensitivity estimation model in Section 3.2.1 requires the computation of higher dimensional integrals (up to pool size), and can be computationally expensive. Therefore, we use the approximation in Section 3.2.2 to derive the sensitivity values for the remaining pool sizes, i.e., $n = \{17, \cdots, 48\}$. Then, we perform a two-dimensional search, over all possible values of $\{(n, s) : n \in \{1, \cdots, 48\}, c_f\ s + c_v\ sn \leq B\}$, to determine the optimal testing pool design, i.e., (n^*, s^*), for the optimization model in Eqn. (3.7) that minimizes the asymptotic variance. To determine the "best" benchmark design, we repeat the two-dimensional search, but without considering pooling dilution, that is, by replacing the parameters, $Sens(n)$, $\forall n \in \mathbb{Z}^+$, with 99.98%, i.e., the sensitivity of individual testing for the HIV ULTRIO Plus Assay; see Table 3.3 for the resulting optimal design and the benchmark design. For each of these designs, we perform a Monte Carlo simulation to derive estimates for the MLE of p, $\hat{p}$ (see Eqn.(3.8)); mean squared error (MSE); and the relative bias (rBias (%)), given by:

$$MSE = (\hat{p} - p)^2, \text{ and } rBias(\%) = 100 \times \left| \frac{\hat{p} - p}{p} \right|. \tag{3.10}$$

These performance metrics relate to the efficiency of prevalence estimation, and are commonly used in the literature, e.g., [57, 58, 111].

In particular, for each testing design, we perform 10,000 simulation replications. In each replication, we randomly generate the infection status of each of the $n^* \times s^*$ specimens, where each specimen carries an infection with probability p; and is infection-free otherwise; and for each infected specimen, we randomly generate a post-exposure time from a Uniform distribution with support $[0, \tau]$, and compute the viral load using Eqn. (3.1) and the parameters of Section 3.2.1. Then, we randomly assign the specimens into s^* pools, each of size n^*, and generate the binary test outcomes based on the test sensitivity model given in Eqn. (3.4). Finally, we compute the MLE, MSE, and rBias for each replication using Eqs. (3.8) and (3.10).

3.3.3 Numerical Study Results

Table 3.3 reports the average estimation efficiency of the optimal design and the benchmark design, over 10,000 simulation replications. All performance metrics are reported in the form of *mean ± half-width of 95% confidence interval (CI).*

Table 3.3: Estimation efficiency (*mean ± half-width of 95% CI*) of the optimal design and the benchmark design for HIV prevalence estimation (with an actual prevalence rate of $p = 0.044$).

Performance Metric	**Optimal Design**	**Benchmark Design**
Pool design	$n^* = 37, s^* = 17$	$n^* = 17, s^* = 33$
$\hat{p}$ (MLE)	0.05204 ± 0.00036	0.03041 ± 0.00029
MSE ($\times 10^4$)	3.95 ± 0.11	4.00 ± 0.08
rBias (%)	18.26 ± 0.52	30.88 ± 0.48

As indicated by Table 3.3, the optimal design outperforms the benchmark design, and the differences are statistically significant. The benchmark design yields especially high bias in comparison to the optimal pool design, mainly due to the assumption of no pooling dilution, leading to biased estimates of the unknown prevalence rate.

3.4 Discussion

Pooled testing is commonly used in public health settings, for both screening and surveillance of diseases and infections. An accurate and tractable method to compute the sensitivity of a pooled test is extremely important in designing the optimal pooled testing scheme for these efforts. As pooled NAT assays are widely used to screen for diseases, several approaches are proposed in the literature to compute the sensitivity of pooled NAT assays. However, these approaches only account for the window period of the infection, and assume perfect sensitivity past the window period, which is a restrictive assumption, especially as pooling dilution plays an important role in the sensitivity of pooled tests. Further, these studies compute the sensitivity of the pooled test based on the assumption of having at most one infected specimen in any testing pool, when the probability of having multiple infected specimens in a pool is, in fact, a function of both the pool size and the prevalence of the disease.

In this paper, we relax the restricting assumptions in the aforementioned studies and propose both exact and approximate models for computing the sensitivity of a pooled test. We expand the doubling time viral load model [17] to mathematically model the various growth phases of an infection; and propose an exact method to compute the conditional sensitivity of a pooled test as a function of the number of infected specimens in the pool and the pool size, by expanding the probit model in [105, 106]. Then, we can use a binomial model for the number of infected specimens in a pool, along with the law of total probability, to

calculate the overall sensitivity of the pooled test given the pool size and the prevalence rate of the disease. We calibrate and validate our exact model using published data on the HIV ULTRIO Plus Assay. Finally, we propose an alternative approximation model to derive the sensitivity of pooled testing that is highly accurate and more analytically tractable than the exact method. We demonstrate the value of our exact and approximate models of pooled testing sensitivity in a case study on HIV prevalence estimation. In particular, we incorporate the proposed models into our testing pool design procedure for prevalence estimation of HIV in Sub-Saharan Africa. Our results show that the sensitivity model is very accurate for the HIV ULTRIO Plus Assay, enabling the design procedure to yield efficient testing pool designs that significantly minimize the estimation error, in comparison to a pool design procedure that utilizes less accurate sensitivity values (i.e., assuming no pooling dilution).

In summary, we develop exact and approximate models for computing the sensitivity of a pooled test by expanding upon the commonly used probit model in [105,106], and relaxing various restricting assumptions, as we previously discuss. Our methodologies are computationally tractable and highly accurate, and can significantly improve the efficiency of testing pool design for prevalence estimation, as demonstrated by our case study, and for public health screening. We further note that the proposed sensitivity estimation methodology is not infection-specific, and can be calibrated with clinical and published data for any infection or disease, e.g., hepatitis B and C viruses. In addition to its application in prevalence estimation, this methodology can be used in conjunction with other optimization models to make optimal decisions for classification efforts (e.g., [5]), and can also be used for setting a classification threshold, i.e., for classifying a subject as infected versus infection-free for the disease in question. Further, as the expanded viral load model considers the life-time of the infection, in regard to the biomarker load in infected subjects, it allows for more precise sensitivity estimation if information is available about the population of interest, e.g., repeat blood donors have lower overall HIV prevalence rates, and, due to their donation history, one can infer which stage of the infection the donor would be in, if infected. Therefore, integrating the sensitivity estimation methodology proposed in this paper with such optimization models would be worthwhile extensions of this research.

Chapter 4

Optimal Pooled Testing Design for Prevalence Estimation

4.1 Introduction

Surveillance is essential for responding to emerging or seasonal diseases, and for assessing the performance of preventative measures and healthcare services for infectious diseases in general. Four basic components of surveillance activities include, collection, analysis, dissemination, and response [35]. Among these components, collection and analysis activities are conducted frequently at many levels of the healthcare system, including local, state, federal, and international levels, by both public and private agencies, with a main goal of *estimating the prevalence rate* of the disease in question [35], which is the focus of this paper. An accurate estimation of the disease prevalence rate is important, as it is a key input to various other surveillance activities, including outbreak detection, response and prevention evaluation, and prediction of the impact of various healthcare services (e.g., [35, 39, 104]). For example, in 2016, following the outbreak of Zika, the Centers for Disease Control and Prevention provided funding to 21 jurisdictions for the surveillance of, and response to, Zika [21]. In general, the testing budget available for prevalence estimation is very small relative to the needs [35]. As a result, prevalence estimation via individually testing of each subject is either infeasible (e.g., [99, 100]), or highly *inefficient* in that it leads to small sample sizes, and hence to potentially inaccurate estimates [54, 99].

A solution to both prevalence estimation and subject identification (i.e., identification of all infected subjects) under limited resources came from Dorfman in the 1940's [38]. Dorfman's idea was to use *pooled testing*, by combining specimens (e.g., blood, urine, tissue swabs) from multiple subjects in a single testing

pool and testing the pool via a single test [38]. Over the years, pooled testing has been extensively studied and shown to be a highly efficient approach under limited resources for both prevalence estimation and subject identification problems, and today it is a widely used testing method for both purposes (e.g., [39,56, 85,95,104]). When pooled testing is used for the purpose of prevalence estimation, it often involves testing of the pools only (i.e., without any follow-up testing on individual subjects in positive-testing pools), as the ultimate goal is to derive an accurate estimate of the disease prevalence rate (e.g., [50, 56, 69, 72]). This is especially true when the goal is to estimate the prevalence of "souces" of vector-borne viral or bacterial diseases, e.g., mosquitoes carrying Zika virus or West Nile virus [83], or romaine lettuce carrying *E. coli* bacteria [22]. In general, the test measures the pool's concentration of a certain bio-marker that serves as an indicator for the presence of the virus or bacteria of interest, and provides a binary outcome: *positive*, indicating the presence of at least one infected specimen in the pool, and *negative* otherwise; and inference on the unknown prevalence rate is made based on the collected testing data.

The efficiency and effectiveness of the estimation process depends critically on the testing *pool design*, which involves determining the *number of pools* to test, and the *pool size* (i.e., the number of specimens to combine in each pool) (e.g., [25,75,92,96]). This is a challenging problem, because there is often limited, and highly uncertain, information on the status and dynamics of a disease prior to a surveillance study, especially for emerging or seasonal diseases, but an initial estimate on the disease prevalence rate is an important input to pool design. Further, research that develops optimal pool designs for prevalence estimation is quite limited. The majority of the relevant literature focuses on the "estimation" component, i.e., derivation of an efficient prevalence estimate from testing data, for a *given* pool design.This stream of research includes studies that investigate the characteristics of the widely used maximum likelihood estimator (MLE) of the prevalence rate in various settings (e.g., [26,27,96]), and that develop various approaches for bias reduction in the MLE (e.g., [16, 44, 54]), as well as studies that investigate alternative methods for deriving an estimator, such as Bayesian analyses (e.g., [40, 81, 94]), or regression analyses (e.g., [28, 41, 55, 102, 110]). On the other hand, only a few studies discuss the pool design component, and mainly in the context of pool size determination for a *fixed (exogenous)* number of testing pools, under perfect tests (e.g., [57, 58, 89]), and imperfect tests (e.g., [45, 69, 99, 100, 112]). However, the aforementioned studies on pool size determination are mainly numerical in nature, and the optimal pool size is not fully characterized in an analytical manner. [69] is a notable exception, and derives various properties of the asymptotic variance function, which is a commonly used objective for pool design. In this paper, we expand upon the properties developed in [69], as we discussed subsequently. It is shown that a pool design that relies highly on an initial point estimate of the prevalence rate, or that corresponds to an exogenously fixed number of testing pools, can result in highly inaccurate estimates of the prevalence rate [57,58,75].

Motivated by these gaps in the literature, in this paper we propose two novel models for testing pool design under uncertainty and limited resources. Our models are flexible, in that they can incorporate both a fixed number and a variable number of testing pools into pool design, and they explicitly accounts for the limited testing budget. Specifically, we study the pool design problem in two settings: the setting where the number of testing pools is fixed exogenously and cannot be altered (*single-variable pool design problem*), and the setting where the tester has control over both the number of pools and the pool size (*joint pool design problem*). The first setting applies, for example, when there is a limited number of testing kits, which can be an important constraint for testing of certain diseases (e.g., [69]), and the tester can only control the pool size. This first setting is important in its own right, as it provides a contribution to the literature in that almost all existing studies on pool design for prevalence estimation focus on this first setting, as discussed above. On the other hand, not surprisingly, our study of the joint pool design problem indicates that a joint optimization of both the pool size and the number of pools can provide substantial benefit, providing a more accurate estimate at the same testing budget. We establish important structural properties of optimal pool designs, which allow us to analytically characterize the form of an optimal pool design and obtain an optimal pool design in a highly efficient manner in each setting. We complement our analytical results with a case study on prevalence estimation of West Nile virus in mosquitoes, which illustrates that the use of joint pool design optimization can overcome the inaccuracies in input parameters, improving estimation accuracy without requiring an increase in the testing budget.

The remainder of this paper is organized as follows. In Section 4.2, we present the notation and modeling assumptions, and formulate the pool design optimization models. In Section 4.3, we establish key structural properties of the objective function in each pool design optimization model, which allow us to analytically characterize optimal pool designs and efficiently solve the optimization models in Section 4.4. We then demonstrate the benefits of utilizing the proposed pool design optimization models with a case study of West Nile virus prevalence estimation in Section 4.5. Finally, we conclude in Section 4.6 with a discussion of our findings and suggestions for future research. To facilitate the presentation, all mathematical proofs are relegated to the Appendix.

4.2 Notation, Assumptions, and Models

In Section 4.2.1, we introduce the notation and discuss the modeling assumptions. Then in Section 4.2.2, we present the pool design models that we develop. A summary of the notation can be found in Appendix C.1.

4.2.1 Notation and Assumptions

Throughout, we denote random variables in upper–case letters and their realization in lower–case letters.

Consider a disease with an unknown prevalence rate, P, which needs to be estimated. Depending on the setting of the *pool design problem (PD)*, the tester needs to determine both the pool size, m, and the number of pools to test, n (*joint pool design problem (PD-J)*); or only the pool size, m, for a *given* number of pools (*single-variable pool design problem (PD-S)*), under a limited testing budget, so as to obtain an accurate estimate of the unknown prevalence rate, i.e., to minimize the asymptotic variance of the estimator, as we discuss below. Testing incurs a fixed testing cost (e.g., cost of the testing kit), of c_f per pool, and a variable cost (e.g., collection cost), of c_v per specimen tested, with $c_f > c_v$, and the tester has a testing budget of B. Then, in *PD-J*, the feasible set for decision variables m and n is given by, $\mathbb{F}(m,n) \equiv \{m,n \in \mathbb{Z}^+ : c_f n + c_v mn \leq B\}$, while in *PD-S*, the feasible set for the single decision variable m, for a *given* value of $n \in Z^+$, is given by, $\mathbb{F}(m;n) \equiv \left\{m \in \mathbb{Z}^+ : m \leq \overline{M}^S(n) \equiv \left\lfloor \frac{B-c_f n}{c_v n} \right\rfloor \right\}$.

The objective is to design testing pools so as to minimize the asymptotic variance of the estimator, which is a commonly used objective for pool design (e.g., [58, 69, 98, 100, 102]), and is also commonly used for assessing the efficiency of an estimator (e.g., [66, 93]). The asymptotic variance, $\sigma^2(m,n;p)$, corresponding to a *true* prevalence rate p, represents the limiting behavior of the mean squared error (MSE) (i.e., variance plus bias square) of an estimator $\hat{P}$, $MSE(\hat{P},m,n;p)$, as the number of pools, n, becomes large, and is commonly used because, in general, the MSE is analytically intractable (e.g., [58, 59, 69]). More specifically, it has been shown that $\lim_{n\to\infty} MSE(\hat{P},m,n;p) \to \sigma^2(m,n;p)$; hence, for a sufficiently large number of pools, the pool size that minimizes the MSE of the prevalence rate estimator converges to the pool size that minimizes the asymptotic variance [58]. The asymptotic variance also provides a strong lower bound (i.e., through the Cramer-Rao lower bound) on the Fisher's information obtained from the prevalence estimate (e.g., [18]), and hence is also utilized in the pool design literature when the objective is to maximize the Fisher's information, because obtaining an explicit analytical expression for Fisher's information often proves to be intractable (e.g., [57, 104]). Following the common terminology, we refer to a pool design as *efficient* if it minimizes the asymptotic variance (e.g., [57, 58, 104]).

On the testing side, we consider a test that can be applied to pools of specimens collected from subjects (i.e., $m \geq 2$) as well as to individual specimens (i.e., $m = 1$). We assume that the test is perfectly reliable and provides a binary outcome, that is, the test has perfect *sensitivity* (true positive probability) and *specificity* (true negative probability), thus providing a *positive* outcome only if there is at least one true-positive specimen in the pool, and a *negative* outcome only if all specimens in the pool are true-negative.

We estimate the unknown prevalence rate via the commonly adopted maximum likelihood estimator

(MLE) (e.g., [58]), given by:

$$\hat{P} = 1 - \left(1 - \frac{T(m,n)}{n}\right)^{\frac{1}{m}}, \tag{4.1}$$

where $T(m,n)$ denotes the random number of positive-testing pools among n pools, each containing m specimens, that is, for $m = 1$ (i.e., individual testing), $T(1,n) \sim \text{Binomial}(n,p)$, while for $m \geq 2$ (i.e., pooled testing), $T(m,n) \sim \text{Binomial}(n, 1-(1-p)^m)$, where p denotes the (unknown) true prevalence rate, and the term $(1-(1-p)^m)$ denotes the probability that a random pool tests positive, i.e., the probability that it contains at least one true-positive specimen. The asymptotic variance of the MLE then follows (e.g., [58]):

$$\sigma^2(m,n;p) = \frac{1-(1-p)^m}{nm^2(1-p)^{m-2}} \ . \tag{4.2}$$

Note that for the special case of individual testing, we have that $\sigma^2(1,n;p) = \frac{p(1-p)}{n}$.

4.2.2 Models for Pool Design Optimization

We formulate and study the following optimization models:

PD-J Model:	*PD-S* Model:
$\underset{(m,n)\in\mathbb{F}(m,n)}{\text{minimize}} \quad \sigma^2(m,n;p_0)$	$\underset{m\in\mathbb{F}(m;n)}{\text{minimize}} \quad \sigma^2(m;n,p_0),$

where p_0 denotes an initial estimate of P. We use problem indices S and J to respectively refer to the single-variable pool design problem, *PD-S*, and the joint pool design problem, *PD-J*. We refer to the pool design problem as Problem X, $X \in \{S,J\}$, when an expression or a result applies to both problems *PD-S* and *PD-J*. We also use the superscript * to denote an optimal solution, e.g., (m_J^*, n_J^*) denotes the optimal solution of the *PD-J* Model.

To our knowledge, only the *PD-S* Model is utilized in the existing literature, i.e., for a given number of pools, n, under different objective functions, including the minimization of the asymptotic variance [58, 69], maximization of the Fisher's information via the Cramer-Rao lower bound, which reduces to a function of the asymptotic variance [57, 111], or maximization of the probability of a random pool testing positive [99], but research that analytically analyzes the model's structural properties and characterizes its optimal solution is rather limited, with the notable exception of [69], as we discuss below. Thus, our analysis of the *PD-S* Model provides a contribution to the literature in its own right; and the *PD-J* Model that we study is novel. In particular, we build upon the properties developed by [69], while considering perfect tests ([69] considers imperfect tests.) However, instead of assuming a given number of pools (n), or a given number of specimens tested (mn), as in [69], i.e., using a single-variable optimization model, we also jointly optimize over both

the pool size, m, and the number of pools, n, while explicitly accounting for the budget constraint (the *PD-J* Model). We fully characterize the optimal solutions to both the *PD-S* and *PD-J* Models by expanding upon the properties developed in [69].

4.3 Properties of the Asymptotic Variance Function

In this section, we establish key properties of the asymptotic variance function, $\sigma^2(m, n; p)$. In particular, we first study the setting with a fixed (exogenous) n value (Section 4.3.1), and then study the setting where n is optimally set (endogenous) so as to minimize the asymptotic variance (Section 4.3.2). All mathematical proofs can be found in Appendix C.2.

4.3.1 Asymptotic Variance Function for an Exogenous Number of Pools

First we consider the *single-variable pool design problem*, *PD-S*, i.e., with an exogenous number of pools, n. This is the setting that has been studied, to some extent, in the literature, as we elaborate below.

Definition 1. (From [69]) For any $n, m_1, m_2 \in \mathbb{Z}^+$: $m_1 < m_2$, the *prevalence threshold*, $\pi_0^S(m_1, m_2, n) \in (0, 1)$, is defined as the prevalence rate at which $\sigma^2\big(m_1, n; \pi_0^S(m_1, m_2, n)\big) = \sigma^2\big(m_2, n; \pi_0^S(m_1, m_2, n)\big)$.

For the exogenous n setting, the literature provides the following proposition on the prevalence threshold, $\pi_0^S(m_1, m_2, n)$ (for both perfect and imperfect tests), and studies the behavior of the asymptotic variance function mainly through numerical studies.

Proposition 1. (From [69]) For any $n, m_1, m_2 \in \mathbb{Z}^+$: $m_1 < m_2$, there exists a unique $\pi_0^S(m_1, m_2, n) \in (0, 1)$. Further:

$$\sigma^2(m_1, n; p) \begin{cases} > \sigma^2(m_2, n; p), & \forall p < \pi_0^S(m_1, m_2, n) \\ < \sigma^2(m_2, n; p), & \forall p > \pi_0^S(m_1, m_2, n) \end{cases}.$$

As discussed in [69], Proposition 1 implies that, for a given n, a smaller pool (of size m_1) is more efficient than a larger pool (of size m_2), in terms of minimizing the asymptotic variance, only if the prevalence rate p is sufficiently high, i.e., $p > \pi_0^S(m_1, m_2, n)$, and vice versa.

In the following, we establish various other properties of the asymptotic variance function, $\sigma^2(m, n; p)$, and the prevalence threshold, $\pi_0^S(m_1, m_2, n)$, for a given n. These properties not only enable us to solve the *PD-S* Model to optimality, i.e., with a given number of pools, but also confirm the numerical observations by [69], further contributing to the literature on the pool size problem for prevalence estimation. In particular, [69] numerically observe that when n is exogenously fixed, the prevalence rate below which pooled testing (i.e.,

with $m \geq 2$) is more efficient than individual testing (i.e., with a $m = 1$), i.e., $\pi_0^S(1, m, n)$, decreases in m [69]. In Lemma 2, we formally establish this result for any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, for perfect tests.

We first provide an analytical expression for the prevalence threshold function. Observe that by Definition 1 and Eqn. (4.2), when n is exogenous, $\pi_0^S(m_1, m_2, n)$ becomes independent of n, which we represent simply by $\pi_0^S(m_1, m_2)$; thus, all subsequent results in this section hold for any $n \in \mathbb{Z}^+$.

Lemma 1. For any $m_1, m_2 \in \mathbb{Z}^+$: $m_1 < m_2$, $\pi_0^S(m_1, m_2)$ is the unique solution to:

$$\left(\frac{1}{1-\pi_0^S(m_1,m_2)}\right)^{m_1} = 1 + \left(\frac{m_1}{m_2}\right)^2 \left[\left(\frac{1}{1-\pi_0^S(m_1,m_2)}\right)^{m_2} - 1\right]. \tag{4.3}$$

Lemma 2. $\pi_0^S(m_1, m_2)$ is decreasing in each of m_1 and m_2, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$.

The following corollaries follow as direct consequences of Lemma 2.

Corollary 1. $\pi_0^S(m-1, m) > \pi_0^S(m, m+1)$, $\forall m \in \mathbb{Z}^+ : m \geq 2$.

Corollary 2. The prevalence threshold function has the following properties:

1. $\pi_0^S(m_1, m_2) \leq \pi_0^S(1, 2) = \frac{2}{3}$, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$.
2. $\pi_0^S(1, m)$ is decreasing in m, $\forall m \in \mathbb{Z}^+ : m \geq 2$.

Remark 1. Proposition 1 and Corollary 2 establish the necessary and sufficient condition for individual testing to be the most efficient estimation method, that is, individual testing outperforms pooled testing, with any pool size and for any given number of pools, i.e., $\sigma^2(1, n; p) < \sigma^2(m, n; p), \forall m \geq 2$, $\forall n \in \mathbb{Z}^+$, if and only if $p > \frac{2}{3}$.

Corollary 2 further indicates that as pool size, m, increases, the prevalence rate below which pooled testing is more efficient than individual testing reduces, that is, larger pools are more efficient than individual testing at smaller prevalence rates. Corollary 2 also analytically establishes the behavior of $\pi_0^S(1, m)$, numerically observed in [69], as discussed above. Prevalence rates of almost all diseases are well below the threshold of $\frac{2}{3}$, and pooled testing is a commonly used method for disease surveillance. Thus, Corollary 2 provides an analytical justification for the widespread utilization of pooled testing for prevalence estimation of diseases.

Next we turn our attention to the asymptotic variance function.

Lemma 3. For any $n \in Z^+$, $\sigma^2(m, n; p)$ has the following properties:

1. For $m \geq 2$, $\sigma^2(m, n; p)$ is increasing in p.
2. $\sigma^2(1, n; p)$ is increasing in p, $\forall p < \frac{1}{2}$.

3. $\sigma^2(m,n;p)$ is decreasing in n.

4. $\sigma^2(m,n;p)$ is strictly convex in m.

Our results in this section characterize the structure of the prevalence threshold and asymptotic variance functions, allowing us to determine an optimal pool size for the *PD-S* Model in Section 4.4.

4.3.2 Asymptotic Variance Function for an Endogenous Number of Pools

Next we study properties of the asymptotic variance function in the *joint pool design problem*, *PD-J*, i.e., when the tester can set both n and m optimally. To this end, we first derive an expression on the optimal number of testing pools, n^*, in *PD-J* by relaxing the restriction that n is integer, i.e., we consider that n is continuous. (We discuss how to incorporate the integrality constraint on n in Section 4.5.)

Lemma 4. Consider that n is continuous. Then, without loss of optimality, the feasible region for the *PD-J* Model, given by $\mathbb{F}(m,n)$, can be replaced by $\mathbb{F}(m) = \left\{m \in \mathbb{Z}^+ : m \leq \overline{M}^J \equiv \left\lfloor \frac{B-c_f}{c_v} \right\rfloor\right\}$, with $n^*(m) \equiv \frac{B}{c_f+c_v m}$, and

$$\sigma^2(m,n^*(m);p) = \left(\frac{c_f+c_v m}{B}\right)\left(\frac{1-(1-p)^m}{m^2(1-p)^{m-2}}\right). \tag{4.4}$$

Thus, the *PD-J* Model reduces to one with a single decision variable, m, and its asymptotic variance function reduces to a function of m and p only, i.e., $\sigma^2(m,n^*(m);p)$.

We next show that many of the properties established for the *PD-S* Model (Section 4.3.1) extend to the joint *PD-J* Model. To this end, we first define the prevalence threshold in the joint setting, which, based on Lemma 4, can be represented as a function of pool sizes only.

Definition 2. For any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, the *prevalence threshold*, $\pi_0^J(m_1,m_2) \in (0,1)$, is defined as the prevalence rate at which $\sigma^2\big(m_1,n^*(m_1);\pi_0^J(m_1,m_2)\big) = \sigma^2\big(m_2,n^*(m_2);\pi_0^J(m_1,m_2)\big)$.

Proposition 2. For any $m_1, m_2 \in \mathbb{Z}^+$: $m_1 < m_2$, there exists a unique $\pi_0^J(m_1,m_2) \in (0,1)$. Further:

$$\sigma^2(m_1,n^*(m_1);p) \begin{cases} > \sigma^2(m_2,n^*(m_2);p), & \forall p < \pi_0^J(m_1,m_2) \\ < \sigma^2(m_2,n^*(m_2);p), & \forall p > \pi_0^J(m_1,m_2) \end{cases}.$$

Lemma 5. For any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, $\pi_0^J(m_1,m_2)$ is the unique solution to:

$$\left(\frac{1}{1-\pi_0^J(m_1,m_2)}\right)^{m_1} = 1 + \left(\frac{m_1}{m_2}\right)^2\left(\frac{c_f+c_v m_2}{c_f+c_v m_1}\right)\left[\left(\frac{1}{1-\pi_0^J(m_1,m_2)}\right)^{m_2} - 1\right]. \tag{4.5}$$

Lemma 5 leads to the following corollaries.

Lemma 6. $\pi_0^J(m_1, m_2)$ is decreasing in each of m_1 and m_2, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$.

Corollary 3. $\pi_0^J(m-1, m) > \pi_0^J(m, m+1)$, $\forall m \in \mathbb{Z}^+ : m \geq 2$.

Corollary 4. The prevalence threshold function has the following properties:

1. $\pi_0^J(m_1, m_2) \leq \pi_0^J(1, 2) = \frac{2c_f}{3c_f+2c_v}$, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$.
2. $\pi_0^J(1, m)$ is decreasing in m, $\forall m \in \mathbb{Z}^+ : m \geq 2$.

The following properties of $\sigma^2(m, n^*(m); p)$ will also be useful in characterizing an optimal pool design for the *PD-J* Model.

Lemma 7. $\sigma^2(m, n^*(m); p)$ has the following properties:

1. For $m \geq 2$, $\sigma^2(m, n^*(m); p)$ is increasing in p.
2. $\sigma^2(1, n; p)$ is increasing in p, $\forall p < \frac{1}{2}$.
3. $\sigma^2(m, n^*(m); p)$ is strictly convex in m, $\forall p < \frac{1}{3}$.

Our results in this section fully characterize the structure of the prevalence threshold and asymptotic variance functions for the *PD-J* Model, allowing us to determine an optimal pool design for the *PD-J* Model in Section 4.4.

4.3.3 Comparison of Prevalence Thresholds in *PD-S* and *PD-J* Models

Our comparison of the prevalence thresholds for the *PD-S* and *PD-J* Models in this section provides important insight on the impact of joint optimization on pool design. For this purpose, we define $\gamma \equiv c_f/c_v$, and first study how the prevalence threshold functions change with cost parameters.

Lemma 8. For any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, we have that $\pi_0^S(m_1, m_2)$ is independent of γ, while $\pi_0^J(m_1, m_2)$ is increasing in γ.

Lemma 9. For any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, we have that $\pi_0^J(m_1, m_2) < \pi_0^S(m_1, m_2)$.

Corollary 5. $\pi_0^J(1, 2) = \frac{2c_f}{3c_f+2c_v} < \pi_0^S(1, 2) = \frac{2}{3}$.

Thus, for any pair of pool sizes, $m_1 < m_2$, a smaller pool size (m_1) is more efficient than a larger pool size (m_2) for a wider range of prevalence values in *PD-J*, compared to *PD-S*. This in turn allows for a larger number of pools to be tested in *PD-J*, increasing the efficiency of the estimation.

4.4 Pool Design Optimization

With the properties established in Section 4.3, we are ready to characterize the optimal solutions to the pool design optimization models, *PD-S* and *PD-J*. In this section, we limit our analysis to the case where p does not exceed $\frac{1}{2}$; this is the most realistic case for disease surveillance studies, as discussed above. Thus, all results in this section apply to the case where $p < \frac{1}{2}$.

Lemma 10. For any $m \in \mathbb{Z}^+$: $m \geq 1$:

1. When n is exogenously fixed, we have the following properties:

   (i) $\sigma^2(m, n; p) < \sigma^2(k, n; p)$, $\forall k \in \mathbb{Z}^+ : k \neq m$, and $\forall n \in \mathbb{Z}^+$, if and only if $\pi_0^S(m, m+1) < p < \pi_0^S(m-1, m)$.

   (ii) $\sigma^2(m, n; p) = \sigma^2(m+1, n; p)$, if and only if $p = \pi_0^S(m, m+1)$.

   (iii) $\sigma^2(m, n; p) = \sigma^2(m-1, n; p)$, if and only if $p = \pi_0^S(m-1, m)$.

2. When n can be optimally set, i.e., $n = n^*(m)$, we have the following properties:

   (i) $\sigma^2(m, n^*(m); p) < \sigma^2(k, n^*(k); p)$, $\forall k \in \mathbb{Z}^+ : k \neq m$, if and only if $\pi_0^J(m, m+1) < p < \pi_0^J(m-1, m)$.

   (ii) $\sigma^2(m, n^*(m); p) = \sigma^2(m+1, n^*(m+1); p)$, if and only if $p = \pi_0^J(m, m+1)$.

   (iii) $\sigma^2(m, n^*(m); p) = \sigma^2(m-1, n^*(m-1); p)$, if and only if $p = \pi_0^J(m-1, m)$.

Lemma 10 allows us to fully characterize the optimal *PD-S* and *PD-J* solutions. Recall that the testing budget constraint imposes an upper bound on pool size m, i.e., $m \leq \overline{M}^S(n) \equiv \left\lfloor \frac{B - c_f n}{c_v n} \right\rfloor$ for the *PD-S* Model, and $m \leq \overline{M}^J \equiv \left\lfloor \frac{B - c_f}{c_v} \right\rfloor$ for the *PD-J* Model. In the *PD-S* Model, since the number of pools, n, is fixed exogenously, the optimal pool size, m_S^*, depends on n only through the upper bound, $\overline{M}^S(n)$.

Theorem 1. For a given p_0 and an exogenously fixed n, an optimal solution to the *PD-S* Model follows a threshold policy:

$$m_S^*(p_0) = \begin{cases} \overline{M}^S(n), & \text{if} \quad p_0 \le \pi_0^S(\overline{M}^S(n)-1, \overline{M}^S(n)) \\ \vdots & \\ m+1, & \text{if} \quad \pi_0^S(m+1,m+2) \le p_0 \le \pi_0^S(m,m+1) \\ m, & \text{if} \quad \pi_0^S(m,m+1) \le p_0 \le \pi_0^S(m-1,m) \\ m-1, & \text{if} \quad \pi_0^S(m-1,m) \le p_0 \le \pi_0^S(m-2,m-1) \\ \vdots & \\ 1, & \text{if} \quad \pi_0^S(1,2) \le p_0 < 1, \end{cases}$$

where $\pi_0^S(1,2) = \frac{2}{3}$.

Theorem 2. For a given p_0, an optimal solution to the *PD-J* Model follows a threshold policy:

$$m_J^*(p_0) = \begin{cases} \overline{M}^J, & \text{if} \quad p_0 \le \pi_0^J(\overline{M}^J-1, \overline{M}^J) \\ \vdots & \\ m+1, & \text{if} \quad \pi_0^J(m+1,m+2) \le p_0 \le \pi_0^J(m,m+1) \\ m, & \text{if} \quad \pi_0^J(m,m+1) \le p_0 \le \pi_0^J(m-1,m) \\ m-1, & \text{if} \quad \pi_0^J(m-1,m) \le p_0 \le \pi_0^J(m-2,m-1) \\ \vdots & \\ 1, & \text{if} \quad \pi_0^J(1,2) \le p_0 < 1, \end{cases}$$

where $\pi_0^J(1,2) = \frac{2c_f}{3c_f+2c_v}$, and $n_J^*(p_0)$ can be determined by Lemma 4, that is, $n_J^*(p_0) = \frac{B}{c_f+c_v m_J^*(p_0)}$.

Thus, the optimal *PD-X*, $X \in \{S, J\}$, solution is unique if the tester's initial point estimate, p_0, does not correspond to a prevalence threshold point ($p_0 \neq \pi_0^X(m,m+1)$, $\forall m \in \mathbb{Z}^+$); and there are dual optimal solutions if p_0 corresponds to a prevalence threshold point.

Using the convexity of the asymptotic variance function leads to the following alternative characterization for the optimal solutions to the *PD-S* and *PD-J* Models.

Lemma 11. Consider *PD-X*, $X \in \{S, J\}$, and let m' denote the solution to the first-order condition (FOC), that is, $\left\{m' : \frac{\partial}{\partial m}\sigma^2(m,n;p_0)\big|_{m=m'} = 0\right\}$ in *PD-S*, and $\left\{m' : \frac{\partial}{\partial m}\sigma^2(m,n^*(m);p_0)\big|_{m=m'} = 0\right\}$ in *PD-J*, or equivalently,

$$m' : \begin{cases} m' = \frac{2[(1-p_0)^{m'}-1]}{\log(1-p_0)}, & \text{for } PD\text{-}S; \\ m' = \frac{\left(1+\frac{c_f}{c_f+c_v m'}\right)[(1-p_0)^{m'}-1]}{\log(1-p_0)}, & \text{for } PD\text{-}J. \end{cases}$$

Then $m^*_{PD-X} \in \left\{m', \lceil m' \rceil, \lfloor m' \rfloor, \overline{M}^X\right\}$.

Theorems 1 and 2, along with Lemmas 3 and 7, lead to the following results.

Corollary 6. For $X \in \{S, J\}$, *PD-X* optimal solution, $m^*_X(p_0)$, is non-increasing in p_0, with $m^*_X(p_0) = 1$, $\forall p_0 \geq \pi^X_0(1, 2)$.

Thus, testing large pools becomes more efficient as p_0 decreases. This property, that pools should get smaller at larger prevalence rates, is so as to gather some "useful" information from testing. In particular, if pools are very large when the disease prevalence is large, then it is likely that many pools will test positive (i.e., contain at least one infected specimen); and similarly, if pools are very small when the disease prevalence is small, then it is likely that many pools will test negative (i.e., will not contain any infected specimen). Thus, an optimal pool design balances these risks, and attempts to gather some useful information from testing. However, when p_0 is a poor estimate of the true p value, $m^*_S(p_0)$ can still be highly inefficient, and, further, an outcome of all positive-testing pools or all negative-testing pools may occur, as we discuss in Section 4.5.

Next, we study how the testing cost structure affects the *PD-X*, $X \in \{S, J\}$, optimal solution.

Corollary 7. For $X \in \{S, J\}$, *PD-X* optimal solution, m^*_X, is non-decreasing in γ.

Thus, an increase in the specimen collection cost, c_v, for a given value of c_f (i.e., a reduction in the value of $\gamma = \frac{c_f}{c_v}$), makes it more efficient to use smaller pool sizes, in turn leading to an increase in the number of pools tested in *PD-J*.

4.5 Case Study: Prevalence Estimation of West Nile Virus in Mosquitoes

Our goal in this section is to gain insights and derive principles on testing pool design for prevalence estimation. For this purpose, we utilize the optimization models, *PD-S* and *PD-J*, to design pooled testing schemes so as to efficiently estimate the prevalence of mosquitoes carrying West Nile virus (WNV), and compare the outcomes with those of a benchmark design.

The WNV disease is a vector-borne disease, and the primary source of disease transmission to humans is a mosquito bite [52]. The mosquito population carrying WNV in any given year depends on various factors, including the temperature and humidity [39, 52]. As a result, the prevalence rate of WNV in mosquitoes is highly seasonal, and can fluctuate substantially from year to year [39, 52, 65].

The WNV disease has become a seasonal endemic in the United States, causing several fatalities from neuro-invasive diseases [52,65]. Further, as most cases of WNV infections are asymptomatic, WNV infections in humans are significantly under-reported [109], further contributing to the risk of transfusion-transmitted infections via blood transfusion or organ transplantation [60,90]. The prevalence rate of WNV in humans has been shown to be highly correlated with the prevalence rate of WNV in mosquitoes [31,36,64]. Therefore, an accurate estimation of the prevalence rate of WNV-carrying mosquitoes is essential for outbreak prediction and prevention of the WNV disease in humans [52, 53].

We consider the reverse-transcription polymerase chain reaction (RT-PCR) assay for WNV screening in mosquitoes [53]. RT-PCR assays employ the nucleic acid amplification technology for detecting the viral RNA present in the specimens, and are highly sensitive and specific [77]. For illustrative purposes, we assume a testing budget of $B = \$8,160$, which is equivalent to a pool design of $(m = 50, n = 30)$, given the testing cost parameters from [53]; see Table 4.1. This benchmark design is highly relevant in our context, as it utilizes a pool size commonly used for prevalence estimation of WNV via RT-PCR assays (e.g., [48, 49, 73, 81, 83]). All data used in our numerical study come from published studies, and we complement these data with sensitivity analysis; see Table 4.1.

Recall that both *PD-S* and *PD-J* Models require an initial estimate of the unknown prevalence rate, p_0, as input, and we study the performance of both models based on various values of p_0, which we derive based on the support of the unknown prevalence rate, P, denoted by $[p_{LB}, p_{UB}]$. To construct the support of P, we use WNV surveillance data from various parts of the Mid-South region of the United States, where transmission of WNV infection was the most intense during the outbreak from 2002 to 2003 [52]. Since the prevalence rate of WNV-infected mosquitoes is reported to be as high as 8.76% in Tennessee Valley during the 2002-2005 period [34], we use an upper bound of $p_{UB} = 9\%$, and use a lower bound of $p_{LB} = 0.3\%$ [81] in our study. In particular, we consider that the distribution of P is $P \sim$ Uniform (p_{LB}, p_{UB}) (the distribution information is only needed for the Monte-Carlo simulation, as we discuss subsequently), and study three scenarios:

1. Scenario 1: $p_0 = \frac{1}{2}(p_{LB} + p_{UB}) = 0.0465$, i.e., p_0 is set to the true mean of P.
2. Scenario 2: $p_0 = \frac{1}{4}(p_{LB} + p_{UB}) = 0.0233$, i.e., p_0 is an underestimate of the true mean of P.
3. Scenario 3: $p_0 = \frac{3}{4}(p_{LB} + p_{UB}) = 0.0698$, i.e., p_0 is an overestimate of the true mean of P.

Table 4.1: Data and sources for the numerical study.

	Cost Parameters (Source)	
c_f	\$72 [53]	
c_v	\$4 [53]	
	Model Input (Scenario 1)	**Sensitivity Analysis (Scenarios 2 and 3)**
	$p_0 = \frac{1}{2}\left(p_{LB} + p_{UB}\right)$	$p_0 = \frac{1}{4}\left(p_{LB} + p_{UB}\right)$; $p_0 = \frac{3}{4}\left(p_{LB} + p_{UB}\right)$
	Benchmark Design (Source)	**Benchmark Budget (Source)**
	$(m, n) = (50, 30)$ (e.g., [83])	$B = \$8,160$ (e.g., [53, 83])

We perform a Monte-Carlo simulation with 20,000 replications for each scenario. Specifically, for each of the three scenarios, we first determine the optimal pool designs for the *PD-S* and *PD-J* Models based on the inputs provided in Table 4.1. To obtain an optimal integer number of pools in *PD-J*, we utilize Theorem 2 repeatedly within a branch and bound algorithm in MATLAB [67]. Each simulation replication corresponds to a randomly generated realization of P from a Uniform (p_{LB}, p_{UB}) distribution, denoted by p. Based on the generated value of p, we then randomly generate the carrier status of each subject (specimen from a mosquito) (a total of $m^* \times n$ subjects in the *PD-S* Model, and $m^* \times n^*(m^*)$ subjects in the *PD-J* Model), where each specimen has the WNV infection with probability p, and is infection-free with probability $1 - p$. These specimens are then randomly assigned to the testing pools. If a pool contains at least one infected specimen, then the test outcome for the pool will be positive; and otherwise, the test outcome for the pool will be negative. Given a set of test outcomes, we then compute the MLE of P using Eqn. (4.1), i.e., $\hat{p}$. In each replication, we compute the following performance metrics: the prevalence estimate ($\hat{p}$), asymptotic variance ($\sigma^2(m^*, n^*; p)$) (see Eqn. 4.2), MSE, and percent relative bias of the prevalence estimate ($rBias(\%)$), where:

$$MSE = (\hat{p} - p)^2, \text{ and } rBias(\%) = 100 \times \left|\frac{\hat{p} - p}{p}\right|.$$

Tables 4.2 and 4.3 report the average of each performance metric over 20,000 replications. All performance metrics are reported in the form: *average ± half width of the 95% confidence interval (CI)*. These performance metrics are commonly used in the statistics literature to evaluate the efficiency of an estimator (e.g., [57, 58, 111]).

4.5.1 Pool Design Models versus the Benchmark Design

Table 4.2 reports the results for scenario 1, i.e., $p_0 = \frac{1}{2}(p_{LB} + p_{UB})$. Since the number of pools, n, is fixed exogenously in the *PD-S* Model, we consider a range of values for n for the *PD-S* Model; see Table 4.2.

The most important finding from Table 4.2 is that jointly selecting m and n in an optimal manner can significantly improve performance, while *a priori* selecting a sub-optimal n value can lead to a rather poor

Table 4.2: Performance of the benchmark design, and the *PD-S* and *PD-J* Models with $p_0 = \frac{1}{2}(p_{LB} + p_{UB})$ (average ± half width of 95% CI.)

n	**Benchmark Design**	
30	m	*50*
	$\hat{p}$	*0.2364 ± 0.00541*
	$\sigma^2(m^*, n; p)[\times 10^6]$	*262 ± 4.30*
	$MSE[\times 10^6]$	*177,000 ± 4,806*
	$rBias(\%)$	*278.70 ± 7.22*
n	***PD-S* Model**	
20	m^*	33
	$\hat{p}$	0.1174 ±0.00355
	$\sigma^2(m^*, n; p)[\times 10^6]$	240.0 ± 3.10
	$MSE[\times 10^6]$	$65,500 \pm 3,140$
	$rBias(\%)$	114.90 ±4.45
30	m^*	33
	$\hat{p}$	0.08294 ±0.00257
	$\sigma^2(m^*, n; p)[\times 10^6]$	159.0 ± 2.04
	$MSE[\times 10^6]$	$32,600 \pm 2,254$
	$rBias(\%)$	64.13 ±3.08
50	m^*	22
	$\hat{p}$	0.04780 ±0.00039
	$\sigma^2(m^*, n; p)[\times 10^6]$	86.70 ± 0.09
	$MSE[\times 10^6]$	102 ± 5.56
	$rBias(\%)$	17.04 ±0.21
100	m^*	2
	$\hat{p}$	0.04675 ±0.00041
	$\sigma^2(m^*, n; p)[\times 10^6]$	226.0 ± 1.93
	$MSE[\times 10^6]$	230 ± 5.45
	$rBias(\%)$	31.80 ±0.43
***PD-J* Model**		
(m^*, n^*)		(19,55)
$\hat{p}$		0.04750 ±0.00038
$\sigma^2(m^*, n^*; p)[\times 10^6]$		80.40 ± 0.81
$MSE[\times 10^6]$		88.20 ± 2.89
$rBias(\%)$		16.64 ±0.21

performance. Furthermore, the benchmark design (m=50, n=30) performs very poorly compared to all the *PD-S* Models (if n is set to 30, a pool size of 50 is not efficient given the support of P, and, thus, it is best not to use the full budget). We also note that when $n = 50$ or $n = 100$, the *PD-S* Model chooses the maximum feasible pool size as the optimal pool size, i.e., $m^* = \overline{M}^S(n)$.

4.5.2 Sensitivity Analysis

Next, we compare the performance of the *PD-S* and *PD-J* Models in all three scenarios that we consider, i.e., $p_0 = \frac{1}{2}(p_{LB} + p_{UB})$, $p_0 = \frac{1}{4}(p_{LB} + p_{UB})$, and $p_0 = \frac{3}{4}(p_{LB} + p_{UB})$; see Table 4.3. In cases where $m^* = \overline{M}^S(n)$ in the *PD-S* Model, i.e., when $n = 50$ or $n = 100$, the pool size in the *PD-S* Model does not change with

p_0. Therefore, in Table 4.3, we report the performance of the *PD-S* Model for the different scenarios only for the $n = 20$ and $n = 30$ cases.

Table 4.3: Performance of the *PD-S* and *PD-J* Models with various values of p_0 (average ± half width of 95% CI.)

***PD-S* Model**				
n		$p_0 = \frac{1}{2}(p_{LB} + p_{UB})$	$p_0 = \frac{1}{4}(p_{LB} + p_{UB})$	$p_0 = \frac{3}{4}(p_{LB} + p_{UB})$
20	m^*	33	68	22
	$\hat{p}$	0.1174 ±0.00355	0.4639 ± 0.00672	0.0567 ± 0.00131
	$\sigma^2(m^*, n; p)[\times 10^6]$	240.0 ± 3.10	873.0 ± 17.70	216.0 ± 2.32
	$MSE[\times 10^6]$	$65,500 \pm 3,140$	391,000 ± 6,007	7,760 ± 1,098
	$rBias(\%)$	114.90 ±4.45	682.68 ± 10.98	37.77 ± 1.51
30	m^*	33	50	22
	$\hat{p}$	0.08294 ±0.00257	0.2404 ± 0.00545	0.05012 ± 0.000710
	$\sigma^2(m^*, n; p)[\times 10^6]$	159.0 ± 2.04	265.0 ± 4.35	144.0 ± 1.53
	$MSE[\times 10^6]$	$32,600 \pm 2,254$	180,000 ± 4,839	1,820 ± 514
	$rBias(\%)$	64.13 ±3.08	283.06 ± 7.24	24.57 ± 0.72
***PD-J* Model**				
(m^*, n^*)		(19,55)	(29,43)	(14,63)
$\hat{p}$		0.04750 ±0.00038	0.05292 ±0.00105	0.04724 ±0.00038
$\sigma^2(m^*, n^*; p)[\times 10^6]$		80.40 ± 0.81	104 ± 1.25	77.90 ± 0.71
$MSE[\times 10^6]$		88.20 ± 2.89	$4,740 \pm 860$	84.70 ± 2.52
$rBias(\%)$		16.64 ±0.21	23.94 ±1.13	17.13 ±0.21

Table 4.3 shows that the performance of both *PD-S* and *PD-J* Models deteriorates when p_0 is an underestimate of the true mean of P, i.e., when $p_0 = \frac{1}{4}(p_{LB} + p_{UB})$. However, the performance of the *PD-J* Model is much more robust, in comparison to that of the *PD-S* Model, in this case. Therefore, the *PD-J* Model is preferable in cases where the true prevalence rate is highly uncertain, and, hence, p_0 can be a poor estimate of the true prevalence rate. Interestingly, we observe that when the true prevalence rate follows a distribution (such as the setting considered in our Monte-Carlo simulation), using an overestimate of the true mean of P, i.e., $p_0 = \frac{3}{4}(p_{LB} + p_{UB})$, results in more efficient pool designs in both models over all scenarios considered. This insight is of particular relevance to prevalence estimation of emerging or seasonal infections such as WNV. Since the prevalence rates of these infections are highly uncertain, using a conservative estimate of P, i.e., an overestimate of P, is more beneficial in designing testing pools for prevalence estimation, as also observed by, e.g., [58, 75].

4.6 Discussion

In this paper, we develop and study two pool design optimization models for prevalence estimation under limited resources: the single-variable pool design model, *PD-S*, where the number of pools is fixed exogenously, and the joint pool design model, *PD-J*, where both the pool size and the number of pools are set

optimally. Our models are quite general, and can apply to prevalence estimation of various diseases. Further, our *PD-J* Model has an important advantage over the models studied in the literature: since it allows for both the pool size and the number of pools to be set optimally, the tester does not need to determine the number of testing pools, or testing kits, in advance. This flexibility of the *PD-J* Model is especially desirable for prevalence estimation of emerging or seasonal infections, as our analysis shows that it provides "good" testing designs even when input parameters are inaccurate. This is not the case for the *PD-S* Model, which is often used in the existing literature, since it is highly dependent on an initial point estimate of the prevalence rate, i.e., it can yield poor solutions if the point estimate is not accurate. From that perspective, the *PD-J* Model applies especially well to prevalence estimation of emerging or seasonal diseases, such as Zika or West Nile virus disease, for which initial information, prior to testing, is often highly unreliable. We develop a threshold-type policy for determining the optimal testing pool design, or optimal pool size, for each model. We also establish key structural properties of optimal pool designs and analytically characterize the form of optimal pool designs in various settings, further contributing to the existing literature of testing pool design for prevalence estimation.

Our case study, on estimating the prevalence of West Nile virus in mosquitoes, compares our models, both with the number of pools set exogenously, i.e., *PD-S*, and jointly optimized with the pool size, i.e., *PD-J*, against a benchmark design from the literature. There are several important findings from this study. First, our designs significantly outperform the benchmark design used in the literature; in fact, our comparison of the benchmark design with the associated *PD-S* designs shows that it is not always the best strategy to maximize the pool size given the remaining budget; sometimes using smaller pools than what the budget allows is better. Our findings also underscore the importance of *jointly* optimizing over pool size and number of pools in order to accurately estimate P. When estimating the prevalence of emerging and/or seasonal diseases, the distribution and support of P are highly uncertain at the outset, and thus input parameters are likely to be inaccurate. It is in these realistic cases that the *PD-J* Model performs significantly better than the *PD-S* Model. These findings have important implications for designing surveillance studies.

As immediate, and important, extensions of our study, one can relax some of our modeling assumptions, including that the screening test is perfectly reliable. In addition, the requirement of an initial point estimate, i.e., p_0, as model input can be removed by applying robust optimization models and methodologies to testing pool design. Further, as healthcare policy-makers may need to allocate their testing budget to prevalence estimation activities for a number of diseases, or over various regions, each potentially having a different prevalence rate of the disease in question, extending our models to study pool design and budget allocation for prevalence estimation for multiple diseases or multiple regions is also an important direction for future research.

Chapter 5

Robust Pooled Testing Design for Prevalence Estimation of Emerging and Seasonal Diseases

5.1 Introduction

Emerging and/or seasonal diseases pose significant challenges to public health policy and health systems management due to their highly stochastic nature, in terms of disease status and dynamics; and outbreaks of these diseases are extremely costly to the society. As examples, the 2016 Zika outbreak in the Latin American and Caribbean region cost approximately $3.5 billion in treatment, prevention, and lost revenues ([107]); the annual treatment costs alone for the highly seasonal West Nile virus and Lyme disease are estimated to be around $778 million and $1.3 billion, respectively, in the United States ([1,9]). An accurate estimate of the disease prevalence rate is crucial for outbreak detection, disease response and prevention, and healthcare services management (e.g., [35, 39, 104]), and is the problem studied in this paper.

While an effective surveillance of emerging and seasonal diseases is essential for the welfare of the society, the funds and resources that can be allocated to testing activities needed to estimate the prevalence of the disease in question are often very small in comparison to the needs ([35]). As a result, prevalence estimation via individual testing of each subject is either infeasible, or highly inefficient, leading to a small sample sizes and to potentially inaccurate estimates ([54, 99, 100]). An effective and efficient solution to prevalence estimation under limited resources comes in the form of *pooled testing*, i.e., combining specimens (e.g., blood,

urine, tissue swabs) from multiple subjects in a testing pool and testing the pool via a single test ([38]). Since its introduction by Dorfman in the 1940's ([38]), pooled testing has been shown to be highly efficient for both prevalence estimation and subject identification problems (the latter problem seeks to identify all infected subjects, which is not the focus of this paper), and is now a widely used testing method for both purposes (e.g., [37, 39, 56, 85, 104]). If used for subject identification, pooled testing is typically followed by individual testing of the positive-testing pools, but this additional step is typically not conducted for prevalence estimation, as the ultimate goal is to derive an accurate estimate of the disease prevalence rate (e.g., [50, 56, 69, 70, 72]). This is especially true when the goal is to estimate the prevalence of the "sources" of vector-borne viral or bacterial diseases, e.g., mosquitoes carrying Zika virus or West Nile virus ([83]), ticks carrying Lyme disease or Babesiosis ([20]), romaine lettuce carrying *E. coli* bacteria ([22]). Thus, in prevalence estimation, the test measures the pool's concentration of a certain bio-marker that serves as an indicator for the presence of the virus or bacteria of interest, and provides a binary outcome: *positive*, indicating the presence of at least one disease-positive specimen in the pool, and *negative* otherwise; and inference on the unknown prevalence rate is made based on the collected testing data.

While the accuracy of the prevalence estimate depends highly on the testing *pool design*, i.e., number of pools to test and pool size (the number of specimens to combine in each pool) ([25, 75, 92, 96]), the pool design problem itself requires some initial estimate of the disease prevalence rate, which is highly unreliable prior to surveillance, especially for emerging and seasonal diseases. This creates challenges for obtaining an optimal pool design, and leads to potentially inaccurate estimates. On the other hand, the literature on prevalence estimation mainly focuses on how to derive an efficient prevalence estimate from the testing data, for a *given* pool design, i.e., the "estimation" component of the prevalence estimation problem (e.g., [26–28, 40, 41, 54, 55, 81, 96, 102, 110]). On the pool design component, there is a limited number of studies, most of which study only the pool size optimization problem, i.e., for a *given* number of testing pools, under perfect tests (e.g., [57, 58, 89]), and imperfect tests (e.g., [45, 69, 99, 100]). These studies require an initial point estimate of the unknown prevalence rate for determining the pool size, but they do not offer a mechanism to hedge against the uncertainty in the initial estimate (e.g., [45, 69, 99, 100, 111]), except for a few studies that use sequential approaches, as we discuss below. However, it has been shown that a pool design that relies highly on an initial point estimate of the prevalence rate, or that corresponds to an exogenously fixed number of testing pools, can result in highly inaccurate estimates of the prevalence rate ([57, 58, 75, 76]). In particular, [76] shows that *jointly* optimizing both the pool size and the number of pools can provide substantial benefit, leading to a more accurate estimate at the same testing budget, and this is the approach we take in this paper.

As mentioned above, one approach to account for uncertainty in the initial estimate of the prevalence

rate is to use a *sequential (multi-stage)* estimation procedure, which allows the prevalence rate estimate to be updated as testing proceeds, based on the testing data collected thus far, so that the remaining tests can be conducted with a pool design that is based on a more accurate estimate (e.g., [57, 58, 75]). However, sequential estimation procedures lead to new operational challenges, including the need to determine the split of the testing budget among the different testing stages, which have not been addressed adequately in the literature. This is mainly because the testing outcome at the end of each stage, which is an input to the subsequent stage, is random, and closed-form analytical expressions are not available for the resulting estimator and its efficiency. In addition, even within a sequential procedure, one still needs to solve the pool design problem at the beginning of each testing stage, under uncertainty, and the current research in this area is limited, as discussed above.

Motivated by these gaps in the literature, in this paper we propose *robust optimization* approaches to testing pool design under uncertainty and limited resources. Our novel robust optimization models for pool design do not require an initial point estimate, and are based solely on the support of the unknown prevalence rate, which is much easier to estimate accurately than its distribution or moments; and provide robust solutions that perform well even when the support information is not perfectly accurate. Further, our models are flexible, in that they can be integrated into both single and multi-stage estimation frameworks. We establish important structural properties of optimal robust pool designs, and complement our analytical results with a case study on the prevalence estimation of West Nile virus in mosquitoes, which illustrates the value of robust pool designs in both single and multi-stage estimation frameworks. Our findings indicate that using a robust pool design within a single-stage estimation framework can reap almost all benefits of multi-stage frameworks, which can be very difficult to implement, and this finding continues to hold even when there is limited information on the current status of the disease prior to testing.

The remainder of this paper is organized as follows. In Section 5.2, we present the notation, modeling assumptions, single and multi-stage estimation frameworks; and formulate the robust pool design optimization models. In Section 5.3, we analytically characterize the optimal robust pool designs, and develop exact algorithms that can efficiently obtain the optimal pool designs. We then demonstrate the benefits of the proposed robust pool designs in both single and multi-stage estimation frameworks, through a case study of West Nile virus prevalence estimation in mosquitoes in Section 5.4. Finally, we conclude in Section 5.5 with a discussion of our findings and suggestions for future research. To facilitate the presentation, all mathematical proofs and some tables are relegated to the Appendix.

5.2 Notation, Assumptions, and Models

This section is organized as follows. In Section 5.2.1, we introduce the notation and discuss the modeling assumptions. Then in Sections 5.2.2 and 5.2.3, we respectively outline the single and multi-stage estimation frameworks, and the pool design optimization models.

5.2.1 Notation and Assumptions

Throughout, we denote random variables in upper–case letters, their realization in lower–case letters, and vectors in bold-face. A summary of the notation can also be found in Appendix D.1.

Our goal is to determine a pool design so as to accurately estimate an unknown prevalence rate, P, of a disease. Our approach is distribution-free, that is, P is allowed to follow any continuous distribution within the family of all continuous distributions having support $\left[p_{LB}, p_{UB}\right]$, where $0 < p_{LB} < p_{UB} < 1$. Further, our approach can be used both within single and multi-stage estimation frameworks. In particular, in each testing stage s, $s = 1, 2, \cdots, S$, the tester determines the pool design, i.e., pool size, $m^{(s)}$, and number of pools to test, $n^{(s)}$, under a limited testing budget, $B^{(s)}$, and uses this pool design to test subjects and collect testing data, which is then used to derive an estimate of the unknown prevalence rate at the conclusion of all testing. The objective is to determine a pool design that minimizes a function of the asymptotic variance of the estimator, as we detail below. Testing incurs a fixed cost (e.g., cost of the testing kit), of c_f per testing pool, and a variable cost (e.g., collection cost), of c_v per specimen tested, with $c_f > c_v$, and the total testing budget available to the tester is given by B. Then, in the pool design problem in stage s, $s = 1, 2, \cdots, S$, the feasible set for decision variables $m^{(s)}$ and $n^{(s)}$ is given by, $\mathbb{F}(m^{(s)}, n^{(s)}) \equiv \left\{(m^{(s)}, n^{(s)}) \in \mathbb{Z}^+ : c_f n^{(s)} + c_v m^{(s)} n^{(s)} \leq B^{(s)}\right\}$, where $\sum_{s=1}^{S} B^{(s)} = B$; and we use the superscript * to denote an optimal solution in stage s, i.e., $(m^{(s)*}, n^{(s)*})$, with $(\mathbf{m}^*, \mathbf{n}^*) = \{(m^{(s)*}, n^{(s)*}), s = 1, \cdots, S\}$ denoting an optimal pool design vector. In this paper, we focus on the single-stage, i.e., $S = 1$, and two-stage, i.e., $S = 2$, estimation frameworks; our models readily extend to other multi-stage estimation frameworks, i.e., with $S > 2$. Then, for a given budget allocation factor λ, $0 < \lambda \leq 1$, the total budget B can be split into $B^{(1)} = \lambda B$ and $B^{(2)} = (1 - \lambda)B$ for $S = 2$, and for $S = 1, \lambda = 1$.

In each stage of the estimation framework, the objective is to minimize a function of the asymptotic variance of the estimator, denoted by $\sigma^2(m^{(s)}, n^{(s)}; p)$, which is a commonly used metric for both pool design and estimation efficiency (e.g., [58, 66, 69, 93, 98, 100, 102]). Specifically, $\sigma^2(m^{(s)}, n^{(s)}; p)$, corresponding to a *true* prevalence rate p, represents the limiting behavior of the mean squared error (MSE) (i.e., variance plus bias square) of an estimator $\hat{P}$ as the number of pools, $n^{(s)}$, becomes large. The asymptotic variance also provides an approximation for the Cramer-Rao lower bound on the Fisher's information obtained from

the prevalence estimate (e.g., [18]). Therefore, the asymptotic variance is commonly utilized in the pool design literature, because, in general, computing the MSE or the Fisher's information is often intractable (e.g., [57, 58, 104]).

The test can be applied to pools of specimens collected from subjects (i.e., $m \geq 2$) as well as to individual specimens (i.e., $m = 1$). We assume that the test is perfectly reliable, and provides a binary outcome, that is, the test has perfect *sensitivity* (true positive probability) and *specificity* (true negative probability), thus providing a *positive* outcome only if there is at least one true-positive specimen in the pool, and a *negative* outcome only if all specimens in the pool are true-negative.

5.2.2 The Sequential Estimation Framework

In what follows, we provide an outline of the two-stage sequential (two-stage) estimation framework, a special case of which reduces to the single-stage estimation framework ([57,58,75]). To this end, let $T(m^{(s)*}, n^{(s)*})$ denote the random number of positive-testing pools among $n^{(s)}$ pools, each containing $m^{(s)}$ specimens, for $s = 1, \cdots, S$, where $S = 1$ for the single-stage framework, and $S = 2$ for the two-stage framework.

Stage s, $\mathbf{s = 1, \cdots, S}$:

1. Determine the optimal pool design, $(m^{(s)*}, n^{(s)*})$ (see Section 5.2.3).
2. Test specimens using the design, $(m^{(s)*}, n^{(s)*})$, and obtain the test outcome, $T(m^{(s)*}, n^{(s)*})$.
3. Compute $\hat{P}^{(s)}$ via the maximum likelihood estimator (MLE) function, which is the unique solution to the following equation (e.g., [58]):

$$\sum_{j=1}^{s} \frac{(m^{(j)*}) \times T(m^{(j)*}, n^{(j)*})}{1 - (1 - \hat{P}^{(s)})^{m^{(j)*}}} = \sum_{j=1}^{s} m^{(j)*} \times n^{(j)*}. \tag{5.1}$$

4. If $s = 1$, then update the input for the second stage pool design model based on $\hat{P}^{(1)}$ (see Section 5.2.3), and repeat the process for stage $s = 2$. If $s = 2$, then terminate the testing, with a final estimate, $\hat{P}^{(2)}$.

To simplify the notation, we denote the final estimate of the prevalence rate as $\hat{P}$, i.e., $\hat{P} = \hat{P}^{(1)}$ for $S = 1$, and $\hat{P} = \hat{P}^{(2)}$ for $S = 2$.

In the next section, we discuss the pool design optimization models that are to be used within this estimation framework. All pool design optimization models utilize a function of the asymptotic variance of the MLE in their objective function. The asymptotic variance of the MLE, corresponding to a pool design

(m, n) and a true prevalence rate p, follows (e.g., [58]):

$$\sigma^2(m, n; p) = \frac{1 - (1-p)^m}{nm^2(1-p)^{m-2}} \ . \tag{5.2}$$

The asymptotic variance of the final MLE, given an optimal pool design vector $(\mathbf{m}^*, \mathbf{n}^*)$, and a true prevalence rate p, follows (e.g., [58]):

$$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p) = \left[\sum_{s=1}^{S} \frac{n^{(s)*}(m^{(s)*})^2(1-p)^{m^{(s)*}-2}}{1-(1-p)^{m^{(s)*}}} \right]^{-1}, \tag{5.3}$$

where the case with $S = 1$ reduces to Eqn. (5.2).

5.2.3 Pool Design Optimization Models

We formulate and study two robust optimization models, which differ in their objective function: the *Mini-max Pool Design Model* (*MM*) and the *Regret-based Pool Design Model* (*RM*), both of which require only the support of P, given by $[p_{LB}, p_{UB}]$. In order to quantify the benefit of the robust pool designs, we compare them with a *benchmark* model from the literature, referred to as the *Deterministic Pool Design Model* (*DM*), which requires an initial point estimate of P, denoted by p_0 ([76]). In what follows, we use the subscript M, R, and D to respectively refer to *MM, RM,* and *DM*, and the superscript $s = 1, 2$, to refer to estimation stage s, e.g., $(m_X^{(s)*}, n_X^{(s)*})$ denotes the optimal pool design for Model X, $X \in \{M, R, D\}$, in stage s, $s = 1, 2$. We drop index s when a model or result applies to all estimation stages.

Mini-max Pool Design Model *(MM)*:

$$\begin{aligned} \text{minimize} \quad & \max_{p \in [p_{LB}, p_{UB}]} \left\{ \sigma^2(m, n; p) \right\} \\ \text{subject to} \quad & c_f n + c_v mn \leq B \qquad \text{(C1)} \\ & m, n \in \mathbb{Z}^+ \qquad \text{(C2)} \end{aligned}$$

Regret-based Pool Design Model *(RM)*:

$$\begin{aligned} \text{minimize} \quad & \max_{p \in [p_{LB}, p_{UB}]} \left\{ Regret(m, n; p) \right\}, \\ \text{subject to} \quad & \text{(C1) – (C2)} \end{aligned}$$

Benchmark Deterministic Pool Design Model (DM) ([76])

$$\begin{aligned} \text{minimize} \quad & \sigma^2(m, n; p_0) \\ \text{subject to} \quad & \text{(C1) – (C2)} \end{aligned}$$

The *Regret* function used in *RM* is defined similar to the literature (e.g., [10, 39, 79]):

$$Regret(m, n; p) = \sigma^2(m, n; p) - \sigma^2(m_D^*(p), n_D^*(p); p), \ \forall p \in [p_{LB}, p_{UB}], \forall (m, n) \in \mathbb{F}(m, n), \tag{5.4}$$

that is, for a given prevalence rate p, $Regret(m,n;p)$ represents the "cost" (i.e., increase in the asymptotic variance) resulting from the use of some pool design (m,n), instead of an optimal design $(m_D^*(p), n_D^*(p))$ had the tester known the true p value, i.e., the optimal solution to *DM* with $p_0 = p$.

Observe that the robust model, *MM*, is concerned only with the worst-case outcome, i.e., *mini-max* $\{\sigma^2(m,n;p)\}$. It is well-known that such mini-max type objectives can lead to overly conservative solutions, especially under an interval uncertainty set (e.g., [6, 11, 39, 79]), which we also utilize, through the use of the support of P. The regret-based objective in *RM* provides an alternative approach, and can reduce the conservativeness of the mini-max solution (e.g., [39,79,84]), and we consider both formulations in our study. Among the three pool design optimization models that we consider, both robust pool design models, *MM* and *RM*, are novel, while the benchmark model, *DM*, as well as *DM*-variations are commonly utilized in the literature, both for a given number of pools, n, under different objective functions, including the minimization of the asymptotic variance ([58, 69]), maximization of the Fisher's information via the Cramer-Rao lower bound, which reduces to a function of the asymptotic variance ([57,111]), or maximization of the probability of a random pool testing positive ([99]); and for joint pool design optimization ([76]).

Each pool design optimization model can be used both within both the single and two-stage estimation frameworks presented in Section 5.2.2. When used within a two-stage framework, one can use various approaches for updating the input parameters for each stage's pool design model; we discuss the approach we that we use in our case study in Section 5.4.

5.3 Optimal Pool Designs

The optimal solution to Model *DM* has been fully characterized in [76]. Therefore, in this section, we focus on characterizing the optimal solutions to the robust models, *MM* and *RM*. We use some of the properties established in [76] on the asymptotic variance function (see Appendix D.2), as well as establish new structural properties. To this end, we define $\overline{M}(n) \equiv \left\lfloor \frac{B - c_f n}{c_v n} \right\rfloor$, for $n \in \mathbb{Z}^+$, i.e., the maximum feasible pool size for a given a number of pools, n. Observe that, by this definition, $\overline{M}(n)$ is non-increasing in n. We also define, similar to [69] and [76], the *threshold prevalence rate*, $\pi_0(m_1, m_2)$, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, where $\sigma^2(m_1, n; p) = \sigma^2(m_2, n; p)$. From [69] and [76], we have that:

$$\sigma^2(m_1, n^*(m_1); p) \begin{cases} > \sigma^2(m_2, n; p), & \forall p < \pi_0(m_1, m_2) \\ < \sigma^2(m_2, n; p), & \forall p > \pi_0(m_1, m_2) \end{cases},$$

and $\pi_0(m_1, m_2)$ is decreasing in each of m_1 and m_2 (see Appendix D.2).

In the remainder of the paper, we restrict our analysis to the case where $p_{UB} < \frac{1}{2}$; this is the most realistic case for disease surveillance studies, as disease prevalence rates are typically very low.

5.3.1 Characterization of the Optimal Solution to the Mini-max Model

We first characterize the optimal solution to Model *MM*.

Theorem 3. For any given $n \in \mathbb{Z}^+$, $n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, Model *MM* can be equivalently formulated as Model *DM* with $p_0 = p_{UB}$. Thus, an optimal *MM* solution follows a threshold policy:

$$m_M^*(p_{UB}) = \begin{cases} \overline{M}(n), & \text{if} \quad p_{UB} \leq \pi_0(\overline{M}(n) - 1, \overline{M}(n)) \\ \vdots & \\ m+1, & \text{if} \quad \pi_0(m+1, m+2) \leq p_{UB} \leq \pi_0(m, m+1) \\ m, & \text{if} \quad \pi_0(m, m+1) \leq p_{UB} \leq \pi_0(m-1, m) \\ m-1, & \text{if} \quad \pi_0(m-1, m) \leq p_{UB} \leq \pi_0(m-2, m-1) \\ \vdots & \\ 1, & \text{if} \quad \pi_0(1, 2) \leq p_{UB} < 1. \end{cases}$$

For $n > \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, no feasible solution to *MM* exists.

Therefore, for any given $n \in \mathbb{Z}^+$, $n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, the optimal *MM* solution is unique if the upper bound, p_{UB}, does not correspond to a prevalence threshold point ($p_{UB} \neq \pi_0(m, m+1)$, $\forall m \in \mathbb{Z}^+$). There are dual optimal solutions if p_{UB} corresponds to a prevalence threshold point. To obtain the optimal number of pools for *MM*, we use Theorem 3 repeatedly within a search over n, $n \in \mathbb{Z}^+ : n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$. Theorem 3 and Lemma A1 (see Appendix D.2) lead to the following results.

Corollary 8. The *MM* optimal solution, m_M^*, has the following properties:

1. m_M^* is non-increasing in p_{UB}, with $m_M^* = 1$, $\forall p_{UB} \geq \pi_0(1, 2)$.
2. For a given c_v, m_M^* is non-decreasing in c_f.
3. For a given c_f, m_M^* is non-increasing in c_v.

Thus, when the support of the unknown the prevalence rate is expanded to include higher prevalence rate possibilities, testing more pools of smaller size becomes more efficient, in terms of reducing the asymptotic variance. The result then follows because this implies a higher likelihood of having a true-positive specimen

in a testing pool, which, in turn, leads to a higher likelihood of a positive test outcome for any pool. However, a testing outcome of all positive-testing pools leads to a poor prevalence rate of 1, and pool sizes reduce to counteract this. Similarly, when the upper bound of the prevalence rate gets smaller (i.e., p_{UB} decreases), testing fewer pools of larger size becomes more efficient, as this reduces the likelihood of having all negative-testing pools. With regards to the testing cost structure, as the fixed cost of testing (c_f) increases, testing fewer pools of larger size becomes more efficient, while, as the variable cost of testing (c_v) increases, testing more pools of smaller size becomes more efficient.

Next, we compare the optimal solutions to *DM* and *MM*.

Corollary 9. For any problem instance, the optimal solutions to *DM* and *MM* are such that: $m_D^*(p_0) \geq m_M^*(p_{UB})$, and $n_D^*(p_0) \leq n_M^*(p_{UB})$, $\forall p_0 \in [p_{LB}, p_{UB}]$.

Our extensive numerical study indicates that, in general, *MM* also calls for smaller pool sizes than *RM*, i.e., $m_M^* \leq m_R^*$, but a larger number of pools tested, i.e., $n_M^* \geq n_R^*$.

5.3.2 Characterization of the Optimal Solution to the Regret-based Pool Design Model

We next turn our attention to the *Regret-based* model, *RM*. For this purpose, we first characterize the optimal *RM* pool size when the number of pools, n, is fixed. We first study the inner maximization problem, i.e., $\max_{p \in [p_{LB}, p_{UB}]} Regret(m, n; p)$, for a given $(m, n) \in \mathbb{F}(m, n)$. To this end, we first derive the first order condition (FOC):

$$\frac{\partial}{\partial p} Regret(m, n; p) = \frac{1}{n} \left\{ \frac{1}{m^2} \Big[\frac{(m-2)}{(1-p)^{m-1}} + 2(1-p) \Big] - \frac{1}{\big(m_D^*(n,p)\big)^2} \Big[\frac{\big(m_D^*(n,p) - 2\big)}{(1-p)^{m_D^*(n,p)-1}} + 2(1-p) \Big] \right\} = 0, \quad (5.5)$$

where $m_D^*(n, p)$ is the optimal solution to *DM* with $p_0 = p$, and for a given $n \in \mathbb{Z}^+$, $n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$. Observe that any solution to the FOC (see Eqn. (5.5)) is independent of n. We define $\tilde{p}(m) \equiv \arg\min_{p \in (0,1)} \left\{ \frac{\partial}{\partial p} Regret(m, n; p) = 0 \right\}$, $\forall m, n \in \mathbb{Z}^+ : m \geq 2, (m, n) \in \mathbb{F}(m, n)$. By this definition, if $\tilde{p}(m)$ exists, then it is the *smallest* solution to the FOC, and, hence, is independent of n. We then have the following result.

Lemma 12. For any given $m, n \in \mathbb{Z}^+ : m \geq 2, (m, n) \in \mathbb{F}(m, n)$, $Regret(m, n; p)$ function has the following properties:

1. $Regret(m, n; p) = 0$, $\forall p \in \big[\pi_0(m, m+1), \pi_0(m-1, m)\big]$.

2. If $\tilde{p}(m)$ exists, then it is unique, belongs to the interval $\big(0, \pi_0(m, m+1)\big)$, and corresponds to a local maximum.

3. $Regret(m, n; p)$ is strictly increasing in p, $\forall p \in \big(\pi_0(m-1, m), 1\big)$.

As an example, Figure 5.1 demonstrates the behavior of the $Regret(m = 4, n = 1; p)$ function with respect to p. In this case, $\tilde{p}(4)$ exists, and is the unique maximum of the function in the interval $\big(0, \pi_0(4, 5)\big)$; further, $Regret(m = 4, n = 1; p)$ function is strictly increasing in p, $\forall p \in \big(\pi_0(3, 4), 1\big)$.

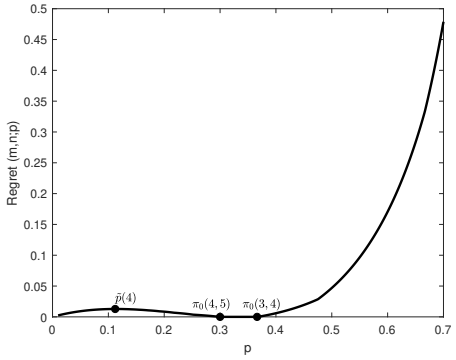


Figure 5.1: $Regret(m = 4, n = 1; p)$ versus p for $c_f = 5$, $c_v = 1$.

Lemma 13. For any given $m \in \mathbb{Z}^+ : 2 \leq m \leq \overline{M}(1)$, the solution to $\left\{ \max_{p \in [p_{LB}, p_{UB}]} \{Regret(m, n; p)\} \right\}$ is attained at one of the points, p_{LB}, p_{UB}, or $\tilde{p}(m)$, $\forall n \in \mathbb{Z}^+$, that is:

$$p^*(m) \equiv \underset{p \in \{p_{LB}, p_{UB}, \tilde{p}(m)\}}{\arg\max} \{Regret(m, n; p)\}.$$

Lemma 13 allows us to compute $p^*(m)$ only once for any $m \in \mathbb{Z}^+ : 2 \leq m \leq \overline{M}(1)$, and to use these values of $p^*(m)$ in the exhaustive search over n, $n \in \mathbb{Z}^+, n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, in order to find (m_R^*, n_R^*). Thus, we have the following result.

Theorem 4. For any n, $n \in \mathbb{Z}^+, n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, use Lemma 13 repeatedly for $m = 1, \cdots, \overline{M}(n)$ to obtain the corresponding objective function value, i.e., $Regret(m, n; p^*(m))$, then:

$$(m_R^*, n_R^*) \equiv \underset{(m,n) \in \mathbb{F}(m,n)}{\arg\min} \Big\{ Regret(m, n; p^*(m)) \Big\}.$$

Theorem 4 enables us to develop an efficient solution algorithm for determining an optimal RM solution.

5.4 Case Study: Prevalence Estimation of West Nile Virus in Mosquitoes

Our goals in this section are: (i) to compare the performance of the robust pool designs obtained by *MM* and *RM* with the benchmark pool design, obtained by *DM* ([76]); and (ii) to compare the single with two-stage estimation frameworks (with various values of λ), with and without robust pool designs, so that we can provide guidelines to practitioners. For this purpose, we apply the proposed robust pool design optimization models, embedded into single and two-stage estimation frameworks, to estimate the prevalence of mosquitoes carrying West Nile virus (WNV).

This section is organized as follows. In Section 5.4.1, we describe the numerical study and data sources. Then, in Section 5.4.2, we present and discuss the numerical results. Additional numerical results are provided in Appendix D.4.

5.4.1 Description of the Numerical Study

WNV-related diseases have become a seasonal endemic in the United States, leading to several fatalities from neuro-invasive diseases ([52, 65]). The primary source of WNV transmission to humans is a mosquito bite ([52]), and, therefore, the prevalence rate of WNV in mosquitoes is shown to be a leading indicator of the prevalence rate of WNV in humans ([31, 36, 64]). The WNV disease in humans is significantly under-reported, because of a lack of specific symptoms, and it can even be asymptomatic ([109]). This is problematic because the WNV disease can be transmitted through blood transfusion or organ transplantation (e.g., [60, 90]). Therefore, an accurate estimation of the prevalence rate of WNV-carrying mosquitoes is essential for outbreak prediction and disease prevention ([52, 53]). The WNV disease fits well with our models, because its prevalence rate in both mosquitoes and humans is highly seasonal, with substantial fluctuation from year to year, and in different regions ([39, 52, 65]).

In our study, we consider the reverse-transcription polymerase chain reaction (RT-PCR) assay for WNV screening in mosquitoes ([53]). RT-PCR assay detects the viral RNA present in the specimens using the nucleic acid amplification technology, and is highly accurate ([77]). All data used in our numerical study come from published studies, and we complement these data with various sensitivity analyses; see Table 5.1 for the parameter values used.

In our numerical study, we use three different testing budgets, $B = \{\$8,160, \$16,320, \$65,280\}$, which, given the testing cost parameters from [53], respectively correspond to pool designs that test 30, 60, and 240 pools of mosquito specimens, each of size 50 ([81, 83]). These pool designs are highly relevant to our

Table 5.1: Data and sources for the numerical study

	Parameters (Source)
c_f	\$72 ([53])
c_v	\$4 ([53])
B	$\{\$8,160,\$16,320,\$65,280\}$ (e.g., [53, 83])
p_{LB}	0.003 ([81])
p_{UB}	0.09 ([34])

Model Input		**Simulation Input**	
		Accurate Input Setting	Inaccurate Input Settings
DM:	$p_0 = \frac{1}{2}\left(p_{LB} + p_{UB}\right)$ $= 0.0465$	$P \sim$ Uniform (0.003, 0.09); $\mu_P = 0.0465$;	$P \sim$ Beta (1.59, 32.56); $\mu_P = 0.0465$; $P \in (0,1)$
MM, *RM*:	$P \in [p_{LB}, p_{UB}]$ $= [0.003, 0.09]$	$P \in (0.003, 0.09)$	$P \sim$ Beta (3.52, 46.92); $\mu_P = 0.0698$; $P \in (0,1)$ $P \sim$ Beta (3.52, 46.92); $\mu_P = 0.0233$; $P \in (0,1)$

study, as testing pools of size 50 are used commonly in studies that estimate the prevalence rate of WNV in mosquitoes via RT-PCR assays (e.g., [48, 49, 73, 81, 83]).

We derive the support of the unknown prevalence rate, P, from WNV surveillance data from various parts of the Mid-South region of the United States, where WNV disease was the most prevalent during the 2002-2003 outbreak ([52]). As the prevalence rate of WNV-infected mosquitoes can be as high as 8.76% (Tennessee Valley between 2002 and 2005 ([34]), we use an upper bound of 9%. We use a lower bound of 0.3%, which has been used in previous WNV prevalence studies in mosquitoes (e.g., [81, 83]). Throughout, we assume that the tester estimates the support of P as $P \in [0.003, 0.09]$ (which is used as an input for *MM* and *RM*), and estimates the mean of P as $\mu_P = 0.0465$ (which is used as an input for *DM*, i.e., $p_0 = \mu_P$). Observe that the distribution of P is not needed for any of the models, *MM*, *RM*, and *DM*. To determine the performance of each model, we perform a Monte Carlo simulation, as detailed below, in which we assume a certain distribution of P; see Table 5.1. We first study an *accurate input setting* in which the tester estimates both the support and the first moment of P accurately (i.e., the setting with simulation inputs of $P \sim$ Uniform (0.003, 0.09); see Table 5.1). Then, we study the more realistic, *inaccurate input settings*, where the tester does not estimate the support and the first moment of P accurately (i.e., the settings with simulation inputs of $P \sim$ Beta (α, β), with various values of α and β; see Table 5.1). For this purpose, we define a *scenario* by an information setting (accurate input setting, or one of the inaccurate input settings) and a testing budget, B ($B \in \{\$8, 160, \$16, 320, \$65, 280\}$).

We perform a Monte Carlo simulation, with 20,000 replications for each scenario. Specifically, for each scenario, we first determine the optimal pool designs for the various models based on the model inputs provided in Table 5.1. Each simulation replication corresponds to a randomly generated realization of P, denoted by p, from a specified *true* distribution of P (see the simulation inputs in Table 5.1). Based on

the generated value of p, we then randomly generate the carrier status of each subject (specimen) (a total of $m^{(s)*} \times n^{(s)*}(m^{(s)*})$ specimens in each stage $s = 1, \cdots, S$, for each estimation framework), where each specimen has the WNV disease with probability p, and is disease-free with probability $1-p$. These specimens are then randomly assigned to the testing pools. If a pool contains at least one infected specimen, then the test outcome for the pool will be positive; and otherwise, the test outcome for the pool will be negative. Given a set of test outcomes, we then compute the MLE of P using Eqn. (5.1), i.e., $\hat{p}^{(1)}$. If $\lambda = 1$, i.e., in a single-stage estimation framework, we stop at this step and compute the performance metrics, as discussed below. If $0 < \lambda < 1$, i.e., in a two-stage estimation framework, we update the input for the optimization models according to $\hat{p}^{(1)}$, the outcome of stage 1. In particular, for Model *DM*, we update p_0 to be $\hat{p}^{(1)}$. For the robust models *MM* and *RM*, we estimate the 95% confidence interval of $\hat{p}^{(1)}$ using its asymptotic variance, i.e., we compute $\sigma^2(m^{(1)*}, n^{(1)*}; \hat{p}^{(1)})$ using Eqn. (5.2). Then, we let $[p_{LB} = \hat{p}^{(1)} - 1.96\sigma(m^{(1)*}, n^{(1)*}; \hat{p}^{(1)})$, $p_{UB} = \hat{p}^{(1)} + 1.96\sigma(m^{(1)*}, n^{(1)*}; \hat{p}^{(1)})]$. We then repeat the pool design optimization and testing steps in order to compute the final prevalence rate estimate $\hat{p}^{(2)}$ using Eqn. (5.1).

We compute the following performance metrics, which are commonly used in the statistics literature to evaluate the efficiency of an estimator (e.g., [57, 58, 111]):

1. The prevalence estimate, $\hat{p}$ (see Eqn. (5.1))
2. The final asymptotic variance, $\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$ (see Eqn. (5.3)), where $(\mathbf{m}^*, \mathbf{n}^*) \equiv \{(m^{(1)*}, n^{(1)*}), (m^{(2)*}, n^{(2)*})\}$ for two-stage estimation frameworks ($0 < \lambda < 1$), and $(\mathbf{m}^*, \mathbf{n}^*) \equiv (m^{(1)*}, n^{(1)*})$ for single-stage estimation frameworks ($\lambda = 1$)
3. The mean squared error of the prevalence estimate: $MSE = (\hat{p} - p)^2$
4. The percent relative bias of the prevalence estimate: $rBias(\%) = 100 \times \left|\frac{\hat{p}-p}{p}\right|$.

In each of the following tables, we report the performance metrics in the form: *average ± half width of 95% confidence interval (CI)*, based on 20,000 replications.

5.4.2 Numerical Results

In our numerical study, we consider both single and two-stage estimation frameworks, i.e., with $\lambda \in \{0.25, 0.5, 1\}$, for both the accurate and inaccurate input settings (Table 5.1). In all cases, the tester determines the "optimal" pool designs based on the assumed support of P, i.e., $P \in [0.003, 0.09]$ (for *MM* and *RM*), or based on the assumed point estimate of P, i.e., $p_0 = 0.0465$ (for *DM*). Then, in the simulation, P is generated from Uniform (0.003, 0.09) in the accurate input setting, and from Beta(α, β), with support of $[0, 1]$, and for varying values of α and β, in the inaccurate input settings; see Table 5.1.

Table 5.2 displays our results for the scenario with B=\$8,160 and accurate input parameters, i.e., $P \sim$ Uniform (0.003, 0.09), for a single-stage estimation framework, i.e., with $\lambda = 1$.

Table 5.2: Comparison of *DM, MM* and *RM* for a single-stage estimation framework ($\lambda = 1$) with $B = \$8,160$ and accurate input parameters (average ± half width of 95% CI)

True Distribution and Moments of P	**Model** (Input Parameters)	***DM*** (p_0=0.0465)	***MM*** ($P \in$ (0.003, 0.09))	***RM*** ($P \in$ (0.003, 0.09))
	(m^*, n^*)	(19,55)	(12,68)	(13,65)
Uniform $\sim$	$\hat{p}$	0.04750 ±0.00038	0.04706 ±0.00038	0.04693 ±0.00037
(0.003, 0.09)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)[\times 10^6]$	80.40 ± 0.81	78.10 ± 0.69	77.90 ± 0.70
$\mu_P = 0.0465$	$MSE[\times 10^6]$	88.20 ± 2.89	81.70 ± 2.20	81.30 ± 2.19
$\sigma_P^2 = 0.000675$	$rBias(\%)$	16.64 ±0.21	17.58 ±0.23	17.33 ±0.22

As observed in Table 5.2, when input parameters are accurate, *DM*, *MM*, and *RM* perform well with respect to all performance metrics, even at the lowest budget level of $B = \$8,160$, and even in a single-stage estimation framework, i.e., $\lambda = 1$. However, we note that the robust models *MM* and *RM* yield slightly more efficient pool designs in comparison to *DM*, with respect to minimizing the estimation error.

In the remainder of this section, we focus on the scenarios with inaccurate input settings, i.e., $P \sim$ Beta (α, β), with support of $[0, 1]$, and varying values of α and β (see Table 5.1), in single and two-stage estimation frameworks, i.e., with $\lambda \in \{0.25, 0.5, 1\}$; see Table 5.3. For two-stage frameworks, the pool designs in stage 2 are random. Therefore, in Table 5.3, we only report the optimal pool designs in stage 1, $(m^{(1)*}, n^{(1)*})$. Since $(m^{(1)*}, n^{(1)*})$ is determined based on the assumed support and first moment of P, the pool designs reported in Table 5.3 remain the same for each model. To facilitate the presentation, the results for scenarios with $B \in \{\$16,320, \$65,280\}$ are given in Appendix D.4.

Not surprisingly, when input parameters are not accurate, all models produce worse results; however, *RM* and *MM* designs are much more robust, with *MM* design having the lowest degradation due to inaccurate parameters. For instance, when the true distribution of P is Beta $\sim$ (1.59, 32.56), *MM* produces, on average, a better estimate of P compared to *DM* (0.04739 vs. 0.05152), and reduces the asymptotic variance by 22.93%, MSE by 92.04%, and relative bias by 7.89%. Observe that, in this case, the input parameter for *DM* is accurate (p_0=μ_P), which is not always realistic. The robust models continue to perform better than the other models (with *MM* performing the best) when the true mean is higher or lower than the assumed mean, see Table 5.1 and Appendix D.4.

From Table 5.3, we also see that, while the robustness of Model *DM* can be significantly improved by using the model within the two-stage estimation framework, e.g., with DM-S ($\lambda = 0.5$), MM-S ($\lambda = 0.5$) does not significantly improve the robustness of Model *MM* ($\lambda = 1$) in all cases of inaccurate input parameters we study. Further, the estimation errors of DM-S and MM-S, both with $\lambda = 0.5$, are also not significantly lower than those of RM-S ($\lambda = 1$). We also note that, while not substantial, MM-S ($\lambda = 0.5$) has the best

Table 5.3: Comparison of *DM, MM* and *RM* for two-stage estimation frameworks ($\lambda \in \{0.25, 0.5, 1\}$) with $B = \$8,160$ and inaccurate input parameters; $\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$ and MSE values are multiplied by 10^6 (average ± half width of 95% CI)

	Model	***DM-S*** ($\lambda = 1$)	***MM-S*** ($\lambda = 1$)	***RM-S*** ($\lambda = 1$)	***DM-S*** ($\lambda = 0.5$)	***MM-S*** ($\lambda = 0.5$)	***DM-S*** ($\lambda = 0.25$)	***MM-S*** ($\lambda = 0.25$)
	(m^*, n^*)	(19,55)	(12,68)	(13,65)	(18,28)	(12,34)	(18,14)	(12,17)
Beta ∼	$\hat{p}$	0.05152 ± 0.00103	0.04739 ±0.00056	0.04686 ±0.00058	0.04840 ± 0.00169	0.04777 ± 0.00163	0.04815 ±0.00169	0.04847 ±0.00170
(1.59, 32.56)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	116 ± 4.61	89.40 ± 1.87	89.90 ± 1.92	97.00 ± 19.20	87.40 ± 11.70	102 ± 19.00	92.40 ± 14.20
$\mu_P = 0.0465$	MSE	$3,230 \pm 632$	257 ± 114	268 ± 136	121 ± 6.98	96.50 ± 4.80	116 ± 6.93	106 ± 5.05
$\sigma_P^2 = 0.00135$	$rBias(\%)$	20.53 ± 0.50	18.91 ±0.28	18.94 ±0.27	18.06 ± 0.82	18.10 ± 0.81	19.41 ±1.20	19.50 ±0.89
Beta ∼	$\hat{p}$	0.07853 ± 0.00130	0.07092 ±0.00056	0.07071 ±0.00057	0.07009 ± 0.00168	0.07059 ± 0.00169	0.07221 ±0.00175	0.06923 ±0.00155
(3.52, 46.92)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	192 ± 4.51	143 ± 1.82	146 ± 2.00	154 ± 28.78	141 ± 14.20	171 ± 31.50	138 ± 15.80
$\mu_P = 0.0698$	MSE	$5,990 \pm 865$	257 ± 114	287 ± 123	173 ± 6.85	151 ± 5.24	225 ± 9.07	150 ± 4.71
$\sigma_P^2 = 0.00135$	$rBias(\%)$	18.62 ± 0.63	14.05 ±0.18	14.12 ±0.18	13.83 ± 0.51	12.02 ± 1.52	14.86 ±0.55	13.96 ±0.49
Beta ∼	$\hat{p}$	0.02789 ±0.00114	0.02414 ±0.00065	0.02433 ±0.00065	0.02259 ± 0.00149	0.02395 ± 0.00164	0.02354 ±0.00169	0.02342 ±0.00164
(0.40,16.61)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	126 ± 54.94	51.20 ± 3.28	56.50 ± 9.46	43.50 ± 14.00	47.50 ± 9.38	58.30 ± 19.30	46.50 ± 10.30
$\mu_P = 0.0233$	MSE	$3,690 \pm 655$	531 ± 220	578 ± 234	516 ± 5.29	465 ± 5.89	698 ± 8.79	498 ± 5.48
$\sigma_P^2 = 0.00135$	$rBias(\%)$	59.73 ±5.03	59.59 ±7.71	55.47 ± 2.39	72.11 ± 29.42	48.33 ± 3.21	59.13 ±1.65	50.85 ±4.19

performance, in terms of minimizing estimation errors, of all models we consider. However, this difference diminishes as the budget level, B, increases; see Appendix D.4. With respect to the value of λ, using $\lambda = 0.25$ can lead to more efficient pool designs, for both DM-S and MM-S, when B is sufficiently high (see, e.g., Table D.2 and D.3), in comparison to $\lambda = 0.5$. When B is low, i.e., $B = \$8,160$, using $\lambda = 0.25$ can lead to inefficient pool designs, as the budget for stage 1 is not sufficient, leading to inaccurate updating of model inputs for pool design optimization in stage 2; see Table 5.3. Therefore, when using a two-stage estimation framework, i.e., when $\lambda < 1$, λ needs to be set in accordance with the total testing budget, B, which is a considerable challenge in implementing the two-stage estimation framework.

5.5 Discussion

In this paper, we develop and study robust pool design optimization models for prevalence estimation under limited resources, within single-stage and two-stage estimation frameworks. Our models are quite general, and enable us to relax various restrictive assumptions common in the existing literature, including the use of a fixed number of pools and a point estimate as model inputs. Further, our robust models have an important advantage over the models studied in the literature, in that they require only the support of the unknown prevalence rate, which is relatively easy to estimate, in comparison to estimating the distribution or point estimate of the unknown prevalence rate. More importantly, our analysis shows that the proposed robust models provide "good" testing designs even when the estimated support is inaccurate. This is not the case for the pool design models studied in the literature that are highly dependent on an initial point estimate of the prevalence rate, in that these models can yield poor solutions if the point estimate is not accurate. From that perspective, our robust models apply especially well to prevalence estimation of emerging or seasonal diseases, such as Zika or West Nile virus disease, for which initial information, prior to testing, is often highly unreliable. Further, our numerical study indicates that the robust pool design models perform well even within a single-stage estimation framework, compared to a two-stage estimation framework with robust pool designs. We also establish key structural properties of optimal pool designs and completely characterize optimal pool designs for various problem settings, which impose different objectives on the pool design problems.

Our case study, on estimating the prevalence of West Nile virus in mosquitoes, compares our robust models, against the benchmark deterministic model from the literature, within single-stage and two-stage estimation frameworks. There are several important findings from this study. When estimating the prevalence of emerging and seasonal diseases, the distribution and support of P are highly uncertain at the outset, and thus input parameters are likely to be inaccurate. It is in these realistic cases that both robust models

(*MM* and *RM*) perform significantly better than their deterministic counter-part, with *DM* having the worst performance within a single-stage estimation framework. In these cases of inaccurate inputs, *DM*, when used within a two-stage estimation framework, i.e., DM-S, with $\lambda = 0.5$, has significantly improved performance, while *MM*, used within a two-stage estimation framework, i.e., MM-S, with $\lambda < 1$, still yields the best performance in all cases of inaccurate input parameters and across all budget levels. However, in comparison to the single-stage robust models (MM-S and RM-S with $\lambda = 1$), the improvement in estimation accuracy provided by the two-stage frameworks (DM-S and MM-S with $\lambda < 1$) is not quite as substantial as the improvement provided by the robust models, in comparison to the deterministic model, within a single-stage framework. Therefore, given the operational challenges of the two-stage estimation framework discussed in Section 5.1, the single-stage estimation framework with our robust models, either *MM* and *RM*, can be a much more desirable approach to surveillance studies of emerging and seasonal diseases in large scale. It is important to note that the robust models, in both single-stage and two-stage frameworks, outperform other deterministic models, while requiring minimal information prior to testing, and perform consistently well even when this information is not perfectly accurate. These findings have important implications for designing surveillance studies.

As immediate, and important, extensions of our study, one can relax some of our modeling assumptions, including that the screening test is perfectly reliable, and that the subjects are independent. We consider prevalence estimation for a single disease in a specific region. In practice, healthcare policy-makers may need to allocate their testing budget among the prevalence estimation activities for a number of diseases in a specific region, or in various regions, each potentially having a different prevalence rate of the disease in question. Therefore, extending our models to study pool design optimization and budget allocation for prevalence estimation for multiple diseases or multiple regions is an important research direction. Additionally, practitioners may need to select a screening test, among a set of commercially available tests, including combo tests, i.e., tests that can simultaneous detect a number of diseases. Thus, another interesting future research direction is to study the models for test selection and pool design optimization for prevalence estimation of multiple diseases. We believe that such models will further enhance the accuracy of disease prevalence estimation and lead to efficient budget allocation and utilization for surveillance study of emerging and seasonal diseases.

Chapter 6

Conclusions

Prevalence estimation of emerging and seasonal diseases is an important input for various functions of public health, epidemiology, and healthcare system management. Consequently, designing efficient testing pools for prevalence estimation is also an important decision for policy-makers and practitioners working in these areas. The testing pool design problem, studied in this dissertation, requires a highly uncertain input (an initial estimate of the unknown prevalence rate), and is highly constrained by limited testing budgets and resources. Despite this, not much emphasis has been placed on addressing these issues in the existing literature. In our study, we develop and study novel frameworks for optimal pool design for prevalence estimation, with a focus on emerging and seasonal diseases. We relax several restricting assumptions commonly used in the relevant literature, in order to establish and provide insight and guidelines into optimal pool designs for prevalence estimation under limited testing budgets, and uncertain, and potentially inaccurate, model input parameters.

In particular, we develop a sequential and adaptive estimation procedure that directly utilizes the continuous outcomes of the pooled tests, detailed in Chapter 2. In order to facilitate our analysis of the sequential estimation procedure, we also develop a novel methodology for estimating the sensitivity of pooled testing, using viral load progression models and explicitly accounting for the dilution effect of pooling, outlined in Chapter 3. Our numerical studies on HIV prevalence estimation using the proposed sequential estimation procedure and the proposed sensitivity estimation methodology indicate that our models can lead to highly efficient pool designs for prevalence estimation, and, hence, to accurate and robust prevalence estimates even when model input parameters deviate from the initial assumptions. The proposed sequential estimation procedure can also be useful in surveillance studies of insect-borne diseases in plants and crops, as demonstrated in the case study on prevalence estimation of the Tomato Spotted Wilt virus in thrips.

While the use of continuous test outcomes can improve the accuracy of the prevalence rate estimate, a common practice in surveillance studies of diseases is to use binary test outcomes to estimate the unknown prevalence rate. Therefore, in Chapters 4 and 5, we study the testing pool design optimization problem under binary test outcomes. Specifically, in Chapter 4, we compare two deterministic models for pool design: with an exogenously fixed number of pools, and with an optimally set number of pools, both under a limited testing budget. In order to solve these models to optimality, we establish key structural properties of the asymptotic variance function, and fully characterize the optimal solutions for the two deterministic models. Our numerical study, on the prevalence estimation of West Nile virus in mosquitoes, indicates that jointly optimizing over both the pool size and the number of pools leads to highly efficient pool designs, in terms of minimizing the estimation error, compared to pool designs restricted by a fixed number of pools as well as other commonly used pool designs.

In Chapter 5, we further extend our study on optimal pool design, and use robust optimization to hedge against the uncertainty in model input parameters. In particular, we develop a mini-max model and a regret-based model for pool design optimization, both of which require only the support of the unknown prevalence rate. It is important to note that both robust models are distribution-free, and as such, they do not require any distribution or moment information about the unknown prevalence rate, which can be very difficult to estimate accurately. Our numerical study, on the prevalence estimation of West Nile virus in mosquitoes, suggests that the robust models substantially outperform their deterministic counterparts, and continue to yield efficient pool designs (with accurate prevalence estimates) even when their input parameters are inaccurate. More interestingly, the robust models also generate pool designs that are as efficient, in terms of minimizing the estimation error, as those obtained from a sequential estimation framework with a deterministic pool design model. From this perspective, in the context of prevalence estimation, robust pool design optimization models, within single-stage estimation frameworks, can reap almost all benefits of sequential frameworks, which are more difficult to implement; and these benefits are realized even when there is little information on the current status of the disease prior to testing. These findings have important implications, especially for the prevalence estimation of the sources of emerging and/or seasonal diseases, e.g., prevalence estimation of West Nile virus in mosquitoes.

In conclusion, our research underscores the value of optimization methodologies for efficient testing pool design for prevalence estimation. This research also quantifies the value of robust optimization, as well as sequential and adaptive approaches, in comparison to other commonly used testing practices, especially for prevalence estimation of emerging and seasonal diseases. In addition, Chapters 2 and 3 show that using continuous test outcomes, as opposed to binary test outcomes, can lead to substantial improvement in estimation accuracy. Thus, an immediate and important extension of this research is to develop and

study robust pool design optimization models considering continuous test outcomes; we believe that this will further enhance the estimation efficiency of the proposed models. We hope that our novel approaches to testing pool design for prevalence estimation will lead to further research on the topic, and our insights and principles on testing pool design will impact the current practices in surveillance studies for emerging and seasonal diseases.

Bibliography

[1] E. R. Adrion, J Aucott, K. W. Lemke, and J. P. Weiner. Health Care Costs, Utilization and Patterns of Care Following Lyme disease. *PLoS One*, 10(2):e0116767, 2015.

[2] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] American Red Cross. *Blood Testing*, accessed May 27, 2017. `http://www.redcrossblood.org/learn-about-blood/what-happens-donated-blood/blood-testing`.

[4] H. Aprahamian, D. R. Bish, and E. K. Bish. Residual Risk and Waste in Donated Blood with Pooled Nucleic Acid Testing. *Statistics in Medicine*, 35(28):5283–5301, 2016.

[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] I. Averbakh and Y. Zhao. Explicit Reformulations for Robust Optimization Problems with General Uncertainty Sets. *SIAM Journal on Optimization*, 18(4):1436–1466, 2008.

[7] AVERT. *HIV and AIDS in East and Southern Africa Regional Overview*, accessed July 19, 2017. `https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview`.

[8] AVERT. *HIV and AIDS in West and Central Africa Overview*, accessed July 19, 2017. `https://www.avert.org/hiv-and-aids-west-and-central-africa-overview`.

[9] A.D.T. Barrett. Economic Burden of West Nile virus in the United States. *The American Journal of Tropical Medicine and Hygiene*, 90(3):389, 2014.

[10] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, 2011.

[11] D. Bertsimas and M. Sim. The Price of Robustness. *Operations Research*, 52(1):35–53, 2004.

[12] C. R. Bilder, B. Zhang, F. Schaarschmidt, and J. M. Tebbs. binGroup: A Package for Group Testing. *The R Journal*, 2(2):56, 2010.

[13] E. K. Bish, P. K. Ragavan, D. R. Bish, A. D. Slonim, and S. L. Stramer. A probabilistic method for the estimation of residual risk in donated blood. *Biostatistics*, 15(4):620–635, 2014.

[14] R. Biswas, E. Tabor, C. C. Hsia, D. J. Wright, M. E. Laycock, E. W. Fiebig, L. Peddada, R. Smith, G. B. Schreiber, J. S. Epstein, G. J. Nemo, and M. P. Busch. Comparative sensitivity of HBV NATs and HBsAg assays for detection of acute HBV infection. *Transfusion*, 43(6):788–798, 2003.

[15] R. Brookmeyer. Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence. *Biometrics*, 55(2):608–612, 1999.

[16] P. M. Burrows. Improved Estimation of Pathogen Transmission Rates by Group Testing. *Phytopathology*, 77(2):363–365, 1987.

[17] M. P. Busch. HIV, HBV and HCV: New developments related to transfusion safety. *Vox Sanguinis*, 78:253–256, 2000.

[18] G. Casella and R. L. Berger. Statistical Inference. Volume 2, chapter 7, page 335. Duxbury Pacific Grove, CA, 2002.

[19] Centers for Disease Control and Prevention. Diagnoses of HIV Infection in the United States and Dependent Areas, 2014. *HIV Surveillance Report*, 26:5–11, 2015.

[20] Centers for Disease Control and Prevention. *Geographic Distribution of Ticks that Bite Humans*, accessed January 10, 2019. `https://www.cdc.gov/ticks/geographic_distribution.html`.

[21] Centers for Disease Control and Prevention. *CDC Awards Nearly $184 Million to Continue the Fight against Zika*, accessed June 11, 2018. `https://www.cdc.gov/media/releases/2016/p1222-zika-funding.html`.

[22] Centers for Disease Control and Prevention. *Multistate Outbreak of E. coli O157:H7 Infections Linked to Romaine Lettuce*, accessed June 14, 2018. `https://www.cdc.gov/ecoli/2018/o157h7-04-18/index.html`.

[23] Centers for Disease Control and Prevention. *HIV Testing*, accessed November 20, 2018. `https://www.cdc.gov/hiv/testing/index.html`.

[24] Centers for Disease Control and Prevention. *Final Annual Maps & Data for 1999-2016*, accessed September 12, 2018. `https://www.cdc.gov/westnile/statsmaps/finalmapsdata/index.html`.

[25] Y.P. Chaubey and W. Li. Estimation of Fraction Defectives Through Batch Sampling. *Proc. Qual. Productvty Sect. Am. Statist Ass*, pages 198–206, 1993.

[26] C. L. Chen and W. H. Swallow. Using Group Testing to Estimate a Proportion, and to Test the Binomial Model. *Biometrics*, 46(4):1035–1046, 1990.

[27] C. L. Chen and W. H. Swallow. Sensitivity Analysis of Variable-Size Group Testing and its Related Continuous Models. *Biometrical Journal*, 37(2):173–181, 1995.

[28] P. Chen, J. M. Tebbs, and C. R. Bilder. Group Testing Regression Models with Fixed and Random Effects. *Biometrics*, 65(4):1270–1278, 2009.

[29] J.J. Cho, R.F.L. Mau, R.T. Hamasaki, and D. Gonsalves. Detection of Tomato Spotted Wilt Virus in Individual Thrips by Enzyme-Linked Immunosorbent Assay. *Phytopathology*, 78(10):1348–1352, 1988.

[30] J.J. Cho, W.C. Mitchell, R.F.L. Mau, and K. Sakimura. Epidemiology of Tomato Spotted Wilt Virus Disease on Crisphead Lettuce in Hawaii. *Plant Disease*, 71(6):505–508, 1987.

[31] W. M. Chung, C. M. Buseman, S. N. Joyner, S. M. Hughes, T. B. Fomby, J. P. Luby, and R. W. Haley. The 2012 West Nile Encephalitis Epidemic in Dallas, Texas. *JAMA*, 310(3):297–307, 2013.

[32] CNN. *Exploding Tick Population – and Illnesses They Bring – Worries Government*, accessed May 07, 2019. `https://www.cnn.com/2018/11/14/health/tick-report-hhs-bn/index.html`.

[33] Consumers Union. *Outrageous Health Costs: Blood Test*, accessed November 02, 2016. `http://consumersunion.org/outrageous-health-costs/blood-test/`.

[34] E. W. Cupp, H. K. Hassan, X. Yue, W. K. Oldland, B. M. Lilley, and T. R. Unnasch. West Nile Virus Infection in Mosquitoes in the Mid-South USA, 2002–2005. *Journal of Medical Entomology*, 44(1):117–125, 2007.

[35] J. R. Davis, J. Lederberg, et al. Public Health Systems and Emerging Infections: Assessing the Capabilities of the Public and Private sectors. Workshop Summary. In *Public Health Systems and Emerging Infections: Assessing the Capabilities of the Public and Private sectors. Workshop Summary.* National Academy Press, 2000.

[36] N. B. DeFelice, E. Little, S. R. Campbell, and J. Shaman. Ensemble Forecast of Human West Nile Virus Cases and Mosquito Infection Rates. *Nature Communications*, 8:14592, 2017.

[37] S. Deo, K. Rajaram, S. Rath, U. S. Karmarkar, and M. B. Goetz. Planning for HIV Screening, Testing, and Care at the Veterans Health Administration. *Operations Research*, 63(2):287–304, 2015.

[38] R. Dorfman. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[39] H. El-Amine, E. K. Bish, and D. R. Bish. Robust Postdonation Blood Screening under Prevalence Rate Uncertainty. *Operations Research*, 66(1):1–17, 2018.

[40] X. Fang, W. W. Stroup, and S. Zhang. Improved Empirical Bayes Estimation in Group Testing Procedure for Small Proportions. *Communications in Statistic–Theory and Methods*, 36(16):2937–2944, 2007.

[41] C. P. Farrington. Estimating Prevalence by Group Testing Using Generalized Linear Models. *Statistics in Medicine*, 11(12):1591–1597, 1992.

[42] E. W. Fiebig, D. J. Wright, B. D. Rawal, P. E. Garrett, R. T. Schumacher, L. Peddada, C. Heldebrant, R. Smith, A. Conrad, S. H. Kleinman, and M. P. Busch. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS*, 17(13):1871–1879, 2003.

[43] Food and Drug Administration. *Procleix Ultrio Assay*, accessed November 19, 2018. `https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-bio-gen/documents/document/ucm335285.pdf`.

[44] J. J Gart. An Application of Score Methodology: Confidence Intervals and Tests of Fit for One-hit Curves. *Handbook of Statistics*, 8:395–406, 1991.

[45] J. L. Gastwirth and P. A. Hammick. Estimation of the Prevalence of a Rare Disease, Preserving the Anonymity of the Subjects by Group Testing: Application to Estimating the Prevalence of AIDS Antibodies in Blood Donors. *Journal of Statistical Planning and Inference*, 22(1):15–27, 1989.

[46] J. L. Gastwirth and W. O. Johnson. Screening with Cost-Effective Quality Control: Potential Applications to HIV and Drug Testing. *Journal of the American Statistical Association*, 89(427):972–981, 1994.

[47] S. A. Glynn, D. J. Wright, S. H. Kleinman, D. Hirschkorn, Y. Tu, C. Heldebrant, R. Smith, C. Giachetti, J. Gallarda, and M. P. Busch. Dynamics of viremia in early hepatitis C virus infection. *Transfusion*, 45(6):994–1002, 2005.

[48] W. Gu, T. R. Unnasch, C. R. Katholi, R. Lampman, and R. J. Novak. Fundamental Issues in Mosquito Surveillance for Arboviral Transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(8):817–822, 2008.

[49] T. L. Hadfield, M. Turell, M.P. Dempsey, J. David, and E.J Park. Detection of West Nile Virus in mosquitoes by RT-PCR. *Molecular and Cellular Probes*, 15(3):147–150, 2001.

[50] P. A. Hammick and J. L. Gastwirth. Group Testing for Sensitive Characteristics: Extension to Higher Prevalence Levels. *International Statistical Review/Revue Internationale de Statistique*, pages 319–331, 1994.

[51] T. E. Hanson, W. O. Johnson, and J. L. Gastwirth. Bayesian Inference for Prevalence and Diagnostic Test Accuracy Based on Dual-Pooled Screening. *Biostatistics*, 7(1):41–57, 2006.

[52] E. B. Hayes, N. Komar, R. S. Nasci, S. P. Montgomery, D. R. O'Leary, and G. L. Campbell. Epidemiology and Transmission Dynamics of West Nile Virus Disease. *Emerging Infectious Diseases*, 11(8):1167, 2005.

[53] J. M. Healy, W. K. Reisen, V. L. Kramer, M. Fischer, N. P. Lindsey, R. S. Nasci, P. A. Macedo, G. White, R. Takahashi, L. Khang, et al. Comparison of the Efficiency and Cost of West Nile Virus Surveillance Methods in California. *Vector-Borne and Zoonotic Diseases*, 15(2):147–155, 2015.

[54] G. Hepworth and R. Watson. Debiased Estimation of Proportions in Group Testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):105–121, 2009.

[55] X. Huang. An Improved Test of Latent-Variable Model Misspecification in Structural Measurement Error Models for Group Testing Data. *Statistics in Medicine*, 28(26):3316–3327, 2009.

[56] J. M. Hughes-Oliver. Pooling Experiments for Blood Screening and Drug Discovery. In A. Dean and S. Lewis, editors, *Screening*, chapter 3, pages 48–68. Springer, 2006.

[57] J. M. Hughes-Oliver and W. F. Rosenberger. Efficient Estimation of the Prevalence of Multiple Rare Traits. *Biometrika*, 87(2):315–327, 2000.

[58] J. M. Hughes-Oliver and W. H. Swallow. A Two-Stage Adaptive Group-Testing Procedure for Estimating Small Proportions. *Journal of the American Statistical Association*, 89(427):982–993, 1994.

[59] M. Hung and W. H. Swallow. Robustness of Group Testing in the Estimation of Proportions. *Biometrics*, 55(1):231–237, 1999.

[60] M. Iwamoto, D. B. Jernigan, A. Guasch, M. J. Trepka, C. G. Blackmore, W. C. Hellinger, S. M. Pham, S. Zaki, R. S. Lanciotti, S. E. Lance-Parker, et al. Transmission of West Nile Virus from an Organ Donor to Four Transplant Recipients. *New England Journal of Medicine*, 348(22):2196–2203, 2003.

[61] Johns Hopkins Medicine. *Blood Transfusions Still Overused and May Do More Harm Than Good in Some Patients*, accessed May 07, 2019. https://www.hopkinsmedicine.org/news/media/releases/blood_transfusions_still_overused_and_may_do_more_harm_than_good_in_some_patients.

[62] Johns Hopkins Medicine. *The Johns Hopkins Hospital Estimated Average Charges for Common Procedures*, accessed May 07, 2019. https://www.hopkinsmedicine.org/the_johns_hopkins_hospital/_docs/jhh_charges.pdf.

[63] M. Y. Karris, C. M. Anderson, S. R. Morris, D. M. Smith, and S. J. Little. Cost Savings Associated with Testing of Antibodies, Antigens, and Nucleic Acids for Diagnosis of Acute HIV Infection. *Journal of Clinical Microbiology*, 50(6):1874–1878, 2012.

[64] A. M. Kilpatrick and W. J. Pape. Predicting Human West Nile Virus Infections with Mosquito Surveillance Data. *American Journal of Epidemiology*, 178(5):829–835, 2013.

[65] C. T. Korves, S. J. Goldie, and M. B. Murray. Cost-effectiveness of Alternative Blood-screening Strategies for West Nile Virus in the United States. *PLoS Medicine*, 3(2):e21, 2006.

[66] C. T. Le. A New Estimator for Infection Rates Using Pools of Variable Size. *American Journal of Epidemiology*, 114(1):132–136, 1981.

[67] J. D. C. Little, K. G. Murty, D. W. Sweeney, and C. Karel. An Algorithm for the Traveling Salesman Problem. *Operations Research*, 11(6):972–989, 1963.

[68] E. Litvak, X. M. Tu, and M. Pagano. Screening for The Presence of a Disease by Pooling Sera Samples. *Journal of the American Statistical Association*, 89(426):424–434, 1994.

[69] A. Liu, C. Liu, Z. Zhang, and P. S. Albert. Optimality of Group Testing in the Presence of Misclassification. *Biometrika*, 99(1):245–251, 2012.

[70] S.C. Liu, K.S. Chiang, C.H. Lin, W.C. Chung, S.H. Lin, and T.C. Yang. Cost Analysis in Choosing Group Size when Group Testing for Potato Virus Y in the Presence of Classification Errors. *Annals of Applied Biology*, 159(3):491–502, 2011.

[71] S. May, A. Gamst, R. Haubrich, C. Benson, and D. M. Smith. Pooled Nucleic Acid Testing to Identify Antiretroviral Treatment Failure during HIV Infection. *Journal of Acquired Immune Deficiency Syndromes*, 53(2):194, 2010.

[72] C. S. McMahan, J. M. Tebbs, and C. R. Bilder. Regression Models for Group Testing Data with Pool Dilution Effects. *Biostatistics*, 14(2):284–298, 2012.

[73] Navy Entomology Center for Excellence. *West Nile Virus Surveillance and Control Guide For U.S. Navy and Marine Corps Installation 2014*, accessed January 17, 2019. https://www.med.navy.mil/sites/nmcphc/Documents/nece/WNV-Surveillance-and-Control-Guide-2014.pdf.

[74] N. T. Nguyen, H. Aprahamian, D. R. Bish, and E. K. Bish. A Methodology for Deriving the Sensitivity of Pooled Testing, based on Viral Load Progression and Pooling Dilution. *Journal of Translational Medicine (in press)*, 2019.

[75] N. T. Nguyen, E. K. Bish, and H. Aprahamian. Sequential Prevalence Estimation with Pooling and Continuous Test Outcomes. *Statistics in Medicine*, 37(15):2391–2426, 2018.

[76] N. T. Nguyen, E. K. Bish, and D. R. Bish. Optimal pooled testing design for prevalence estimation. Working paper. *Department of Industrial and Systems Engineering, Virginia Tech*, 2019.

[77] L. Overbergh, A. Giulietti, D. Valckx, B. Decallonne, R. Bouillon, and C. Mathieu. The Use of Real-Time Reverse Transcriptase PCR for the Quantification of Cytokine Gene Expression. *Journal of Biomolecular Techniques: JBT*, 14(1):33, 2003.

[78] Pan American Health Organization. *Regional Zika Epidemiological Update (Americas) October 20, 2016*, accessed November 02, 2016. http://www.paho.org/hq/index.php?option=com_content\&id=11599\&Itemid=41691.

[79] G. Perakis and G. Roels. Regret in the Newsvendor Model with Partial Information. *Operations Research*, 56(1):188–203, 2008.

[80] C. D. Pilcher, G. Joaki, I. F. Hoffman, F. E. A. Martinson, C. Mapanje, P. W. Stewart, K. A. Powers, S. Galvin, D. Chilongozi, S. Gama, M. A. Price, A. F. Fiscus, and Cohen M. S. Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. *AIDS*, 21(13):1723, 2007.

[81] N. A. Pritchard and J. M. Tebbs. Bayesian Inference for Disease Prevalence using Negative Binomial Group Testing. *Biometrical Journal*, 53(1):40–56, 2011.

[82] W.K. Roth, M.P. Busch, A. Schuller, S. Ismay, A. Cheng, C.R. Seed, C. Jungbauer, P.M. Minsk, D. Sondag-Thull, S. Wendel, et al. International survey on NAT testing of blood donations: expanding implementation and yield from 1999 to 2009. *Vox Sanguinis*, 102(1):82–90, 2012.

[83] C. R. Rutledge, J. F. Day, C. C. Lord, L. M. Stark, and W. J. Tabachnick. West Nile Virus Infection Rates in Culex Nigripalpus (Diptera: Culicidae) Do not Reflect Transmission Rates in Florida. *Journal of Medical Entomology*, 40(3):253–258, 2003.

[84] L. J. Savage. The Theory of Statistical Decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951.

[85] P. Sham, J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. DNA Pooling: a Tool for Large-Scale Association Studies. *Nature Reviews Genetics*, 3(11):862–871, 2002.

[86] V. Shyamala. Nucleic Acid Technology (NAT) Testing for Blood Screening: Impact of Individual Donation and Mini Pool–NAT Testing on Analytical Sensitivity, Screening Sensitivity and Clinical Sensitivity. *ISBT Science Series*, 9(2):315–324, 2014.

[87] D. Simkiss. Zika Virus. *Journal of Tropical Pediatrics*, 62(1):1–2, 2016.

[88] Sino Biological Inc. *What is ELISA? Enzyme-linked immunosorbent assay (ELISA)*, accessed August 11, 2017. `http://www.elisa-antibody.com/ELISA-Introduction`.

[89] M. Sobel and R.M. Elashoff. Group Testing with a New Goal, Estimation. *Biometrika*, 62(1):181–193, 1975.

[90] S. L. Stramer, C. T. Fang, G. A. Foster, A. G. Wagner, J. P. Brodsky, and R. Y. Dodd. West Nile Virus among Blood Donors in the United States, 2003 and 2004. *New England Journal of Medicine*, 353(5):451–459, 2005.

[91] S. L. Stramer, D. E. Krysztof, J. P. Brodsky, T. A. Fickett, B. Reynolds, R. Y. Dodd, and S. H. Kleinman. Comparative Analysis of Triplex Nucleic Acid Test Assays in United States Blood Donors. *Transfusion*, 53(10.2):2525–2537, 2013.

[92] W. H. Swallow. Group Testing for Estimating Infection Rates and Probabilities of Disease Transmission. *Phytopathology*, 75(8):882–889, 1985.

[93] J. M. Tebbs and C. R. Bilder. Confidence Interval Procedures for the Probability of Disease Transmission in Multiple-Vector-Transfer Designs. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(1):75, 2004.

[94] J. M. Tebbs, C. R. Bilder, and B. K. Moser. An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics-Theory and Methods*, 32(5):983–995, 2003.

[95] J. M. Tebbs, C. S. McMahan, and C. R. Bilder. Two-Stage Hierarchical Group Testing for Multiple Infections with Application to the Infertility Prevention Project. *Biometrics*, 69(4):1064–1073, 2013.

[96] J. M. Tebbs and W. H. Swallow. Estimating Ordered Binomial Proportions with the Use of Group Testing. *Biometrika*, pages 471–477, 2003.

[97] The Henry J. Kaiser Family Foundation. *U.S. Federal Funding for HIV/AIDS: Trends Over Time*, accessed November 02, 2016. http://kff.org/global-health-policy/fact-sheet/u-s-federal-funding-for-hivaids-trends-over-time/.

[98] K. H. Thompson. Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics*, 18(4):568–578, 1962.

[99] X. M. Tu, E. Litvak, and M. Pagano. Screening Tests: Can We Get More by Doing Less? *Statistics in Medicine*, 13(19-20):1905–1919, 1994.

[100] X. M. Tu, E. Litvak, and M. Pagano. On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika*, 82(2):287–297, 1995.

[101] U.S. Department of Health and Human Services. *Procleix Ultrio Plus Assay*, accessed January 12, 2018. https://www.fda.gov/downloads/BiologicsBloodVaccines/BloodBloodProducts/ApprovedProducts/LicensedProductsBLAs/BloodDonorScreening/InfectiousDisease/UCM092120.pdf.

[102] S. Vansteelandt, E. Goetghebeur, and T. Verstraeten. Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples. *Biometrics*, 56(4):1126–1133, 2000.

[103] M. S. Warasi, J. M. Tebbs, C. S. McMahan, and C. R. Bilder. Estimating the Prevalence of Multiple Diseases from Two-Stage Hierarchical Pooling. *Statistics in Medicine*, 35:3851–3864, 2016.

[104] L. M. Wein and S. A. Zenios. Pooled Testing for HIV Screening: Capturing the Dilution Effect. *Operations Research*, 44(4):543–569, 1996.

[105] J. Weusten, H. Van Drimmelen, and N. Lelie. Mathematic modeling of the risk of HBV, HCV, and HIV transmission by window-phase donations not detected by NAT. *Transfusion*, 42(5):537–548, 2002.

[106] J. Weusten, M. Vermeulen, H. Van Drimmelen, and N. Lelie. Refinement of a Viral Transmission Risk Model for Blood Donations in Seroconversion Window Phase Screened by Nucleic Acid Testing in Different Pool Sizes and Repeat Test Algorithms. *Transfusion*, 51(1):203–215, 2011.

[107] World Bank Group. *The Short-term Economic Costs of Zika in Latin America and the Caribbean*, accessed April 03, 2019. http://pubdocs.worldbank.org/en/410321455758564708/The-short-term-economic-costs-of-Zika-in-LCR-final-doc-autores-feb-18.pdf.

[108] World Health Organization. *HIV/AIDS Prevalence in Sub-Saharan Africa Data by Sex and Residence* , accessed December 14, 2016. http://apps.who.int/gho/data/node.main.247?lang=en.

[109] World Health Organization. *West Nile virus*, accessed September 05, 2018. http://www.who.int/news-room/fact-sheets/detail/west-nile-virus.

[110] M. Xie. Regression Analysis of Group Testing Samples. *Statistics in Medicine*, 20(13):1957–1969, 2001.

[111] S. A. Zenios and L. M. Wein. Pooled Testing for HIV Prevalence Estimation: Exploiting the Dilution Effect. *Statistics in Medicine*, 17(13):1447–1467, 1998.

[112] Z. Zhang, C. Liu, S. Kim, and A. Liu. Prevalence estimation subject to misclassification: the mis-substitution bias and some remedies. *Statistics in Medicine*, 33(25):4482–4500, 2014.

Appendix A

Appendix to Chapter 2

A.1 Summary of Notation

See Table A.1.

A.2 The Numerical Procedure for Computing the MLE of the Prevalence Rate under Continuous Test Outcomes

We use the following algorithm, which expands upon the iterative algorithm proposed by Zenios and Wein (1998) [111], to solve for $\hat{p}_{MLE}$ in each estimation stage, when a dual-configuration pooling design is utilized. Let p_0 denote the initial estimate of p at the beginning of stage s, and let $\hat{p}_{MLE}$ denote the MLE obtained at the end of stage s, $s = 1, 2$. The algorithm that is used to solve for $\hat{p}_{MLE}$, given p_0, follows:

1. Initialize: $t = 0$; $\hat{p}_{MLE,(t)} = p_0$, i.e., the initial estimate of p.

2. Solve for $\hat{p}_{MLE,(t+1)}$ using the following equation:

$$\hat{p}_{MLE,(t+1)} = \frac{1}{(\sum_{i=1}^{2} n_i^* m_i^*)} \sum_{i=1}^{2} \sum_{j=1}^{n_i^*} \sum_{k=0}^{m_i^*} k\tau^{(m_i^*)}(k; y_j^{(m_i^*)}, \hat{p}_{MLE,(t)}).$$

3. If $\left|\hat{p}_{MLE,(t+1)} - \hat{p}_{MLE,(t)}\right| \leq \epsilon$, terminate; $\hat{p}_{MLE} = \hat{p}_{MLE,(t+1)}$. Otherwise, increment t ($t \leftarrow t + 1$) and return to Step 2.

In the estimation procedures implemented in Section 2.3, we use a tolerance level of $\epsilon = 10^{-7}$.

Table A.1: Summary of Notation – Chapter 2

Random Variables	
Y^+	Bio-marker concentration of a random infected subject, with pdf $f_{Y^+}(\cdot)$
Y^-	Noise level coming from a random uninfected subject, with pdf $f_{Y^-}(\cdot)$
Y_i	Bio-marker concentration of subject $i, i = 1, \cdots, m$, whose specimen is included in a pool of size m
$Y^{(m)}$	Average bio-marker concentration plus noise of a pool of size m, with pdf $f_{Y^{(m)}}$
$W(m;p)$	Number of infected specimens in a pool of size m ($\sim$ Binomial (m,p))
$S^{(m;k)}$	Conditional sum of bio-marker concentration plus noise of a pool of size m, given k infected specimens in the pool, with pdf $f_{S^{(m;k)}}$, $\forall k \in \mathbb{N}, k \leq m$
$Y^{(m;k)}$	Conditional average bio-marker concentration plus noise of a pool of size m, given k infected specimens in the pool, with pdf $f_{Y^{(m;k)}}$, $\forall k \in \mathbb{N}, k \leq m$
$\hat{p}^{(1)}_{MLE}$	Random outcome (MLE of p) from stage 1 ($p_0^{(2)} = \hat{p}^{(1)}_{MLE}$)
N^+	Number of pools with a positive test outcome, used in the single-stage estimation procedure with binary test outcomes
Random Vectors	
$\mathbf{Y(m,n)}$	$= \left\{\mathbf{Y}^{(m_1,n_1)}, \cdots, \mathbf{Y}^{(m_C,n_C)}\right\}$
$\mathbf{Y}^{(m_i,n_i)}$	$= \left(Y_j^{(m_i)}\right)_{j=1,\cdots,n_i}, i = 1, \cdots, C$
Decision Variables	
$\mathbf{D}_s = (m_{si}, n_{si})$, $s = 1, 2$, $i = 1, \cdots, C$	Pooling design (pool size: m_{si}, number of pools of size m_{si}: n_{si}) for stage s, with C pooling configurations
Parameters	
p	The true prevalence rate, unknown to the decision-maker
$p_0^{(1)}$	The initial estimate of p at the beginning of stage 1. For single-stage estimation procedures, $p_0^{(1)} = p_0$
C	Number of pooling configurations
B	Testing budget
$B^{(s)}$	Testing budget available for stage s, $s = 1, 2$
λ	Budget allocation factor, such that $B^{(1)} = \lambda B$ and $B^{(2)} = (1-\lambda)B$
c_f	Per pool testing cost
c_v	Per specimen collection cost
Th	Pre-set threshold used in estimation procedures that utilize binary test outcomes
$Se(m,Th)$	Test sensitivity under binary test outcomes, given a pool of size m and a threshold of Th
$Sp(m,Th)$	Test specificity under binary test outcomes, given a pool of size m and a threshold of Th
Performance Metrics	
MLE	The average of $\hat{p}_{MLE}$ across all simulation replications, i.e., an estimate of $E[\hat{p}_{MLE}; p_0]$
MSE	The average of $(\hat{p}_{MLE} - p)^2$ across all simulation replications, i.e., an estimate of $MSE(\hat{p}_{MLE}; p)$
rBias (%)	The relative bias (in percentage) of MLE; i.e., $rBias(\%) = 100\left\lvert \frac{MLE-p}{p} \right\rvert$

A.3 Numerical Corrections to the MLE for Estimation Procedures with Binary Test Outcomes

The vertically corrected estimate of p, denoted by $\hat{p}_V$, is calculated as, $\hat{p}_V = E[\hat{p}_{MLE} \mid p = \hat{p}_{MLE}]$, while the horizontally corrected estimate, $\hat{p}_H$, is calculated as, $\hat{p}_{MLE} = E[\hat{p}_{MLE} \mid p = \hat{p}_H]$; see [54] for details. When testing errors are present, given $\hat{p}_{MLE}$, the test sensitivity of $Se(m,Th)$ and specificity of $Sp(m,Th)$, the pooling design of (m,n), and a preset threshold of Th, we have the following expression for $\hat{p}_V$:

$$\hat{p}_V = E[\hat{p}_{MLE} \mid p = \hat{p}_{MLE}] = \sum_{n^+=0}^{n} \left\{ 1 - \left(\frac{Se(m,Th) - \frac{n^+}{n}}{Se(m,Th) + Sp(m,Th) - 1} \right)^{\frac{1}{m}} \right\} \left\{ \binom{n}{n^+} (f^+)^{n^+} (1-f^+)^{n-n^+} \right\},$$

where f^+ is the probability of a pool being positive, i.e., $f^+ = (1-(1-\hat{p}_{MLE})^m)Se(m,Th) + (1-\hat{p}_{MLE})^m Sp(m,Th)$. On the other hand, $\hat{p}_H$ is computed by solving the following equation:

$$\hat{p}_{MLE} = E[\hat{p}_{MLE} \mid p = \hat{p}_H] = \sum_{n^+=0}^{n} \left\{ 1 - \left(\frac{Se(m,Th) - \frac{n^+}{n}}{Se(m,Th) + Sp(m,Th) - 1} \right)^{\frac{1}{m}} \right\} \left\{ \binom{n}{n^+} (f^+)^{n^+} (1-f^+)^{n-n^+} \right\},$$

where $f^+ = (1-(1-\hat{p}_H)^m)Se(m,Th) + (1-\hat{p}_H)^m Sp(m,Th)$. In the case study of Section 2.3.2, the estimation procedure with binary test outcomes utilizes the following data for the HIV Ultrio Plus Assay: $Th = 1,700$ copies/ml; $Sp(m,Th) = 99.5\%$ [101] and $Se(m,Th)$ values are as reported in Table A.2. The threshold, Th, is calibrated via simulation such that $Se(16,Th) = 88\%$ [91] for 100,000 randomly generated pools, each of size 16. Once Th is determined, it is utilized to compute $Se(m,Th)$ via simulation for various m values based on 100,000 randomly generated pools.

A.4 HIV Viral Load Model

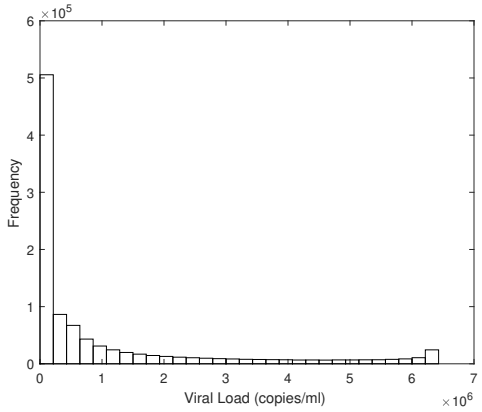
The HIV viral load in an infected individual goes through various phases of growth rate post-infection: pre ramp-up phase, ramp-up phase, where the growth rates accelerate, and post ramp-up phase, where the growth rates decrease to reach a plateau [42, 80, 105, 106]. We extend upon the widely used doubling time viral load model, proposed by Busch [17], and the probit model, studied by Weusten *et al.* [105, 106], to develop a mathematical model that reflects the viral load throughout the lifetime of an HIV-infected individual post-infection [74]. The model is calibrated such that the test sensitivity for a pool of size 16, during the window period of an HIV-infected individual (i.e., the period during which the viral load remains below the pre-set threshold for the HIV Ultrio Plus Assay), is 88% [91]. Further calibration is conducted such that the HIV viral load level peaks at day 17, with an average viral load level of 6.8 $\log_{10}$ copies/ml, and reaches steady state at day 61, with an average viral load level of 5.1 $\log_{10}$ copies/ml [80]. Let t_w, t_p, and t_s respectively denote the time at which the window period ends, the viral load peaks, and the viral load reaches steady state; similarly, let VL_w, VL_p, and VL_s respectively denote the viral load at the end of the window period, at peak viremia, and at steady state. Let $(VL|T=t)$ denote the viral load of an HIV-infected individual, given a random testing time, post-infection, of T, with realization t, i.e., $T = t$.

Then, the viral load is modeled by the following equation [74]:

$$(VL|T=t) = \begin{cases} C_0 2^{t/\gamma}, & \text{if } t \le t_w \\ VL_w + \frac{C_w}{t} \exp\left(-\frac{(ln(t-t_w)-a)^2}{b}\right), & \text{if } t_w < t \le t_s \\ VL_s, & \text{if } t > t_s \end{cases},$$

where the parameters are calibrated as, $a = 1.98$, $b = 1.73$, $C_0 = 9$, $\gamma = 0.85$, $C_w = 1.096 \times 10^8$ [74, 106]. Using this viral load model, we generated 1,000,000 realizations of the viral load. The resulting distribution of Y^+ is depicted in Figure A.1.

Figure A.1: HIV viral load (Y^+) distribution for a random infected individual, when the testing time is uniformly distributed between 0 and 100 days



A.5 Functional Form and Regression Coefficients of $\mu(m, p_0^{(s)})$ and $\sigma^2(m, p_0^{(s)})$ in Case Study 1

For case study 1, we have the following functional forms:

$$\mu(m, p_0^{(s)}) = a_2 {p_0^{(s)}}^2 + a_1 p_0^{(s)} + a_0, \text{ and}$$
$$\sigma^2(m, p_0^{(s)}) = a_0 + a_1 p_0^{(s)} + a_2 {p_0^{(s)}}^2 + a_3 {p_0^{(s)}}^3 + a_4 {p_0^{(s)}}^4 + a_5 {p_0^{(s)}}^5 + a_6 {p_0^{(s)}}^6,$$

where the values of a_0, $a_1 \cdots$, a_6 are reported in Tables A.3 and A.4.

A.6 Functional Form and Regression Coefficients of $\mu(m, p_0^{(s)})$ and $\sigma^2(m, p_0^{(s)})$ in Case Study 2

For case study 2, we have the following functional forms:

$$\mu(m, p_0^{(s)}) = a_0 + a_1 x + a_2 x^2 + a_3 x^3, \text{ and}$$
$$\sigma^2(m, p_0^{(s)}) = a_0 + a_1 p_0^{(s)} + a_2 {p_0^{(s)}}^2 + a_3 {p_0^{(s)}}^3 + a_4 {p_0^{(s)}}^4 + a_5 {p_0^{(s)}}^5,$$

where x denotes $\log\left(p_0^{(s)}\right)$, and the values of a_0, $a_1 \cdots$, a_5 are reported in Tables A.5 and A.6.

A.7 Additional Numerical Results for the Case Studies

The optimal pooling designs in Case Studies 1 and 2 are respectively given in Tables A.7 and A.12. Note that, for **SE** (i.e., $\lambda = 0.25$ and $\lambda = 0.5$), only stage 1 optimal pooling designs are reported, as stage 2 optimal pooling designs are random variables, and rely on the outcome of stage 1, i.e., $\hat{p}_{MLE}^{(1)}$. Tables A.7 and A.12 also illustrate how B and $p_0^{(1)}$ impact the optimal pooling design.

For Case Study 1, Tables A.8 and A.9 provide the numerical results for scenarios with $p = 0.022$, and Tables A.10 and A.11 provide the results for scenarios with $p = 0.071$ and $p = 0.044$.

For Case Study 2, Table A.13 provides the numerical results of **SE** and the single-stage estimation procedure under incorrectly specified parameters and distributions of Y^+.

A.8 The Impact of the Test's Measurement Error on Estimation Efficiency

In this appendix we model the measurement error in the pooled test's reading as a function that is independent of the number of uninfected specimens in the pool (similar to [111]), and study its impact on the proposed estimation procedure considering the first case study of HIV prevalence estimation. Let $X^{(m)}$ denote the *measured* bio-marker concentration of a pool of size m, with an *actual* (and *unobservable*) bio-marker concentration of $Y^{(m)}$. Then, Eqs. (2.4) and (2.5) continue to hold, with $f_{Y^{(m;k)}}(\cdot)$ replaced by $f_{X^{(m;k)}}(\cdot)$, where $f_{X^{(m;k)}}(\cdot)$ is given by [111], that is:

$$f_{X^{(m;k)}}(x) = \int_0^\infty f_Y^{(m;k)}(y) f(x \mid y) dy.$$

Next, we consider the first case study under measurement error as modeled here. We keep the optimal pooling design as obtained in Section 2.3.1, but modify the computation of the MLE as explained above. We note here that we adapted the measurement error model from [111] and [104] to our HIV Ultrio Plus data. Following [111] and [104], for the HIV viral load model in the first case study, $[X^{(m)} \mid y^{(m)}] \sim$ Normal $(y^{(m)}, \alpha(y^{(m)})^2)$, where α is a testing kit specific parameter. Let x_q represent the actual viral load for which the sensitivity of individual testing equals $q\%$, for $q \in [0, 100]$, that is, $Pr[X^{(1)} > Th \mid y^{(1)} = x_q] = q\%$. Then, α is the solution to: $\alpha = \frac{x_{50} - x_{95}}{x_{95}} z_{0.05}$, where $z_{0.05} = -1.645$, and from [105, 106], $x_{95} = 18.4$ $\log_{10}$ copies/ml, and $x_{50} = 2.7$ $\log_{10}$ copies/ml, leading to $\alpha = 0.6079$.

The simulation for this case study is implemented similarly to that of the original case, outlined in Section 2.3.1, with an additional step to generate the *measured* viral load: once the pool's *actual* viral load, i.e., $y^{(m)}$, is computed, the pool's measured viral load, i.e., $X^{(m)}$, is generated based on $[X^{(m)} \mid y^{(m)}] \sim$ Normal $(y^{(m)}, \alpha(y^{(m)})^2)$.

Table A.14 reports the performance measures for the single-stage estimation procedure and **SE** both for the original model studied in the paper (see Section 2.3.2 and Table 2.3) and in the presence of measurement error as modeled in this section, for a testing budget of $B = \$4,460$. Table A.14 indicates that the inclusion of measurement error changes the MLE only slightly, with minor deviations in the relative bias and MSE for both the **SE** and the single-stage estimation procedures. Table A.14 also indicates that the pooling designs obtained via our original model perform well under measurement error, as long as the expression for the MLE is adjusted to capture the measurement error. However, the HIV Ultrio Plus Assay considered in case study 1 is known to be very accurate [91, 105, 106], as it measures the viral load directly rather than the host's reaction to the infection (e.g., antibodies), thus, the magnitude of measurement error is expected to be very small in our setting.

A.9 Accuracy of the MSE Approximation

In this appendix, we study the accuracy of the MSE approximation provided in Eqn. (2.15). In particular, we compare the MSE estimate, given by the average value of $(\hat{p}_{MLE} - p)^2$ across 20,000 simulation replications, with the MSE approximation computed via Eqn. (2.15), with $p_0^{(s)}$ replaced by the average value of $\hat{p}_{MLE}$ obtained in the simulation, as this average now serves as the estimate of the unknown p; see Table A.15. Table A.15 shows that the approximated MSE and the average value of $(\hat{p}_{MLE} - p)^2$ are relatively close, with the approximated MSE being lower than the average value of $(\hat{p}_{MLE} - p)^2$ in all cases. Table A.15 also shows that the closer the average value of $\hat{p}_{MLE}$ to p, the more accurate the MSE approximation is. Thus, the use of **SE** is especially important in cases where $p_0^{(1)}$ significantly deviates from p, e.g., in prevalence

estimation for emerging infections.

A.10 Confidence Interval Estimation of p

In this appendix, we propose an approach to estimating the confidence interval of p, given the computed final MLE, $\hat{p}^{(2)}_{MLE}$, and the test continuous outcomes. From Section 2.2.3, we found that:

$$Var(\hat{p}^{(2)}_{MLE};p)=Var\left(\frac{1}{nm}\sum_{j=1}^{n}\sum_{k=0}^{m}k\tau^{(m)}(k;y_j^{(m)},\hat{p}^{(2)}_{MLE})\right)=\frac{1}{nm^2}Var\left[E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right],\ \forall j=1,\cdots,n, \text{ and}$$

$$Bias(\hat{p}^{(2)}_{MLE};p)=E\left(\frac{1}{nm}\sum_{j=1}^{n}\sum_{k=0}^{m}k\tau^{(m)}(k;y_j^{(m)},\hat{p}^{(2)}_{MLE})\right)-p=\frac{1}{m}\ E\left[E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right]-\hat{p}^{(2)}_{MLE},\ \forall j=1,\cdots,n.$$

Note that in the two equations above, $\hat{p}^{(2)}_{MLE}$ is replaced by $\hat{p}^{(1)}_{MLE}$ for single-stage estimation procedures. From [58,98], $\left(\hat{p}^{(2)}_{MLE}-p\right)\sim$ Normal $\Big(Bias(\hat{p}^{(2)}_{MLE};p),Var(\hat{p}^{(2)}_{MLE};p)\Big)$, and the confidence interval of p, with a confidence level of $100(1-\alpha)\%$, can be approximated by:

$$\left[\hat{p}^{(2)}_{MLE}-Bias(\hat{p}^{(2)}_{MLE};p)\right]\pm z_{1-\alpha/2}\sqrt{Var(\hat{p}^{(2)}_{MLE};p)},$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. We propose the following approximations for $Bias(\hat{p}^{(2)}_{MLE};p)$ and $Var(\hat{p}^{(2)}_{MLE};p)$ based on the continuous test outcomes and $\hat{p}^{(2)}_{MLE}$: First, we note that, for any given $y_j^{(m)}$, $\forall j=1,\cdots,n$, $E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]=\sum_{k=0}^{m}k\tau^{(m)}(k;y_j^{(m)},\hat{p}^{(2)}_{MLE})$, where $\tau^{(m)}(k;y_j^{(m)},\hat{p}^{(2)}_{MLE})$ is given by Eqn. (2.4). Thus, given the test continuous outcomes, $y_j^{(m)}$, $\forall j=1,\cdots,n$, and $\hat{p}^{(2)}_{MLE}$:

$$\begin{aligned}E\left[E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right]&\approx\frac{1}{n}\sum_{j=1}^{n}E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\\&=\bar{E}\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big];\end{aligned}$$

$$\begin{aligned}Var\left[E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right]&\approx\frac{1}{n-1}\sum_{j=1}^{n}\left(E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]-\bar{E}\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right)^2\\&=s^2\left(E\Big[W(m;\hat{p}^{(2)}_{MLE})\mid y_j^{(m)}\Big]\right).\end{aligned}$$

Therefore, we have the following approximations:

$$Var(\hat{p}^{(2)}_{MLE};p) \approx \frac{1}{nm^2} s^2 \left(E\Big[W(m;\hat{p}^{(2)}_{MLE}) \mid y_j^{(m)} \Big] \right), \text{ and}$$

$$Bias(\hat{p}^{(2)}_{MLE};p) \approx \frac{1}{m} \bar{E}\Big[W(m;\hat{p}^{(2)}_{MLE}) \mid y_j^{(m)} \Big] - \hat{p}^{(2)}_{MLE}, \ \forall j = 1, \cdots, n.$$

Table A.2: Case Study 1 - Estimation Procedure with Binary Test Outcomes: Estimated Sensitivity of the HIV Ultrio Plus Assay for various pool sizes (m), based on a threshold of $Th = 1,700$ copies/ml.

m	$Se(m, Th)$ (%)
9	89. 3
10	88.9
11	88.6
12	88.5
13	88.3
18	87.8
20	87.8
26	87.3
28	87.3
45	83.3
46	83.2
48	83.1

Table A.3: Regression Coefficients for $\mu(m, p_0)$ (Case Study 1)

m \ Coefficients	a_2	a_1	a_0
2	0	0.508	0
3	0	0.996	0
4	0	1.466	0.000662
5	0	1.926	0.001029
6	0	2.380	0.001504
7	0	2.817	0.002569
8	0	3.243	0.003701
9	-2.565	3.911	0
10	-3.288	4.397	0
11	-3.995	4.873	0
12	-4.759	5.341	0.001282
13	-4.977	5.753	0.002119
14	-5.858	6.212	0.002454
15	-6.666	6.686	0.002183
16	-7.707	7.146	0.002163
17	-9.646	7.700	0.001367
18	-9.198	8.014	0.003523
19	-10.370	8.473	0.003600
20	-12.040	8.972	0.003369
21	-12.540	9.384	0.003754
22	-13.920	9.842	0.004467
23	-15.790	10.350	0.003669
24	-16.550	10.750	0.004632
25	-19.280	11.640	0.004766
26	-17.380	11.150	0.004910
27	-20.980	12.120	0.004565
28	-22.080	12.520	0.005745
29	-23.580	12.960	0.006287
30	-24.600	13.360	0.006710
31	-26.130	13.800	0.006478
32	-27.150	14.190	0.007278
33	-29.830	14.720	0.006857
34	-31.720	15.160	0.007354
35	-32.870	15.540	0.007854
36	-34.770	16.000	0.008277
37	-37.370	16.490	0.007682
38	-38.260	16.840	0.008527
39	-40.220	17.280	0.009125
40	-42.270	17.730	0.009140
41	-44.070	18.120	0.009958
42	-46.010	18.550	0.009897
43	-48.050	18.970	0.010740
44	-49.850	19.370	0.011290
45	-52.380	19.850	0.010410
46	-53.860	20.200	0.011520
47	-56.530	20.670	0.012000
48	-57.800	21.000	0.013190

Table A.4: Regression Coefficients for $\sigma^2(m, p_0)$ (Case Study 1)

m \ Coefficients	$a_6(\times 10^4)$	$a_5(\times 10^4)$	$a_4(\times 10^4)$	a_3	a_2	a_1	a_0
2	0	0	0	-0.291	0.1434	0	0
3	0	0	0	-0.759	0.4590	0	0
4	0	0	0	-2.208	0.9701	0	0
5	0	0	0	-5.921	1.8220	0	0
6	0	0	0	-10.080	2.7320	0	0
7	0	0	0	-14.040	3.5600	0.02056	0
8	0	0	0.01155	-41.870	5.9460	0	0
9	0	0	0	-23.110	5.3440	0.05344	0
10	0	0	0	-32.170	6.7330	0.05942	0
11	0	0	0	-41.360	7.9600	0.07977	-0.0003368
12	0	0	0.0408	-122.800	13.2500	0	0
13	0	0	0.0406	-130.300	14.2500	0	0
14	0	0	0.0439	-142.400	15.2100	0.08657	0
15	0	0	0.0640	-192.200	18.7800	0	0
16	0	0	0.0786	-225.300	21.0700	0.07824	0
17	0	0	0.1092	-292.600	25.3500	0	0
18	0	−1.157	0.3668	-500.900	32.5800	0	0
19	0	0	0.1282	-0.034	28.4200	0.12820	0
20	0	−1.346	0.4429	-604.700	37.4700	0	0
21	0	0	0.1356	-0.038	30.6800	0.20460	0
22	0	0	0.1435	-0.039	31.0200	0.27470	-0.0007841
23	0	−1.700	0.5984	-830.700	48.6000	0	0
24	75.65	−23.800	3.0260	-2078.000	77.9700	0	0
25	0	−1.579	0.6935	-977.600	55.0700	0	0
26	0	−1.884	0.6017	-871.300	50.3100	0	0
27	76.34	−25.010	3.3310	-2386.000	89.6000	0	0
28	0	0	0.2462	-617.100	41.7600	0.50400	-0.0013390
29	0	−2.123	0.7776	-1086.000	58.1800	0.35950	0
30	0	−2.425	0.9096	-1263.000	66.3700	0.29360	0
31	0	−3.543	1.1590	-1448.000	70.5500	0.34200	0
32	74.67	−23.100	3.0510	-2290.000	87.3900	0	0
33	80.72	−28.240	4.0740	-3103.000	114.1000	0	0
34	108.30	−35.790	4.8460	-3469.000	121.3000	0	0
35	0	−36.260	1.2580	-1610.000	75.5100	0.55360	0
36	0	−4.055	1.3970	-1757.000	80.2300	0.57770	0
37	0	−4.646	1.5620	-1906.000	84.6300	0.58520	0
38	0	−5.217	1.6890	-2008.000	86.8100	0.64290	0
39	102.60	−34.670	4.9020	-3670.000	127.0000	0	0
40	92.68	−34.750	5.3070	-4104.000	141.2000	0	0
41	−6.21	1.962	-0.2257	92.250	0.8139	0	0
42	0	−7.018	2.2390	-2557.000	103.6000	0.72820	0
43	0	−8.413	2.5440	-2773.000	108.0000	0.77350	0
44	0	−7.129	2.2350	-2508.000	97.2200	0.99260	-0.0020660
45	145.10	−52.300	7.5810	-5471.000	172.1000	0	0
46	0	−7.668	2.4180	-2686.000	100.9000	1.09700	-0.0022030
47	102.90	−41.340	6.6120	-5135.000	166.0000	0	0
48	0	−8.607	2.6180	-2819.000	101.5000	1.26300	-0.0026390

Table A.5: Regression Coefficients for $\mu(m, p_0)$ (Case Study 2)

m \ Coefficients	a_3	a_2	a_1	a_0
2	0.04426	0.7088	4.818	1.717
3	0.03154	0.5346	4.078	2.893
4	0.02102	0.389	3.428	3.206
5	0.01373	0.2785	2.895	3.217
6	0.009524	0.2114	2.543	3.218
7	0.003076	0.1215	2.135	3.056
8	0	0.07441	1.89	2.993
9	0	0.06379	1.782	3.032
10	0	0.0544	1.687	3.055
11	0	0.04868	1.621	3.092
12	0	0.0408	1.546	3.103
13	0	0.03673	1.5	3.138
14	0	0.03319	1.457	3.17
15	0	0.02918	1.415	3.192
16	0	0.02577	1.38	3.221
17	-0.006679	-0.05073	1.103	2.998
18	0	0.02194	1.331	3.284
19	0	0.01914	1.302	3.303
20	0	0.01875	1.29	3.344
21	-0.006337	-0.05473	1.027	3.12
22	0	0.01567	1.255	3.398
23	0	0.01441	1.24	3.425
24	-0.005589	-0.05027	1.009	3.229
25	-0.004957	-0.04458	1.018	3.275
26	0	0.01194	1.206	3.509
27	0	0.01079	1.193	3.529
28	0	0.01046	1.187	3.559
29	-0.004641	-0.04287	0.9981	3.398
30	0	0.009234	1.171	3.611
31	-0.004931	-0.04687	0.9744	3.44
32	-0.0044	-0.04146	0.9866	3.481
33	-0.004594	-0.04444	0.9724	3.498
34	0	0.0074	1.145	3.704
35	0	0.006853	1.139	3.725
36	-0.003991	-0.03803	0.9842	
37	-0.004168	-0.04068	0.9704	3.606
38	-0.003815	-0.03663	0.9821	3.646
39	-0.003796	-0.03742	0.9743	3.661
40	-0.003696	-0.03602	0.9765	3.688
41	-0.003485	-0.03386	0.982	3.718
42	-0.00318	-0.03066	0.9893	3.749
43	0	0.005238	1.11	3.896
44	0	0.004779	1.105	3.913
45	0	0.005064	1.105	3.937
46	0	0.004636	1.101	3.953
47	0	0.004675	1.1	3.972
48	-0.003608	-0.03658	0.9565	3.846
49	-0.003097	-0.03059	0.9757	3.886
50	-0.003133	-0.03144	0.9703	3.899

Table A.6: Regression Coefficients for $\sigma^2(m, p_0)$ (Case Study 2)

m \ Coefficients	a_5	a_4	a_3	a_2	a_1	a_0
2	0	0	1201	-642.9	112.5	10.16
3	0	-7579	4332	-995.1	93.19	7.61
4	0	-5703	3407	-775.6	60.03	6.203
5	0	-6893	3766	-748.5	45.07	5.158
6	0	-5638	3020	-571.6	27.17	4.48
7	0	-4763	2558	-460.8	16.13	3.951
8	0	-4516	2321	-389.3	9.533	3.483
9	0	-4470	2190	-339.2	4.732	3.131
10	35320	-21290	4877	-486.5	5.697	2.818
11	0	-3016	1412	-190.1	-4.513	2.62
12	0	-2532	1150	-140.6	-7.263	2.419
13	29420	-17010	3623	-308.2	-2.965	2.198
14	28210	-16150	3362	-269.4	-4.761	2.048
15	25760	-14510	2911	-209.2	-7.65	1.944
16	20060	-11610	2398	-173.9	-7.729	1.801
17	0	-1189	429.9	-11.73	-12.62	1.726
18	0	-775	211.9	25.74	-14.49	1.66
19	0	-633.4	169.4	27.22	-13.75	1.553
20	0	0	-96.71	63.15	-15.04	1.486
21	0	0	-120.8	69.74	-15.19	1.422
22	0	0	-111.6	65.77	-14.44	1.345
23	0	0	-131.1	71.61	-14.66	1.298
24	0	0	-123.5	67.62	-13.85	1.225
25	0	0	-132.1	69.68	-13.75	1.18
26	0	334.3	-281.9	91.48	-14.68	1.152
27	0	0	-138.8	70.05	-13.19	1.087
28	0	240	-232.8	81.08	-13.39	1.053
29	0	265.3	-244.1	82.02	-13.18	1.017
30	0	362.5	-281.1	85.78	-13.07	0.9832
31	0	554.6	-365.6	97.35	-13.42	0.9551
32	0	581.8	-379.2	99.03	-13.32	0.9285
33	0	642.4	-401.7	101	-13.18	0.9003
34	0	645.5	-403.5	100.9	-13.01	0.8771
35	0	648.1	-401.9	99.52	-12.71	0.85
36	0	630.4	-390.1	96.31	-12.26	0.8186
37	0	602.4	-378	94.23	-12	0.7956
38	0	659.6	-397.7	95.61	-11.86	0.7753
39	0	695.6	-416.4	98.41	-11.89	0.7584
40	0	608.1	-374.8	91.27	-11.31	0.7325
41	0	691.4	-407.4	94.86	-11.33	0.7184
42	0	682.2	-403.8	94.09	-11.18	0.701
43	0	617.4	-371.9	88.39	-10.7	0.6794
44	0	766.6	-439.4	98.39	-11.17	0.6746
45	0	633.1	-375.9	87.78	-10.41	0.6492
46	0	726.1	-416.6	93.47	-10.64	0.642
47	0	701.4	-402.6	90.51	-10.33	0.6253
48	0	704	-403	90.09	-10.19	0.611
49	0	694.6	-396.3	88.41	-9.984	0.598
50	0	699.9	-395.3	87.34	-9.794	0.5848

Table A.7: Case Study 1: Stage 1 optimal pooling design for different values of $p_0^{(1)}$, λ, and B. All optimal pooling designs are reported in the form, (m^*, n^*) for single configurations, and $\{(m_1^*, n_1^*), (m_2^*, n_2^*)\}$ for dual configurations.

Budget	$p_0^{(1)}$	$\lambda = 1$	$\lambda = 0.25$	$\lambda = 0.5$
$\$5,575$	0.0110	(45,14)	$\{(39,2)(40,2)\}$	$\{(38,1),(47,6)\}$
	0.0220	(26,23)	$\{(20,1),(26,5)\}$	$\{(15,1),(26,11)\}$
	0.0330	(26,23)	$\{(20,1),(26,5)\}$	$\{(15,1),(26,11)\}$
	0.0355	(26,23)	$\{(20,1),(26,5)\}$	$\{(15,1),(26,11)\}$
	0.0660	(13,41)	$\{(11,2),(12,9)\}$	$\{(12,8),(13,13)\}$
	0.1065	(10,50)	$\{(8,7),(9,7)\}$	$\{(8,1),(9,26)\}$
$\$4,460$	0.0110	(46,11)	$\{(47,1),(40,2)\}$	$\{(40,3),(45,3)\}$
	0.0220	(26,18)	$\{(15,1),(26,4)\}$	$\{(35,1),(26,8)\}$
	0.0330	(26,18)	$\{(15,1),(26,4)\}$	$\{(19,3),(26,7)\}$
	0.0355	(20,23)	$\{(15,1),(26,4)\}$	$\{(19,3),(26,7)\}$
	0.0660	(11,37)	$\{(12,4),(11,5)\}$	$\{(13,7),(12,10)\}$
	0.1065	(9,43)	$\{(8,3),(9,8)\}$	$\{(7,3),(9,19)\}$
$\$3,345$	0.0110	(48,8)	(48,2)	(48,4)
	0.0220	(28,13)	$\{(28,1),(32,2)\}$	$\{(25,1),(26,6)\}$
	0.0330	(18,19)	$\{(18,2),(26,2)\}$	$\{(25,1),(26,6)\}$
	0.0355	(18,19)	$\{(18,2),(26,2)\}$	$\{(25,1),(26,6)\}$
	0.0660	(12,26)	$\{(15,1),(13,5)\}$	$\{(13,1),(12,12)\}$
	0.1065	(9,32)	$\{(10,1),(9,7)\}$	$\{(10,2),(9,14)\}$

Table A.8: Case Study 1: Performance measures for the single-stage estimation procedure with binary and continuous test outcomes, $p = 0.0220$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Meas.	Binary (No Correction)	Binary (Vertical Correction)	Binary (Horizontal Correction)	Continuous
$\$5,575$	0.0110	MLE	0.0219 ($\pm$0.0105)	0.0198 ($\pm$0.0108)	0.0166 ($\pm$0.0078)	0.0228 ($\pm$0.0075)
		MSE ($\times 10^4$)	1.10 ($\pm$3.15)	1.21 ($\pm$5.24)	0.89 ($\pm$2.32)	0.57 ($\pm$1.04)
		rBias(%)	2.15	10.19	24.37	3.80
	0.0330	MLE	0.0212 ($\pm$0.0079)	0.0208 ($\pm$0.0095)	0.0173 ($\pm$0.0074)	0.0224 ($\pm$0.0070)
		MSE ($\times 10^4$)	0.63 ($\pm$1.17)	0.91 ($\pm$2.20)	0.76 ($\pm$1.15)	0.49 ($\pm$0.83)
		rBias(%)	3.87	5.33	21.31	1.99
$\$4,460$	0.0110	MLE	0.0226($\pm$0.0121)	0.0205 ($\pm$0.0131)	0.0170 ($\pm$0.0092)	0.0232 ($\pm$0.0086)
		MSE ($\times 10^4$)	1.46 ($\pm$3.39)	1.74 ($\pm$5.45)	1.11 ($\pm$2.79)	0.76 ($\pm$1.43)
		rBias(%)	2.66	6.79	22.73	5.43
	0.0330	MLE	0.0216 ($\pm$0.0092)	0.0203 ($\pm$0.0082)	0.0167 ($\pm$0.0065)	0.0227 ($\pm$0.0080)
		MSE ($\times 10^4$)	0.85 ($\pm$1.56)	0.71 ($\pm$1.06)	0.70 ($\pm$0.75)	0.65 ($\pm$1.07)
		rBias(%)	2.02	7.79	24.08	2.99
$\$3,345$	0.0110	MLE	0.0254 ($\pm$0.0208)	0.0219 ($\pm$0.0288)	0.0192 ($\pm$0.0204)	0.0236 ($\pm$0.0104)
		MSE ($\times 10^4$)	4.43 ($\pm$14.2)	8.28 ($\pm$31.2)	4.26 ($\pm$14.2)	1.11 ($\pm$2.27)
		rBias(%)	15.45	0.07	12.73	7.45
	0.0330	MLE	0.0218 ($\pm$0.0099)	0.0208 ($\pm$0.0093)	0.0171 ($\pm$0.0072)	0.0226 ($\pm$0.0090)
		MSE ($\times 10^4$)	0.99 ($\pm$1.76)	0.89 ($\pm$1.39)	0.76 ($\pm$0.87)	0.81 ($\pm$1.40)
		rBias(%)	1.16	5.49	22.40	2.74

Table A.9: Case Study 1: Performance measures for the single-stage estimation procedure and **SE**, with continuous test outcomes, $p = 0.0220$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
				$\lambda = 0.25$	$\lambda = 0.5$
		MLE	0.0228 ($\pm$0.0075)	0.0220($\pm$0.0079)	0.0225 ($\pm$0.0072)
	0.0110	MSE ($\times 10^4$)	0.57 ($\pm$1.04)	0.62 ($\pm$1.09)	0.52 ($\pm$0.84)
		rBias (%)	3.80	0.16	2.24
\$5,575					
		MLE	0.0224 ($\pm$0.0070)	0.0219($\pm$0.0079)	0.0224 ($\pm$0.0070)
	0.0330	MSE ($\times 10^4$)	0.49 ($\pm$0.83)	0.63 ($\pm$1.11)	0.49 ($\pm$0.78)
		rBias (%)	1.99	0.42	1.79
		MLE	0.0232 ($\pm$0.0086)	0.0218($\pm$0.0093)	0.0228 ($\pm$0.0081)
	0.0110	MSE ($\times 10^4$)	0.76 ($\pm$1.43)	0.86 ($\pm$1.39)	0.66 ($\pm$1.11)
		rBias (%)	5.43	0.95	3.42
\$4,460					
		MLE	0.0227 ($\pm$0.0080)	0.0212($\pm$0.0095)	0.0225 ($\pm$0.0078)
	0.0330	MSE ($\times 10^4$)	0.65 ($\pm$1.07)	0.91 ($\pm$1.47)	0.61 ($\pm$0.99)
		rBias (%)	2.99	3.61	2.30
		MLE	0.0236 ($\pm$0.0104)	0.0211($\pm$0.0117)	0.0229 ($\pm$0.0097)
	0.0110	MSE ($\times 10^4$)	1.11($\pm$2.27)	1.37 ($\pm$2.04)	0.95 ($\pm$1.68)
		rBias (%)	7.45	4.12	4.17
\$3,345					
		MLE	0.0226 ($\pm$0.0090)	0.0206($\pm$0.0118)	0.0226 ($\pm$0.0094)
	0.0330	MSE ($\times 10^4$)	0.81 ($\pm$1.40)	1.40 ($\pm$2.04)	0.89 ($\pm$1.54)
		rBias (%)	2.74	6.39	2.88

Table A.10: Case Study 1: Performance measures for the single-stage estimation procedure and **SE**, with continuous test outcomes, $p = 0.0710$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
				$\lambda = 0.25$	$\lambda = 0.5$
		MLE	0.0725 ($\pm$0.0152)	0.0719($\pm$0.0140)	0.0719 ($\pm$0.0143)
	0.0355	MSE ($\times 10^4$)	2.32 ($\pm$3.83)	1.97 ($\pm$2.87)	2.07 ($\pm$3.20)
		rBias (%)	2.15	1.26	1.32
\$5,575					
		MLE	0.0716 ($\pm$0.0136)	0.0716($\pm$0.0137)	0.0715 ($\pm$0.0136)
	0.1065	MSE ($\times 10^4$)	1.86 ($\pm$2.78)	1.88 ($\pm$2.81)	1.85 ($\pm$2.78)
		rBias (%)	0.87	0.80	0.75
		MLE	0.0726 ($\pm$0.0728)	0.0719($\pm$0.0156)	0.0722 ($\pm$0.0160)
	0.0355	MSE ($\times 10^4$)	2.70 ($\pm$2.76)	2.45 ($\pm$3.78)	2.59 ($\pm$4.17)
		rBias (%)	2.18	1.31	1.71
\$4,460					
		MLE	0.0718 ($\pm$0.0152)	0.0718($\pm$0.0720)	0.0716 ($\pm$0.0719)
	0.1065	MSE ($\times 10^4$)	2.31 ($\pm$3.46)	2.41 ($\pm$2.46)	2.34 ($\pm$2.39)
		rBias (%)	1.11	1.13	0.94
		MLE	0.0730 ($\pm$0.0188)	0.0724($\pm$0.0186)	0.0728 ($\pm$0.0189)
	0.0355	MSE ($\times 10^4$)	3.56 ($\pm$6.00)	3.48 ($\pm$5.67)	3.62 ($\pm$6.00)
		rBias (%)	2.80	1.91	2.52
\$3,345					
		MLE	0.0719 ($\pm$0.0177)	0.0720($\pm$0.0186)	0.0719 ($\pm$0.0176)
	0.1065	MSE ($\times 10^4$)	3.13 ($\pm$4.67)	3.48 ($\pm$6.06)	3.12 ($\pm$4.71)
		rBias (%)	1.32	1.40	1.28

Table A.11: Case Study 1: Performance measures for the single-stage estimation procedure and **SE**, with continuous test outcomes, $p = 0.0440$. MLE and MSE are reported in the form: sample average ($\pm$ sample standard deviation).

Budget	$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
				$\lambda = 0.25$	$\lambda = 0.5$
		MLE	0.0450 ($\pm$0.0109)	0.0444($\pm$0.0105)	0.0448 ($\pm$0.0106)
	0.0220	MSE ($\times 10^4$)	1.20 ($\pm$1.98)	1.10 ($\pm$1.74)	1.14 ($\pm$1.75)
		rBias (%)	2.20	1.00	1.83
$5,575					
		MLE	0.0446 ($\pm$0.0103)	0.0445($\pm$0.0107)	0.0445 ($\pm$0.0104)
	0.0660	MSE ($\times 10^4$)	1.06 ($\pm$1.59)	1.16 ($\pm$1.93)	1.09 ($\pm$1.65)
		rBias (%)	1.29	1.13	1.18
		MLE	0.0453 ($\pm$0.0124)	0.0447($\pm$0.0121)	0.0447 ($\pm$0.0118)
	0.0220	MSE ($\times 10^4$)	1.54 ($\pm$2.53)	1.47 ($\pm$2.48)	1.40 ($\pm$2.26)
		rBias (%)	2.94	1.58	1.70
$4,460					
		MLE	0.0445 ($\pm$0.0114)	0.0444($\pm$0.0123)	0.0447 ($\pm$0.0115)
	0.0660	MSE ($\times 10^4$)	1.31 ($\pm$1.99)	1.53 ($\pm$2.78)	1.33 ($\pm$2.07)
		rBias (%)	1.19	0.83	1.52
		MLE	0.0459 ($\pm$0.0145)	0.0445($\pm$0.0146)	0.0451 ($\pm$0.0138)
	0.0220	MSE ($\times 10^4$)	2.14 ($\pm$3.72)	2.14 ($\pm$3.70)	1.92 ($\pm$3.17)
		rBias (%)	4.38	1.22	2.62
$3,345					
		MLE	0.0447 ($\pm$0.0135)	0.0442($\pm$0.0150)	0.0446 ($\pm$0.0135)
	0.0660	MSE ($\times 10^4$)	1.81 ($\pm$2.87)	2.25 ($\pm$3.92)	1.82 ($\pm$2.96)
		rBias (%)	1.68	0.47	1.39

Table A.12: Case Study 2: Stage 1 optimal pooling design for different values of $p_0^{(1)}$ and λ. All optimal pooling designs are reported in the form, (m^*, n^*) for single configurations, and $\{(m_1^*, n_1^*), (m_2^*, n_2^*)\}$ for dual configurations.

$p_0^{(1)}$	$\lambda = 1$	$\lambda = 0.25$	$\lambda = 0.5$
0.04	(42,17)	(42, 4)	$\{(40,1),(39,8)\}$
0.06	(38,19)	$\{(31,1),(32,4)\}$	$\{(40,1),(39,8)\}$
0.18	(21, 23)	$\{(20,1),(21,5)\}$	$\{(20,1),(21,11)\}$
0.20	(17,25)	$\{(19,1),(21,5)\}$	$\{(13,1),(17,12)\}$

Table A.13: Case Study 2: Performance measures for the single-stage estimation procedure and **SE** with continuous test outcomes, with incorrect parameters for $Y+$, and incorrect distributions for Y^+, $p = 0.12$, $B = \$52.5$. MLE and MSE are reported in the form: sample average (± sample standard deviation).

	$p_0^{(1)}$	Perf. Measures	Single Stage	**SE**	
				$\lambda = 0.25$	$\lambda = 0.5$
Incorrect Parameters - Y^+	0.18	MLE	0.1239 (±0.0200)	0.119(±0.0365)	0.1263 (±0.0303)
		MSE ($\times 10^3$)	0.42 (±0.59)	1.34 (±3.19)	1.03 (±1.64)
		rBias (%)	3.22	1.09	5.26
	0.20	MLE	0.1241 (±0.0197)	0.119(±0.0373)	0.1220 (±0.0297)
		MSE ($\times 10^3$)	0.42 (±0.56)	1.55 (±3.31)	0.97 (±1.77)
		rBias (%)	3.44	0.98	1.68
Incorrect Distribution - Y^+	0.18	MLE	0.1236 (±0.0195)	0.1193(±0.0374)	0.1242 (±0.0296)
		MSE ($\times 10^3$)	0.39 (±0.52)	1.40 (±3.22)	0.89 (±1.60)
		rBias (%)	2.99	0.59	3.51
	0.20	MLE	0.1234 (±0.0200)	0.1196(±0.0364)	0.1254 (±0.0287)
		MSE ($\times 10^3$)	0.41 (±0.57)	1.32 (±3.14)	0.85 (±1.39)
		rBias (%)	2.89	0.30	4.49

Table A.14: Case Study 1: Performance measures for the single-stage estimation procedure and **SE** with continuous outcomes, with original model, and the model with measurement error (in Appendix A.8), $B = \$4,460$. MLE and MSE are reported in the form: sample average (± sample standard deviation).

	p	$p_0^{(1)}$	Perf. Measures	Single-stage	**SE**	
					$\lambda = 0.25$	$\lambda = 0.5$
Original Model	0.0220	0.0110	MLE	0.0232 (±0.0086)	0.0218(±0.0093)	0.0228 (±0.0081)
			MSE ($\times 10^4$)	0.76 (±1.43)	0.86 (±1.39)	0.66 (±1.11)
			rBias (%)	5.43	0.95	3.42
	0.0440	0.0220	MLE	0.0453 (±0.0124)	0.0447(±0.0121)	0.0447 (±0.0118)
			MSE ($\times 10^4$)	1.54 (±2.53)	1.47 (±2.48)	1.40 (±2.26)
			rBias (%)	2.94	1.58	1.70
	0.0710	0.0355	MLE	0.0726 (±0.0728)	0.0719(±0.0156)	0.0722 (±0.0160)
			MSE ($\times 10^4$)	2.70 (±2.76)	2.45 (±3.78)	2.59 (±4.17)
			rBias (%)	2.18	1.31	1.71
Model with Measurement Error	0.0220	0.0110	MLE	0.0229 (±0.0085)	0.0216(±0.0093)	0.0226 (±0.0083)
			MSE ($\times 10^4$)	0.73 (±1.35)	0.87 (±1.44)	0.68 (±1.23)
			rBias (%)	4.27	1.85	2.53
	0.0440	0.0220	MLE	0.0448 (±0.0123)	0.0440(±0.0121)	0.0444 (±0.0119)
			MSE ($\times 10^4$)	1.51 (±2.55)	1.48 (±2.51)	1.42 (±2.34)
			rBias (%)	1.92	0.0028	0.91
	0.0710	0.0335	MLE	0.0715 (±0.0159)	0.0709 (±0.0156)	0.0707 (±0.0156)
			MSE ($\times 10^4$)	2.53 (±4.14)	2.42 (±3.72)	2.43 (±3.52)
			rBias (%)	0.72	0.074	0.44

Table A.15: Case Study 1: Comparison of the approximated MSE (in Eqn. (2.15)) and the mean value of $(\hat{p}_{MLE} - p)^2$ (obtained from Monte – Carlo simulation), for single-stage estimation procedures, where the mean value of $(\hat{p}_{MLE} - p)^2$ is denoted by $MSE^{(s)}$, and the approximated MSE is denoted by $MSE^{(a)}$.

Budget	$p_0^{(1)}$	Perf. Measures	$p = 0.071$	$p = 0.044$	$p = 0.022$
		MLE	0.0725	0.0450	0.0228
	$\frac{p}{2}$	$MSE^{(s)}$ ($\times 10^4$)	2.32	1.20	0.57
		$MSE^{(a)}$ ($\times 10^4$)	1.70	0.76	0.36
$\$5,575$					
		MLE	0.0716	0.0446	0.0224
	$\frac{3p}{2}$	$MSE^{(s)}$ ($\times 10^4$)	1.86	1.06	0.49
		$MSE^{(a)}$ ($\times 10^4$)	1.26	0.73	0.32
		MLE	0.0726	0.0453	0.0232
	$\frac{p}{2}$	$MSE^{(s)}$ ($\times 10^4$)	2.70	1.54	0.76
		$MSE^{(a)}$ ($\times 10^4$)	1.90	0.97	0.42
$\$4,460$					
		MLE	0.0718	0.0445	0.0227
	$\frac{3p}{2}$	$MSE^{(s)}$ ($\times 10^4$)	2.31	1.31	0.65
		$MSE^{(a)}$ ($\times 10^4$)	1.60	0.95	0.41
		MLE	0.0730	0.0459	0.0236
	$\frac{p}{2}$	$MSE^{(s)}$ ($\times 10^4$)	3.56	2.14	1.11
		$MSE^{(a)}$ ($\times 10^4$)	2.41	1.44	0.65
$\$3,345$					
		MLE	0.0719	0.0447	0.0226
	$\frac{3p}{2}$	MSE ($\times 10^4$)	3.13	1.81	0.81
		$MSE^{(a)}$ ($\times 10^4$)	2.16	1.25	0.58

Appendix B

Appendix to Chapter 3

Table B.1: Summary of Notation – Chapter 3

Notation used in viral load and sensitivity models	
$VL(t)$	Viral load of an infected subject at time t post-exposure
t_w	The time at which the window period ends
t_p	The time at which the viral load peaks
t_s	The time at which the viral load reaches steady state
λ	Doubling time of the viral load during the window period
τ	The life-time of the infection
C_0, C_w, a, b	Infection-specific calibration parameters
$T^+(n)$	The event that the test outcome is positive for pool size n, $n \in \mathbb{Z}^+$
$N_I(n)$	Number of infected specimens in a pool of size n, $n \in \mathbb{Z}^+$
$Spec$	Specificity of a test (constant for any pool size)
$Sens(n)$	Sensitivity of a pooled test, with pool size n, $n \in \mathbb{Z}^+$
$Sens(n;i)$	Conditional sensitivity of a pooled test, with pool size n, given that the pool contains i infected specimens, $i \in \{0, 1, \cdots, n\}$, $n \in \mathbb{Z}^+$
$\Phi(.)$	The cumulative distribution function (CDF) of the standard normal distribution
z	A constant such that $\Phi(z) = 0.95$, i.e., $z = 1.6449$
χ	The number of nucleic acid copies per viral particle
x_{50}, x_{95}	Viral load measurement at which the probability of testing positive is 50% and 95%, respectively
$\widetilde{Sens}(n;i)$	Approximate conditional sensitivity of a pooled test, with pool size n, given that the pool contains i infected specimens, $i \in \{0, 1, \cdots, n\}$, $n \in \mathbb{Z}^+$
β, α, γ	Calibration parameters for the approximation model
MSE	Mean squared error
Notation used in the case study (prevalence estimation)	
s	Number of testing pools
n	Pool size
p_0	An initial estimate of p
c_f	Fixed testing cost per pool
c_v	Collection cost per specimen
B	Total testing budget
$\overline{N}$	The maximum pool size that can be used
$\hat{p}$	The maximum likelihood estimator (MLE) of p
$\sigma^2(n, s; p)$	The asymptotic variance of the MLE for a pool design (n, s), given a prevalence rate of p
$S_I(s)$	Number of positive-testing pools among s pools
$rBias$	Relative bias of the MLE with respect to p

Table B.2: Summary of Abbreviations – Chapter 3

HIV	Human immunodeficiency virus
NAT	Nucleic acid amplification testing
CDF	Cumulative distribution function
FDA	Food and Drug Administration
MSE	Mean squared error
MLE	Maximum likelihood estimator
CI	Confidence interval

Appendix C

Appendix to Chapter 4

C.1 Summary of Notation

Table C.1: Summary of Notation – Chapter 4

Decision Variables	
m	Pool size, i.e., number of specimens in each pool
n	Number of testing pools (parameter in the *PD-S* Model; decision variable in the *PD-J* Model)
Random Variables	
P	True (unknown) prevalence rate of the disease (with realization p)
$\hat{P}$	Maximum likelihood estimator (MLE) of P (with realization $\hat{p}$)
$T(m,n)$	Number of positive-testing pools among n pools, each containing m specimens
Objective Function and Performance Metrics	
$\sigma^2(m,n;p)$	Asymptotic variance of $\hat{P}$ for pool design (m,n), given a true prevalence rate of p
$MSE(\hat{P},m,n;p)$	Mean square error of $\hat{P}$ for pool design (m,n), given a true prevalence rate of p
$rBias(\%) = 100\left(\left\lvert\frac{\hat{p}-p}{p}\right\rvert\right)$	Relative bias of $\hat{p}$ (in percentage), given a true prevalence rate of p
Model Parameters	
p_0	An initial point estimate of P
c_f	Fixed testing cost per pool tested
c_v	Variable testing cost per specimen tested
B	Testing budget
$\gamma = \frac{c_f}{c_v}$	

C.2 Proofs

Proof of Lemma 1: To simplify the notation, we represent $\pi_0^S(m_1,m_2)$ as π_0. By Definition 1 and

Proposition 1, π_0 is the unique solution to:

$$\begin{aligned}
&\frac{1-(1-\pi_0)^{m_1}}{nm_1^2(1-\pi_0)^{m_1-2}} = \frac{1-(1-\pi_0)^{m_2}}{nm_2^2(1-\pi_0)^{m_2-2}} \\
\Leftrightarrow\ &\left(\frac{1}{1-\pi_0}\right)^{m_1-2} - (1-\pi_0)^2 = \left(\frac{m_1}{m_2}\right)^2\left[\left(\frac{1}{1-\pi_0}\right)^{m_2-2} - (1-\pi_0)^2\right] \\
\Leftrightarrow\ &\left(\frac{1}{1-\pi_0}\right)^{m_1} = 1 + \left(\frac{m_1}{m_2}\right)^2\left[\left(\frac{1}{1-\pi_0}\right)^{m_2} - 1\right].
\end{aligned}$$

This completes the proof. □

Proof of Lemma 2: To simplify the notation, we represent $\pi_0^S(m_1, m_2)$ as π_0, and let $\theta(m_1, m_2) \equiv 1-\pi_0$, for $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$. Then, $\theta(m_1, m_2) \in (0, 1)$. First, note that the equation provided in Lemma 1 is equivalent to:

$$(1-\pi_0)^{m_2} = (1-\pi_0)^{m_2-m_1}\Big(\frac{m_2^2}{m_2^2-m_1^2}\Big) - \Big(\frac{m_1^2}{m_2^2-m_1^2}\Big). \tag{C.1}$$

Part 1. Proof that $\pi_0^{\mathbf{S}}(\mathbf{m_1}, \mathbf{m_2})$ is decreasing in $\mathbf{m_1}$:

Taking the derivative of the RHS of Eqn. (C.1) with respect to m_1 yields:

$$\begin{aligned}
\frac{\partial}{\partial m_1}\left\{(1-\pi_0)^{m_2-m_1}\Big(\frac{{m_2}^2}{{m_2}^2-{m_1}^2}\Big) - \Big(\frac{{m_1}^2}{{m_2}^2-{m_1}^2}\Big)\right\} &= -\,[\theta(m_1,m_2)]^{m_2-m_1}\log[\theta(m_1,m_2)]\Big(\frac{m_2^2}{m_2^2-m_1^2}\Big) \\
&\quad + \frac{2m_1m_2^2}{(m_2^2-m_1^2)^2}[(\theta(m_1,m_2))^{m_1-m_2} - 1] > 0,
\end{aligned}$$

which follows because $\theta(m_1, m_2) < 1$ and $m_1 < m_2$, and hence, we have that $\log(\theta(m_1, m_2)) < 0$ and $(\theta(m_1, m_2))^{m_1-m_2} > 1$. Therefore, the RHS of Eqn (C.1) is increasing in m_1. Then, the LHS of Eqn. (C.1) must also increase in m_1 to preserve the equality, which in turn implies that $\theta(m_1, m_2)$ is increasing in m_1, and, equivalently, $\pi_0^S(m_1, m_2)$ is decreasing in m_1.

Part 2. Proof that $\pi_0^{\mathbf{S}}(\mathbf{m_1}, \mathbf{m_2})$ is decreasing in $\mathbf{m_2}$:

We prove this result by contradiction. Assume, to the contrary, that $\pi_0^S(m-2, m) > \pi_0^S(m-2, m-1)$ for some $m \in \mathbb{Z}^+ : m > 2$. From Part 1, we have that $\pi_0^S(m-1, m) < \pi_0^S(m-2, m)$, $\forall m \in \mathbb{Z}^+ : m > 2$. Therefore, we have two cases:

Case 1. $\pi_0^S(m-2, m-1) < \pi_0^S(m-1, m) < \pi_0^S(m-2, m)$:

Consider any $p \in \Big(\pi_0^S(m-1, m), \pi_0^S(m-2, m)\Big)$. Since $\pi_0^S(m-2, m-1) < \pi_0^S(m-1, m)$ (by assumption of this case), $p > \pi_0^S(m-2, m-1)$. Therefore, by Proposition 1, $\sigma^2(m-2, n; p) < \sigma^2(m-1, n; p)$. Further, since $\pi_0^S(m-1, m) < p < \pi_0^S(m-2, m)$, by Proposition 1, $\sigma^2(m-1, n; p) < \sigma^2(m, n; p)$, and $\sigma^2(m, n; p) < \sigma^2(m-2, n; p)$, leading to a contradiction, i.e., $\sigma^2(m-2, n; p) < \sigma^2(m, n; p)$ *and* $\sigma^2(m-2, n; p) > \sigma^2(m, n; p)$.

Hence, Case 1 is not possible.

Case 2. $\pi_0^S(m-1,m) < \pi_0^S(m-2,m-1) < \pi_0^S(m-2,m)$:
Consider any $p \in \Big(\pi_0^S(m-2,m-1), \pi_0^S(m-2,m)\Big)$. Since $\pi_0^S(m-2,m-1) > \pi_0^S(m-1,m)$ (by assumption of this case), $p > \pi_0^S(m-1,m)$. Therefore, by Proposition 1, $\sigma^2(m-1,n;p) < \sigma^2(m,n;p)$. Further, since $\pi_0^S(m-2,m-1) < p < \pi_0^S(m-2,m)$, by Proposition 1, $\sigma^2(m-2,n;p) < \sigma^2(m-1,n;p)$, and $\sigma^2(m,n;p) < \sigma^2(m-2,n;p)$, leading to a contradiction, i.e., $\sigma^2(m-2,n;p) > \sigma^2(m-1,n;p)$, *and* $\sigma^2(m-2,n;p) < \sigma^2(m-1,n;p)$. Hence, Case 2 is not possible.

From Cases 1 and 2, it follows that $\pi_0^S(m-2,m) < \pi_0^S(m-2,m-1)$, $\forall m \in \mathbb{Z}^+ : m > 2$, completing the proof. □

Proof of Corollary 1: By Lemma 2, $\pi_0^S(m-1,m) > \pi_0^S(m-1,m+1) > \pi_0^S(m,m+1)$, and the result follows. □

Proof of Corollary 2: The result follows directly from Lemmas 1 and 2. □

Proof of Lemma 3: Part 1. From Eqn. (5.2), we can derive:

$$\frac{\partial}{\partial p}\sigma^2(m,n;p) = \frac{1}{nm^2}\Big[\frac{m-2}{(1-p)^{m-1}} + 2(1-p)\Big] > 0 \ \forall m \geq 2 \text{ and } p < 1,$$

and the result follows.

Part 2. From the derivation in Part 1, it follows that $\frac{\partial}{\partial p}\sigma^2(1,n;p) = 1-2p > 0$, $\forall p < \frac{1}{2}$, and the result follows.

Part 3. Since $\sigma^2(m,n;p) = \frac{1-(1-p)^m}{nm^2(1-p)^{m-2}}$, the result trivially follows.

Part 4. We have the following derivatives:

$$\frac{\partial}{\partial m}\sigma^2(m,n;p) = \frac{-2}{m^3(1-p)^{m-2}} - \frac{\log(1-p)}{m^2(1-p)^{m-2}} + \frac{2(1-p)^2}{m^3}, \text{ and}$$

$$\begin{aligned}\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p) = \frac{6}{m^4}&\left\{\frac{m-2}{(1-p)^{m-1}} + 2(1-p)\right\} \\ &+ \frac{m(m-2)\log^2(1-p) - 2m\log(1-p) + 4(m-2)\log(1-p) - 4}{m^3(1-p)^{m-1}}.\end{aligned} \tag{C.2}$$

Consider the following:

$$\frac{\partial}{\partial p}\Big(\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p)\Big) = \frac{6(m-2) + 12m^4(1-p)^m + m^2(m-2)[\log(1-p)]^2 - \log(1-p)(8m-2m^2) - 4m}{m^4(1-p)^{m-1}}. \tag{C.3}$$

To find the root of $\frac{\partial}{\partial p}\Big(\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p)\Big)$, we define $x \equiv \log(1-p)$, and solve the following quadratic equation:

$$m^2(m-2)x^2 - 2m(4-m)x - 4m + 6(m-2) + 12m^4(1-p)^m = 0,$$

where $\Delta = b^2 - 4ac = m^2[48m^4(1-p)^m(2-m) - 4m^2 + 32m - 32] < 0, \forall m \in \mathbb{R}^+$. Thus, the quadratic equation has no real root. Further, $\frac{\partial}{\partial p}\Big(\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p)\Big)\Big|_{p=0} = \frac{12m^4+2m-12}{m^4} > 0, \forall m \geq 1$. Therefore, $\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p)$ starts out as a positive function in p, and is increasing in p, $\forall m \geq 1$, $p \in (0,1)$. Hence, $\frac{\partial^2}{\partial m^2}\sigma^2(m,n;p) > 0$, $\forall m \geq 1$ and $p \in (0,1)$, which implies that $\sigma^2(m,n;p)$ is strictly convex in m, $\forall m \geq 1$ and $p \in (0,1)$. $\square$

Proof of Proposition 2 For any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, we study the ratio $\frac{\sigma^2(m_1,n^*(m_1);p)}{\sigma^2(m_2,n^*(m_2);p)}$, which, from Lemma 4, equals:

$$\frac{\sigma^2(m_1,n^*(m_1);p)}{\sigma^2(m_2,n^*(m_2);p)} = \left(\frac{c_f + c_v m_1}{c_f + c_v m_2}\right)\Big(\frac{m_2}{m_1}\Big)^2 \left\{\frac{(1-p)^{m_2-2}\big[1-(1-p)^{m_1}\big]}{(1-p)^{m_1-2}\big[1-(1-p)^{m_2}\big]}\right\}, \text{ where}$$

$\lim_{p\to 0} \frac{\sigma^2(m_1,n^*(m_1);p)}{\sigma^2(m_2,n^*(m_2);p)} = \frac{c_f m_2 + c_v m_1 m_2}{c_f m_1 + c_v m_1 m_2} > 1$ (by L'Hospital's Rule and since $m_2 > m_1 \geq 1$), and
$\lim_{p\to 1} \frac{\sigma^2(m_1,n^*(m_1);p)}{\sigma^2(m_2,n^*(m_2);p)} = 0.$

Further, from [69], the ratio $\Big(\frac{m_2}{m_1}\Big)^2 \left\{\frac{(1-p)^{m_2-2}\big[1-(1-p)^{m_1}\big]}{(1-p)^{m_1-2}\big[1-(1-p)^{m_2}\big]}\right\}$ is decreasing in p, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$. Thus, Proposition 2 follows. $\square$

Proof of Lemma 5: To simplify the notation, we represent $\pi_0^J(m_1,m_2)$ as π_0. From Definition 2 and Proposition 2, π_0 is the unique solution to:

$$\begin{aligned}
&\left(\frac{c_f + c_v m_1}{B}\right)\left(\frac{1-(1-\pi_0)^{m_1}}{m_1^2(1-\pi_0)^{m_1-2}}\right) = \left(\frac{c_f + c_v m_2}{B}\right)\left(\frac{1-(1-\pi_0)^{m_2}}{m_2^2(1-\pi_0)^{m_2-2}}\right) \\
\Leftrightarrow\ &\left(\frac{1}{1-\pi_0}\right)^{m_1-2} - (1-\pi_0)^2 = \Big(\frac{m_1}{m_2}\Big)^2\Big(\frac{c_f + c_v m_2}{c_f + c_v m_1}\Big)\left[\left(\frac{1}{1-\pi_0}\right)^{m_2-2} - (1-\pi_0)^2\right] \\
\Leftrightarrow\ &\left(\frac{1}{1-\pi_0}\right)^{m_1} = 1 + \Big(\frac{m_1}{m_2}\Big)^2\Big(\frac{c_f + c_v m_2}{c_f + c_v m_1}\Big)\left[\left(\frac{1}{1-\pi_0}\right)^{m_2} - 1\right].
\end{aligned}$$

This completes the proof. $\square$

Proof of Lemma 6: We first show that $\pi_0^J(1,m)$ is decreasing in m, $\forall m \in \mathbb{Z}^+ : m \geq 2$. By Lemma 5, $\forall m \in \mathbb{Z}^+ : m \geq 2$, $\pi_0^J(1,m)$ is the unique solution to:

$$\pi_0^J(1,m) = \left(\frac{c_f + c_v m}{c_f + c_v}\right)\left\{\frac{1-\left(1-\pi_0^J(1,m)\right)^m}{m^2\left(1-\pi_0^J(1,m)\right)^{m-1}}\right\}. \tag{C.4}$$

Taking the derivative of the RHS of Eqn. (C.4) with respect to m yields:

$$\frac{\partial}{\partial m}RHS = \left(\frac{1}{c_f + c_v}\right)\left[\left(\frac{-2c_f - c_v m}{m^3}\right)\left(\frac{1}{\left(1-\pi_0^J(1,m)\right)^{m-1}} - \left(1-\pi_0^J(1,m)\right)\right) + \left(\frac{c_f + c_v m}{m^2}\right)\left(\frac{-\log\left(1-\pi_0^J(1,m)\right)}{\left(1-\pi_0^J(1,m)\right)^{m-1}}\right)\right] > 0,$$

where the last inequality follows because $m \geq 2$, leading to $\frac{c_f+c_v m}{m^2} \geq \frac{2c_f+c_v m}{m^3}$; and because $\log\left(1-\pi_0^J(1,m)\right) < -1$, $\forall \pi_0^J(1,m) < \frac{2}{3}$ (by Corollary 2), leading to $\frac{-\log\left(1-\pi_0^J(1,m)\right)}{\left(1-\pi_0^J(1,m)\right)^{m-1}} > \frac{1}{\left(1-\pi_0^J(1,m)\right)^{m-1}} - \left(1-\pi_0^J(1,m)\right)$. Thus the RHS of Eqn. (C.4) is increasing in m. Further, the RHS of Eqn. (C.4), which equals $\frac{B}{c_f+c_v}\sigma^2(m, n^*(m), \pi_0^J(1,m))$, is also increasing in $\pi_0^J(1,m)$. Thus, $\pi_0^J(1,m)$ must decrease as m increases so as to preserve the equality in Eqn. (C.4).

We are now ready to show that $\pi_0^J(m_1,m_2)$ is decreasing in each of m_1 and m_2, $\forall m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$.

Part 1. Proof that $\pi_0^J(m_1,m_2)$ is decreasing in m_1:

The proof follows by induction and contradiction. To this end, we first show that $\pi_0^J(2,m_2) < \pi_0^J(1,m_2)$, $\forall m_2 > 2$. Suppose, to the contrary, that $\pi_0^J(2,m_2) > \pi_0^J(1,m_2)$. Since $\pi_0^J(1,m_2) < \pi_0^J(1,2)$, $\forall m_2 > 2$, there are two cases:

Case 1. $\pi_0^J(2,m_2) > \pi_0^J(1,2)$: By Proposition 2, we have that $\forall p \in \left(\pi_0^J(1,m_2), \pi_0^J(1,2)\right)$, $\sigma^2(1,n^*(1);p) < \sigma^2(m_2,n^*(m_2);p) < \sigma^2(2,n^*(2);p)$, and $\sigma^2(1,n^*(1);p) > \sigma^2(2,n^*(2);p)$, leading to a contradiction.

Case 2. $\pi_0^J(1,m_2) < \pi_0^J(2,m_2) < \pi_0^J(1,2)$: By Proposition 2, we have that $\forall p \in \left(\pi_0^J(1,m_2), \pi_0^J(2,m_2)\right)$, $\sigma^2(1,n^*(1);p) < \sigma^2(m_2;n^*(m_2);p) < \sigma^2(2,n^*(2);p)$ (since $m_2 > 2$), but since $\pi_0^J(2,m_2) < \pi_0^J(1,2)$ (by assumption of this case), we also have that $\sigma^2(2,n^*(2);p) < \sigma^2(1,n^*(1);p)$ (by Proposition 2), leading to a contradiction.

Thus, it follows that $\pi_0^J(2,m_2) < \pi_0^J(1,m_2)$. It then follows, by induction, that $\pi_0^J(m_2-1,m_2) < \cdots < \pi_0^J(2,m_2) < \pi_0^J(1,m_2)$.

Part 2. Proof that $\pi_0^J(m_1,m_2)$ is decreasing in m_2:

From Part 1, we have that $\forall m_2 \in \mathbb{Z}^+$, $\pi_0^J(m_2-1,m_2) < \pi_0^J(m_2-2,m_2)$. We want to show that $\pi_0^J(m_2-2,m_2-1) > \pi_0^J(m_2-2,m_2)$. Suppose, to the contrary, that $\pi_0^J(m_2-2,m_2-1) < \pi_0^J(m_2-2,m_2)$. We have the following cases:

Case 1. $\pi_0^J(m_2-2,m_2-1) < \pi_0^J(m_2-1,m_2)$:
By Proposition 2, $\forall p \in \big(\pi_0^J(m_2-1,m_2), \pi_0^J(m_2-2,m_2)\big)$, $\sigma^2(m_2-1,n^*(m_2-1);p) < \sigma^2(m_2,n^*(m_2);p) < \sigma^2(m_2-2,n^*(m_2-2);p)$, and $\sigma^2(m_2-2,n^*(m_2-2);p) < \sigma^2(m_2-1,n^*(m_2-1);p)$, leading to a contradiction.

Case 2. $\pi_0^J(m_2-1,m_2) < \pi_0^J(m_2-2,m_2-1) < \pi_0^J(m_2-2,m_2)$:
By Proposition 2, $\forall p \in \big(\pi_0^J(m_2-2,m_2-1), \pi_0^J(m_2-2,m_2)\big)$, $\sigma^2(m_2-2,n^*(m_2-2);p) < \sigma^2(m_2-1,n^*(m_2-1);p) < \sigma^2(m_2,n^*(m_2);p)$, but $\sigma^2(m_2,n^*(m_2);p) < \sigma^2(m_2-2,n^*(m_2-2);p)$, leading to a contradiction.

Thus, it follows that $\pi_0^J(m_2-2,m_2-1) > \pi_0^J(m_2-2,m_2)$, completing the proof. □

Proof of Corollary 3: By Lemma 6, $\pi_0^J(m-1,m) > \pi_0^J(m-1,m+1) > \pi_0^J(m,m+1)$, and the result follows. □

Proof of Corollary 4: The results follow directly from Lemmas 5 and 6. □

Proof of Lemma 7: Parts 1 and 2. Observe that $\sigma^2(m,n^*(m);p) = \left(\frac{c_f+c_v m}{B}\right)\left(\frac{\sigma^2(m,n;p)}{n}\right)$. Then, the proof follows directly from Lemma 3.

Part 3. $\sigma^2(m,n^*(m);p)$ can be expressed as follows:

$$\begin{aligned}\sigma^2(m,n^*(m);p) &= \frac{1}{B}\left\{c_f\left[\frac{1-(1-p)^m}{m^2(1-p)^{m-2}}\right] + c_v\left[\frac{1-(1-p)^m}{m(1-p)^{m-2}}\right]\right\} \\ &= \frac{1}{B}\left[\frac{1}{(1-p)^{m-2}} - (1-p)^2\right]\left(\frac{c_f}{m^2}+\frac{c_v}{m}\right).\end{aligned}$$

Let $g(m;p) \equiv B\sigma^2(m,n^*(m);p)$. We have the following:

$$\frac{\partial}{\partial m}g(m;p) = \left[-\frac{\log(1-p)}{(1-p)^{m-2}}\right]\left(\frac{c_f}{m^2}+\frac{c_v}{m}\right) + \left[\frac{1}{(1-p)^{m-2}} - (1-p)^2\right]\left(\frac{-2c_f}{m^3} - \frac{c_v}{m^2}\right), \text{ and}$$

$$\begin{aligned}\frac{\partial^2}{\partial m^2}g(m;p) = &\left[\frac{\big[\log(1-p)\big]^2}{(1-p)^{m-2}}\right]\left(\frac{c_f}{m^2}+\frac{c_v}{m}\right) + 2\left[\frac{\log(1-p)}{(1-p)^{m-2}}\right]\left(\frac{2c_f}{m^3}+\frac{c_v}{m^2}\right) \\ &+ \left[\frac{1}{(1-p)^{m-2}} - (1-p)^2\right]\left(\frac{6c_f}{m^4}+\frac{2c_v}{m^3}\right).\end{aligned}$$

We now show that $\frac{\partial^2}{\partial m^2} g(m;p) > 0$, $\forall m > 0$ and $\forall p \in (0, \frac{1}{3})$. Noting that $(1-p)^{m-2} > 0$, $\forall m > 0$ and $p \in (0, \frac{1}{3})$, and multiplying $\frac{\partial^2}{\partial m^2} g(m;p)$ by $\left\{(1-p)^{m-2}\right\}$ yields the following :

$$h(m;p) = \Big(\log(1-p)\Big)^2 \left(\frac{c_f}{m^2} + \frac{c_v}{m}\right) + 2\log(1-p)\left(\frac{2c_f}{m^3} + \frac{c_v}{m^2}\right) + \left[1-(1-p)^m\right]\left(\frac{6c_f}{m^4} + \frac{2c_v}{m^3}\right).$$

$$\frac{\partial}{\partial p} h(m;p) = \frac{-2\log(1-p)}{1-p}\left(\frac{c_f}{m^2} + \frac{c_v}{m}\right) - \frac{2}{1-p}\left(\frac{2c_f}{m^3} + \frac{c_v}{m^2}\right) + (1-p)^{m-1}\left(\frac{6c_f}{m^3} + \frac{2c_v}{m^2}\right).$$

Note that the sign of $\frac{\partial}{\partial p} h(m;p)$ equals the sign of the following function:

$$\begin{aligned}
&-2\log(1-p)\left(\frac{c_f}{m^2} + \frac{c_v}{m}\right) - 2\left(\frac{2c_f}{m^3} + \frac{c_v}{m^2}\right) + (1-p)^m\left(\frac{6c_f}{m^3} + \frac{2c_v}{m^2}\right) \\
&= -\frac{2\log(1-p)c_v}{m} + \frac{1}{m^2}\Big(-2c_f\log(1-p) - 2c_v + 2c_v(1-p)^m\Big) + \frac{c_f}{m^3}\Big(-4 + 6(1-p)^m\Big) > 0,
\end{aligned}$$

where the last inequality follows because $c_f > c_v$ and $\log(1-p) < 0$, $\forall p \in \left(0, \frac{1}{3}\right)$. Thus, $\frac{\partial^2}{\partial m^2} g(m;p)$ is increasing in p, $\forall m > 0$ and $\forall p \in \left(0, \frac{1}{3}\right)$. At $p = 0$, $\frac{\partial^2}{\partial m^2} g(m;p) = \frac{6c_f}{m^4} + \frac{2c_v}{m^3} > 0$. Thus, $\frac{\partial^2}{\partial m^2} g(m;p) > 0$, $\forall m > 0$ and $\forall p \in \left(0, \frac{1}{3}\right)$. This completes the proof. $\square$

Proof of Lemma 8:

Part 1. The proof follows directly from Lemma 2.

Part 2. From Lemma 5, we have the following:

$$\frac{m_2^2(1-\pi_0^J(m_1,m_2))^{m_2}\left[1-(1-\pi_0^J(m_1,m_2))^{m_1}\right]}{m_1^2(1-\pi_0^J(m_1,m_2))^{m_1}\left[1-(1-\pi_0^J(m_1,m_2))^{m_2}\right]} = \frac{c_f + c_v m_2}{c_f + c_v m_1} = \frac{\gamma + m_2}{\gamma + m_1} = 1 + \frac{(m_2 - m_1)}{\gamma + m_1}.$$

Note that $\frac{m_2^2(1-\pi_0^J(m_1,m_2))^{m_2}\left[1-(1-\pi_0^J(m_1,m_2))^{m_1}\right]}{m_1^2(1-\pi_0^J(m_1,m_2))^{m_1}\left[1-(1-\pi_0^J(m_1,m_2))^{m_2}\right]}$ is decreasing in $\pi_0^J(m_1,m_2)$ (from Liu *et al* [69]) and is constant in γ, while $1 + \frac{(m_2-m_1)}{\gamma+m_1}$ is decreasing in γ (since $m_2 > m_1$) and is constant in $\pi_0^J(m_1,m_2)$. Thus, as γ increases, $\pi_0^J(m_1,m_2)$ also increases. This completes the proof. $\square$

Proof of Lemma 9: From Lemmas 1 and 5, we observe that when $c_v = 0$, i.e., when $\gamma \to \infty$, $\pi_0^J(m_1,m_2) \to \pi_0^S(m_1,m_2)$. Further, from Lemma 8, $\pi_0^J(m_1,m_2)$ is increasing in γ. Thus, $\pi_0^J(m_1,m_2) < \pi_0^S(m_1,m_2)$, completing the proof. $\square$

Proof of Lemma 10: By Properties 1 and 2, and Corollaries 1 and 3, we have that $\pi_0^X(m,m+1) \leq \pi_0^X(m-1,m)$, $\forall m \in \mathbb{Z}^+ : m > 2$, $X \in \{S, J\}$. Then, we have that $\sigma^2(m,n;p) < \sigma^2(k,n;p)$, $\forall k \in \mathbb{Z}^+ : k \neq m$, if and only if $\pi_0^X(m,m+1) < p < \pi_0^X(m-1,m)$. If $p = \pi_0^X(m,m+1)$, then $\sigma^2(m+1,n;p) = \sigma^2(m,n;p)$,

and, if $p = \pi_0^X(m-1, m)$, then $\sigma^2(m, n; p) = \sigma^2(m-1, n; p)$. This completes the proof. $\square$

of Theorem 1: From Lemma 10, Part 1, we have the following:

$\cdots\ \pi_0^S(m+1, m+2) < p < \pi_0^S(m, m+1)$	$\pi_0^S(m, m+1) < p < \pi_0^S(m-1, m)$	$\pi_0^S(m-1, m) < p < \pi_0^S(m-2, m-1)\ \cdots$
$\sigma^2(m+1, n; p) < \sigma^2(k, n; p)$, $\forall k \in \mathbb{Z}^+ : k \neq m+1$	$\sigma^2(m, n; p) < \sigma^2(k, n; p)$, $\forall k \in \mathbb{Z}^+ : k \neq m$	$\sigma^2(m-1, n; p) < \sigma^2(k, n; p)$, $\forall k \in \mathbb{Z}^+ : k \neq m-1$.

Then, if $\pi_0^S(m, m+1) < p < \pi_0^S(m-1, m)$, then m is the unique optimal solution to *PD-S*. If $p = \pi_0^S(m, m+1)$, then m and $m+1$ are both optimal, and, if $p = \pi_0^S(m-1, m)$, then $m-1$ and m are both optimal for *PD-S*. This completes the proof. $\square$

Proof of Theorem 2: From Lemma 10, Part 2, we have the following:

$\cdots\ \pi_0^J(m+1, m+2) < p < \pi_0^J(m, m+1)$	$\pi_0^J(m, m+1) < p < \pi_0^J(m-1, m)$	$\pi_0^J(m-1, m) < p < \pi_0^J(m-2, m-1)\ \cdots$
$\sigma^2(m+1, n*(m+1); p) < \sigma^2(k, n^*(k); p)$, $\forall k \in \mathbb{Z}^+ : k \neq m+1$	$\sigma^2(m, n^*(m); p) < \sigma^2(k, n^*(k); p)$, $\forall k \in \mathbb{Z}^+ : k \neq m$	$\sigma^2(m-1, n^*(m-1); p) < \sigma^2(k, n^*(k); p)$, $\forall k \in \mathbb{Z}^+ : k \neq m-1$.

Then, if $\pi_0^J(m, m+1) < p < \pi_0^J(m-1, m)$, then m is the unique optimal solution to *PD-J*. If $p = \pi_0^J(m, m+1)$, then m and $m+1$ are both optimal, and, if $p = \pi_0^J(m-1, m)$, then $m-1$ and m are both optimal for *PD-J*. This completes the proof. $\square$

Proof of Corollary 8: The result follows from Lemma 10 and Theorems 1 and 2. $\square$

Proof of Corollary 7: The result follows from Lemma 8. $\square$

Proof of Lemma 11: By Lemmas 3 and 7, m' is the unique solution to the first-order condition:

$$\frac{-2}{(m')^3(1-p_0)^{m'-2}} - \frac{\log(1-p_0)}{(m')^2(1-p_0)^{m'-2}} + \frac{2(1-p_0)^2}{(m')^3} = 0, \text{ for } PD\text{-}S, \text{ and}$$
$$\left(\frac{c_f}{(m')^2} + \frac{c_v}{m'}\right)\left[-\frac{\log(1-p_0)}{(1-p_0)^{m'-2}}\right] - \left(\frac{2c_f}{(m')^3} + \frac{c_v}{(m')^2}\right)\left[\frac{1}{(1-p_0)^{m'-2}} - (1-p_0)^2\right] = 0, \text{ for } PD\text{-}J,$$

which are respectively equivalent to:

$$m' = \frac{2[(1-p_0)^{m'} - 1]}{\log(1-p_0)}, \text{ for } PD\text{-}S, \text{ and}$$
$$m' = \frac{\left(1 + \frac{c_f}{c_f + c_v m'}\right)[(1-p_0)^{m'} - 1]}{\log(1-p_0)}, \text{ for } PD\text{-}J.$$

This completes the proof. $\square$

Appendix D

Appendix to Chapter 5

D.1 Summary of Notation

Table D.1: Summary of Notation – Chapter 5

Decision Variables	
$m^{(s)}$	Pool size, i.e., number of specimens in each pool, used in stage s
$n^{(s)}$	Number of testing pools used in stage s
Random Variables	
P	True (unknown) prevalence rate of the disease (with realization p) (P follows an arbitrary continuous distribution with support $[p_{LB}, p_{UB}]$)
$\hat{P}^{(s)}$	Maximum likelihood estimator (MLE) of P obtained in stage s
$T(m^{(s)}, n^{(s)})$	Number of positive-testing pools among $n^{(s)}$ pools, each containing $m^{(s)}$ specimens
Objective Functions and Performance Metrics	
$\sigma^2(m, n; p)$	Asymptotic variance of $\hat{P}$ for pool design (m, n), given a true prevalence rate of p
$Regret(m, n; p)$	Regret for pool design (m, n), given a true prevalence rate of p
$MSE(\hat{P}, m, n; p)$	Mean squared error of $\hat{P}$ for pool design (m, n), given a true prevalence rate of p
$rBias(\hat{P}, m, n; p)(\%) = 100\left(\left\lvert \frac{\hat{P}-p}{p} \right\rvert\right)$	Relative bias of $\hat{P}$ (in percentage) for pool design (m, n), given a true prevalence rate of p
Model Parameters	
p_{LB}, p_{UB}	Lower and upper bounds of the support of P (input for RM and MM)
p_0	An initial point estimate of P (input for DM)
c_f	Fixed testing cost per pool tested
c_v	Variable testing cost per specimen tested
B	Testing budget
λ	Budget allocation factor

D.2 Properties of the Asymptotic Variance Function

In this section, we summarize some key properties of the asymptotic variance function, $\sigma^2(m, n; p)$, established in [76].

In this setting, we first present some relevant properties of the asymptotic variance function, $\sigma^2(m,n;p)$.

Lemma A1. (From [69] and [76]) In general, $\sigma^2(m,n;p)$ has the following properties:

1. For any $m,n \in \mathbb{Z}^+$, $\sigma^2(m,n;p)$ is increasing in p, $\forall p < \frac{1}{2}$.

2. For a given $n \in \mathbb{Z}^+$, and for any $m_1, m_2 \in \mathbb{Z}^+ : m_1 < m_2$, the *prevalence threshold*, $\pi_0(m_1,m_2) \in (0,1)$, is defined as the prevalence rate at which $\sigma^2\big(m_1,n;\pi_0(m_1,m_2)\big) = \sigma^2\big(m_2,n;\pi_0(m_1,m_2)\big)$. Then, for any $m_1, m_2 \in \mathbb{Z}^+$: $m_1 < m_2$, there exists a unique $\pi_0(m_1,m_2) \in (0,1)$. Further:

$$\sigma^2(m_1,n;p) \begin{cases} > \sigma^2(m_2,n, & \forall p < \pi_0(m_1,m_2) \\ < \sigma^2(m_2,n, & \forall p > \pi_0(m_1,m_2) \end{cases},$$

and $\pi_0(m_1,m_2)$ is decreasing in each of m_1 and m_2, $\forall m_1,m_2 \in \mathbb{Z}^+ : m_1 < m_2$. For a given $n \in \mathbb{Z}^+$, $\pi_0(m_1,m_2)$ is computed using the following equation ([76]):

$$\left(\frac{1}{1-\pi_0}\right)^{m_1} = 1 + \left(\frac{m_1}{m_2}\right)^2 \left[\left(\frac{1}{1-\pi_0}\right)^{m_2} - 1\right].$$

Thus, $\pi_0(1,2) = \frac{2}{3}$, and $\pi_0(m_1,m_2) < \frac{2}{3}$, $\forall m_1,m_2 \in \mathbb{Z}^+ : m_1 < m_2$. Further, the maximum pool size feasible, given a testing budget, B, a testing cost, c_f, a collection cost, c_v, and a fixed n is given by: $\overline{M} \equiv \lfloor \frac{B - c_f n}{c_v n} \rfloor$.

Further, for any $n \in \mathbb{Z}^+$, the optimal pool size to *DM*, $m_D^*(n,p_0)$, follows a threshold policy, as defined in Theorem 3, with p_{UB} replaced by p_0. Further, $m_D^*(p_0)$ is decreasing in p_0, $\forall p_0 \geq \pi_0(1,2)$. For a given c_v, $m_D^*(p_0)$ is non-decreasing in c_f; for a given c_f, $m_D^*(p_0)$ is non-increasing in c_v ([76]).

D.3 Proofs

Proof of Theorem 3: From Lemma A1, it follows that $\max\limits_{P \in [p_{LB}, p_{UB}]} \sigma^2(m,n^*(m);p) \equiv \sigma^2(m,n^*(m);p_{UB})$, for any $m \in \mathbb{Z}^+$ and $p_{UB} < \frac{1}{2}$. Thus, Model *MM* can be reformulated as Model *DM* with $p_0 = p_{UB}$. From [76], it follows that Model *MM* can be solved using the threshold policy stated in Theorem 3. □

Proof of Corollary 8: The proof follows from Lemma A1 with p_0 replaced by p_{UB}. □

Proof of Corollary 9: The proof follows directly from Corollary 8 because $p_0 \in [p_{LB}, p_{UB}]$. □

Proof of Lemma 12: We first derive, for a given $n \in \mathbb{Z}^+, n \leq \left\lfloor \frac{B}{c_f + c_v} \right\rfloor$, $\forall m \in \mathbb{Z}^+ : m \geq 2$:

$$\frac{\partial}{\partial p} Regret(m,n;p) = \frac{1}{n}\left\{ \frac{1}{m^2}\Big[\frac{(m-2)}{(1-p)^{m-1}} + 2(1-p)\Big] - \frac{1}{\big(m^*(n,p)\big)^2}\Big[\frac{\big(m^*(n,p)-2\big)}{(1-p)^{m^*(n,p)-1}} + 2(1-p)\Big]\right\} = 0.$$

By the Envelope Theorem, $\frac{\partial}{\partial p}\sigma^2(m^*(n,p),n;p) = \frac{\partial}{\partial p}\sigma^2(m,n;p)\Big|_{m=m^*(n,p)}$. Thus, we can derive:

$$\begin{aligned}\frac{\partial}{\partial p}Regret(m,n;p) &= \frac{1}{n}\left\{\frac{\partial}{\partial p}\sigma^2(m,n;p) - \frac{\partial}{\partial p}\sigma^2(m^*(n,p),n;p)\right\} \\ &= \frac{1}{n}\left\{\frac{1}{m^2}\Big[\frac{(m-2)}{(1-p)^{m-1}} + 2(1-p)\Big] - \frac{1}{\big(m^*(n,p)\big)^2}\Big[\frac{\big(m^*(n,p)-2\big)}{(1-p)^{m^*(n,p)-1}} + 2(1-p)\Big]\right\}.\end{aligned} \tag{D.1}$$

Without loss of generality, in the remainder of this proof, we assume $n = 1$. To simplify the notation, we represent $m_D^*(1,p)$ as $m^*(p)$, and $\overline{M}(1)$ as $\overline{M}$.

Part 1. From the threshold policy to solve *DM*, $\forall p \in \big(\pi_0(m,m+1), \pi_0(m-1,m)\big)$, $m^*(p) = m$ ([76]). Thus, it follows that $Regret(m,n;p) = \sigma^2(m,n;p) - \sigma^2(m^*(p),n;p) = 0$.

Part 2. By Eqn. (D.1), since $\pi_0(1,2) = \frac{2}{3}$, we have that $m^*(1) = 1$ and $m^*(0) = \overline{M}$, $\forall n \in \mathbb{Z}^+, n \leq \Big\lfloor \frac{B}{c_f + c_v} \Big\rfloor$. Thus, we have the following:

$$\begin{aligned}&\lim_{p\to 0}\frac{\partial}{\partial p}Regret(m,n;p) = \frac{1}{m} - \frac{1}{\overline{M}} \geq 0, \text{ and} \\ &\lim_{p\to 1}\frac{\partial}{\partial p}Regret(m,n;p) = \lim_{p\to 1}\left\{\frac{1}{m^2}\Big[\frac{(m-2)}{(1-p)^{m-1}}\Big] - \Big[\frac{-1}{(1-p)^0}\Big]\right\} \to \infty.\end{aligned}$$

To simplify the notation, we represent $\tilde{p}(m)$ as $\tilde{p}$. Recall that $\tilde{p}$ is the smallest solution to the FOC, i.e, $\tilde{p} = \arg\min\Big\{\frac{\partial}{\partial p}Regret(m,n;p) = 0\Big\}$. Further, by Part 1, since $Regret(m,n;p) \geq 0$, $\forall m \in \mathbb{Z}^+$, $\frac{\partial}{\partial p}Regret(m,n;p) = 0$, $\forall p \in \big(\pi_0(m,m+1), \pi_0(m-1,m)\big)$. Then, since $\lim_{p\to 0}\frac{\partial}{\partial p}Regret(m,n;p) \geq 0$, and $\lim_{p\to 1}\frac{\partial}{\partial p}Regret(m,n;p) > 0$, $\tilde{p} < \pi_0(m,m+1)$, and $\tilde{p}$ must correspond to a local maximum of the $Regret(m,n;p)$ function with respect to p.

Part 3. From Lemma A1, $m^*(p)$ is decreasing in p, for any $0 < p < \frac{1}{2}$. Thus, for a given $m \geq 2$, $m \in \mathbb{Z}^+$, since $\tilde{p} < \pi_0(m,m+1)$, there are two cases: 1. $m^*(\tilde{p}) > m$, and 2. $m^*(\tilde{p}) = m$.

Case 1. $m^*(\tilde{p}) > m$: By definition of $\tilde{p}$, $\tilde{p}$ is the solution to the FOC given by:

$$\begin{aligned}&\frac{1}{m^2}\Big[\frac{m-2}{(1-\tilde{p})^{m-1}}\Big] + \frac{2(1-\tilde{p})}{m^2} = \frac{m^*(\tilde{p})-2}{[m^*(\tilde{p})]^2(1-\tilde{p})^{m^*(\tilde{p})-1}} + \frac{2(1-\tilde{p})}{[m^*(\tilde{p})]^2} \\ \Leftrightarrow &\frac{(m-2)[m^*(\tilde{p})]^2(1-\tilde{p})^{m^*(\tilde{p})-1} - [m^*(\tilde{p})-2]m^2(1-\tilde{p})^{m-1}}{m^2 - [m^*(\tilde{p})]^2} = 2(1-\tilde{p})^{m+m^*(\tilde{p})-1}.\end{aligned} \tag{D.2}$$

We will now show that $\tilde{p}$ is the unique solution to the FOC in the interval $\big(0, \pi_0(m,m+1)\big)$, i.e, when $m^*(\tilde{p}) > m$. Eqn. (D.2) is equivalent to the following:

$$(m^*(\tilde{p})-2)m^2(1-\tilde{p})^{m-1}-(m-2)[m^*(\tilde{p})]^2(1-\tilde{p})^{m^*(\tilde{p})-1}=2\Big\{[m^*(\tilde{p})]^2-m^2\Big\}(1-\tilde{p})^{m+m^*(\tilde{p})-1}. \quad \text{(D.3)}$$

Taking the derivative of both sides of Eqn. (D.3) with respect to $\tilde{p}$, we have:

$$\begin{aligned}\frac{\partial}{\partial\tilde{p}}LHS=&-(m^*(\tilde{p})-2)(m-1)m^2(1-\tilde{p})^{m-2}+(m^*(\tilde{p})-1)[m^*(\tilde{p})]^2(m-2)(1-\tilde{p})^{m^*(\tilde{p})-2}+m^2(1-\tilde{p})^{m-1}\\&-\Big\{(m+2)(1-\tilde{p})^{m^*(\tilde{p})-1}[2m^*(\tilde{p})+\log(1-\tilde{p})]\Big\}\frac{\partial m^*(\tilde{p})}{\partial\tilde{p}};\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\tilde{p}}RHS=&-2\{[m^*(\tilde{p})]^2-m^2\}(1-\tilde{p})^{m+m^*(\tilde{p})-2}(m+m^*(\tilde{p})-1)\\&+\left\{2(1-\tilde{p})^{m+m^*(\tilde{m})-1}\left\{2m^*(\tilde{p})+\left([m^*(\tilde{p})]^2-m^2\right)\log(1-\tilde{p})\right\}\right\}\frac{\partial m^*(\tilde{p})}{\partial\tilde{p}}.\end{aligned}$$

Further, since $m^*(p)$ is decreasing in p for any $0<p<\frac{1}{2}$ ([76]), $\frac{\partial m^*(\tilde{p})}{\partial\tilde{p}}\leq 0$. Furthermore, by the assumption that $m^*(\tilde{p})>m$, $\frac{\partial}{\partial m^*(\tilde{p})}LHS>0$, while $\frac{\partial}{\partial\tilde{p}}RHS<0$, for $0<\tilde{p}<1$. Therefore, there exists a unique $\tilde{p}$ such that $\tilde{p}<\pi_0(m,m+1)$, i.e., $\tilde{p}$ is the unique solution to the FOC in the interval $\big(0,\pi_0^S(m,m+1)\big)$.

Case 2. $m=m^*(\tilde{p})$: $\pi_0(m,m+1)\leq\tilde{p}\leq\pi_0(m-1,m)$. This contradicts with the property that $\tilde{p}<\pi_0(m,m+1)$, and with the property that $\tilde{p}$ corresponds to a local maximum of $Regret(m,n;p)$. Therefore, we conclude that there exists a unique $\tilde{p}$ such that $\tilde{p}<\pi_0(m,m+1)$, $\forall m\geq 2$.

We now show that the FOC of the $Regret(m,n;p)$ function has no solution in the interval $\big(\pi_0(m-1,m),1\big)$. We prove this property by contradiction. Assume that there exists a solution, $\bar{p}$, to the FOC of $Regret(m,n;p)$ such that $\bar{p}>\pi_0(m-1,m)$, i.e., $m^*(\bar{p})<m$, for a given $m\geq 2$, $m\in\mathbb{Z}^+$. Since $m^*(\bar{p})<m$, in order for Eqn. (D.2) to hold, the following needs to hold:

$$\begin{aligned}&(m-2)[m^*(\bar{p})]^2(1-\bar{p})^{m^*(\bar{p})-1}>[m^*(\bar{p})-2]m^2(1-\bar{p})^{m-1}\\\Leftrightarrow\ &\frac{m-2}{m^2}\Big[\frac{1}{(1-\bar{p})^{m-1}}\Big]>\frac{m^*(\bar{p})-2}{[m^*(\bar{p})]^2}\Big[\frac{1}{(1-\bar{p})^{m^*(\bar{p})-1}}\Big]\end{aligned} \quad \text{(D.4)}$$

Since $m^*(\bar{p})$ is the *DM* optimal solution for $p_0 = \bar{p}$, we have that:

$$\begin{aligned} &\frac{1}{[m^*(\bar{p})]^2}\Big[\frac{1}{(1-\bar{p})^{m^*(\bar{p})-2}} - (1-\bar{p})^2\Big] < \frac{1}{m^2}\Big[\frac{1}{(1-\bar{p})^{m-2}} - (1-\bar{p})^2\Big] \\ \Leftrightarrow\ &\frac{1}{(1-\bar{p})^{m-1}} < \frac{1}{(1-\bar{p})^{m^*(\bar{p})-1}} \text{ (since } m > m^*(\bar{p}) \text{ by assumption).} \end{aligned} \tag{D.5}$$

Further, $\frac{m-2}{m^2}$ is decreasing in m, equivalently, $\frac{m-2}{m^2} < \frac{m^*(\bar{p})-2}{[m^*(\bar{p})]^2}$, $\forall m > m^*(\bar{p})$. Therefore, Eqs. (D.4) and (D.5) lead to a contradiction. Thus, we conclude that $m < m^*(\bar{p})$, i.e., the FOC of the $Regret(m, n; p)$ function has no solution in the interval $\big(\pi_0(m-1, m), 1\big)$. This completes the proof. $\square$

Proof of Lemma 13: From Lemma 12, the result follows because $Regret(m, n; p)$ has a unique local maximum at $\tilde{p}(m)$, and $p \in [p_{LB}, p_{UB}]$. Thus, $\max\limits_{p\in[p_{LB},p_{UB}]}\Big\{Regret(m, n; p)\Big\}$ is attained at either $\tilde{p}(m)$, p_{LB} or p_{UB} for any $m \in \mathbb{Z}^+ : m \geq 1$. $\square$

Proof of Theorem 4:

From Lemma 13, for any $n \in \mathbb{Z}^+$, $n \leq \Big\lfloor \frac{B}{c_f + c_v} \Big\rfloor$, $p^*(m) \equiv \underset{p\in\{p_{LB},p_{UB},\tilde{p}(m)\}}{\arg\max}\{Regret(m, n; p)\}$. Thus, the result follows by definition of *RM*. $\square$

D.4 Additional Numerical Results

Table D.2: Comparison of *DM, MM* and *RM* for two-stage estimation frameworks ($\lambda \in \{0.25, 0.5, 1\}$) with $B = \$16,320$ and inaccurate input parameters; $\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$ and MSE values are multiplied by 10^6 (average ± half width of 95% CI)

	Models	***DM-S*** ($\lambda = 1$)	***MM-S*** ($\lambda = 1$)	***RM-S*** ($\lambda = 1$)	***DM-S*** ($\lambda = 0.5$)	***MM-S*** ($\lambda = 0.5$)	***DM-S*** ($\lambda = 0.25$)	***MM-S*** ($\lambda = 0.25$)
	(m^*, n^*)	(18,113)	(12,136)	(14,127)	(19,55)	(12,68)	(18,28)	(12,34)
Beta ∼	$\hat{p}$	0.04897 ±0.00108	0.04709 ±0.00072	0.04706 ±0.00078	0.04550 ±0.00159	0.04547 ± 0.00155	0.04641 ±0.00159	0.04735 ±0.00161
(1.59, 32.56)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	54.70 ± 2.48	44.70 ± 1.16	47.00 ± 1.70	45.40 ± 6.94	40.30 ± 6.42	45.60 ± 7.46	43.20 ± 5.17
$\mu_P = 0.0465$	MSE	1220 ± 525	44.90 ± 2.19	210 ± 18.60	49.30 ± 3.41	45.70 ± 2.13	45.00 ± 3.23	43.20 ± 2.38
$\sigma_P^2 = 0.00135$	$rBias(\%)$	13.68 ±0.40	13.23 ±0.26	13.16 ±0.29	12.89 ±0.61	12.78 ± 0.62	12.61 ±0.59	12.48 ±0.51
Beta ∼	$\hat{p}$	0.07146 ±0.00101	0.07046 ±0.00073	0.07011 ±0.00075	0.07179 ±0.00165	0.06990 ± 0.00158	0.06938 ±0.00161	0.07055 ±0.00160
(3.52,46.92)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	88.10 ± 2.39	71.60 ± 1.26	74.60 ± 1.48	81.10 ± 9.38	69.20 ± 9.48	73.30 ± 10.80	70.20 ± 5.76
$\mu_P = 0.0698$	MSE	1030 ± 475	76.20 ± 3.39	139 ± 112	84.00 ± 3.78	72.70 ± 2.46	88.70 ± 3.56	69.90 ± 2.40
$\sigma_P^2 = 0.00135$	$rBias(\%)$	10.62 ±0.32	9.95 ±0.16	9.95 ±0.16	9.81 ±0.34	9.75 ± 0.33	10.01 ±0.36	9.64 ±0.34
Beta ∼	$\hat{p}$	0.02483 ±0.00112	0.02467 ±0.00085	0.02418 ±0.00089	0.02437 ±0.00122	0.02143 ± 0.00154	0.02403 ±0.00162	0.02235 ±0.00155
(0.40,16.61)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	44.40 ± 9.06	26.80 ± 2.30	30.20 ± 3.83	29.30 ± 9.12	20.40 ± 6.47	26.60 ± 7.47	20.20 ± 5.74
$\mu_P = 0.0232$	MSE	1270 ± 521	295 ± 218	450 ± 279	356 ± 3.87	241 ± 2.81	269 ± 3.67	232 ± 2.24
$\sigma_P^2 = 0.00135$	$rBias(\%)$	41.73 ±1.65	46.06 ±3.31	45.51 ±2.92	43.72 ±5.36	48.13 ± 10.21	68.27 ±41.28	40.45 ±6.92

Table D.3: Comparison of *DM, MM* and *RM* for two-stage estimation frameworks ($\lambda \in \{0.25, 0.5, 1\}$) with $B = \$65,280$ and inaccurate input parameters; $\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$ and MSE values are multiplied by 10^6 (average ± half width of 95% CI)

	Models	***DM-S*** ($\lambda = 1$)	***MM-S*** ($\lambda = 1$)	***RM-S*** ($\lambda = 1$)	***DM-S*** ($\lambda = 0.5$)	***MM-S*** ($\lambda = 0.5$)	***DM-S*** ($\lambda = 0.25$)	***MM-S*** ($\lambda = 0.25$)
	(m^*, n^*)	(18,453)	(11,562)	(14,510)	(18,226)	(11,281)	(18,113)	(12,136)
Beta ∼	$\hat{p}$	0.04664 ±0.00070	0.04700 ±0.00071	0.04672 ±0.00070	0.04530 ±0.00146	0.04771 ± 0.00160	0.04687 ±0.00160	0.04521 ±0.00153
(1.59, 32.56)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	13.40 ± 0.95	11.30 ± 0.56	11.60 ± 0.89	10.40 ± 1.21	10.80 ± 0.99	10.90 ± 1.40	9.94 ± 9.21
$\mu_P = 0.0465$	MSE	13.50 ± 0.67	11.20 ± 0.28	11.60 ± 0.40	10.70 ± 0.58	10.50 ± 0.54	10.30 ± 0.72	9.41 ± 0.50
$\sigma_P^2 = 0.00135$	$rBias(\%)$	6.40 ±0.12	6.68 ±0.13	6.48 ±0.13	6.05 ±0.25	6.07 ± 0.24	4.93 ±0.16	4.73 ±0.17
Beta ∼	$\hat{p}$	0.07047 ±0.00053	0.07041 ±0.00050	0.06972 ±0.00050	0.07099 ±0.00159	0.07064 ± 0.00161	0.06865 ±0.00156	0.07076 ±0.00154
(3.52,46.92)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	22.60 ± 0.47	19.70 ± 0.32	17.70 ± 0.21	19.30 ± 1.99	17.40 ± 1.23	17.30 ± 1.61	17.20 ± 1.40
$\mu_P = 0.0698$	MSE	106 ± 93.92	20.20 ± 0.80	17.90 ± 0.52	19.50 ± 0.83	16.90 ± 0.60	18.20 ± 0.69	17.10 ± 0.57
$\sigma_P^2 = 0.00135$	$rBias(\%)$	5.01 ±0.07	4.90 ±0.05	4.93 ±0.06	4.73 ±0.16	4.88 ± 0.16	4.93 ±0.16	4.73 ±0.17
Beta ∼	$\hat{p}$	0.02375 ±0.00085	0.02288 ±0.00068	0.02374 ±0.00070	0.02328 ±0.00149	0.02299 ± 0.00155	0.02231 ±0.00151	0.02295 ±0.00146
(0.40,16.61)	$\sigma^2((\mathbf{m}^*, \mathbf{n}^*); p)$	16.20 ± 13.60	5.87 ± 0.76	6.69 ± 1.61	5.59 ± 1.61	5.28 ± 1.08	5.77 ± 1.09	4.93 ± 0.82
$\mu_P = 0.0232$	MSE	339 ± 248.00	6.05 ± 0.34	7.42 ± 0.46	6.24 ± 0.62	5.55 ± 0.66	5.44 ± 0.94	5.44 ± 0.45
$\sigma_P^2 = 0.00135$	$rBias(\%)$	29.76 ±4.37	29.94 ±1.92	28.99 ±2.94	26.74 ±3.28	26.42 ± 3.67	33.23 ±9.69	22.84 ±1.95