Optimizing Pooled Testing for Estimating the Prevalence of Multiple Diseases

Md S. Warasi[1,*], Laura L. Hungerford[2], Kevin Lahmers[2]

[1]Department of Mathematics and Statistics, Radford University, Radford, VA 24142

[2]Virginia-Maryland College of Veterinary Medicine, Virginia Tech, Blacksburg, VA 24061

**email*: msarker@radford.edu

ABSTRACT: Pooled testing can enhance the efficiency of diagnosing individuals with diseases of low prevalence. Often, pooling is implemented using standard groupings (2, 5, 10, etc.). On the other hand, optimization theory can provide specific guidelines in finding the ideal pool size and pooling strategy. This article focuses on optimizing the precision of disease prevalence estimators calculated from multiplex pooled testing data. In the context of a surveillance application of animal diseases, we study the estimation efficiency (i.e., precision) and cost-efficiency of the estimators with adjustments for the number of expended tests. This enables us to determine the pooling strategies that offer the highest benefits when jointly estimating the prevalence of multiple diseases, such as theileriosis and anaplasmosis. The outcomes of our work can be used in designing pooled testing protocols, not only in simple pooling scenarios but also in more complex scenarios where individual retesting is performed in order to identify positive cases. A software application using the `shiny` package in `R` is provided with this article to facilitate implementation of our methods.

KEYWORDS: Animal testing; Experimental design; Group testing; Screening; Surveillance.

1 Introduction

When disease screening is to be performed for a large population, testing individuals one-at-a-time can be difficult or infeasible due to the restrictions in testing time and cost. To address such challenges, Dorfman (1943) introduced a two-stage hierarchical testing protocol when screening American recruiting soldiers during the Second World War. This protocol is commonly called pooled testing or group testing. Dorfman's approach, which involves testing pooled samples such as blood or swabs in stage 1 and resolving positive pools in stage 2, can be highly efficient when disease prevalence is low. Over the decades, pooled testing has been used in screening human populations for HIV and hepatitis B/C (Pilcher et al. 2005), chlamydia and gonorrhea (Lewis, Lockary, and Kobic 2012), and other infectious diseases. The use of pooled testing has been widespread in the recent COVID-19 pandemic (Daniel et al. 2021; Mutesa et al. 2021). In tracking animal diseases, pooled testing is especially appealing. Area- and country-wide surveillance are commonly used to prevent the spread of transboundary diseases. Animal producers often need disease information on individual animals or herds but cost limits the number of tests that are conducted.

Common goals of pooled testing include screening individuals for disease (known as case identification) or estimating the prevalence of disease at the subgroup or population level. Whether the goal is case identification or estimation, pooled testing can provide many benefits, such as savings in testing cost/time and increasing testing capacity. The focus of our work is estimation; i.e., to develop statistical methods and software so disease prevalence estimation can be performed more easily and efficiently. An interesting feature of pooled testing is that pooled responses alone can provide sufficient information for prevalence estimation (i.e., individual retest information is not required). Thus, when the testing budget is of main concern, one can find estimates using only pooled responses by expending only a small fraction of the testing budget required in one-at-a-time testing. It is, however, worth noting that the estimates become more precise when the additional individual retest responses are incorporated (Zhang,

Bilder, and Tebbs 2013a; Zhang et al. 2020a). In the pooled testing literature, much of the early work involves testing only initial pools (Vansteelandt, Goetghebeur, and Verstraeten 2000; Liu et al. 2012), whereas reccent work focuses more on pooled testing with pooled or individual retesting (Xie 2001; Wang et al. 2014; Wang, McMahan, and Gallagher 2015; McMahan et al. 2017; Liu et al. 2021; Warasi 2021).

While estimation based on pooled testing has been extensively studied with single infections, much less work on estimation is found with multiple infections. The first recognizable work with multiple infections is Hughes-Oliver and Rosenberger (2000), where a statistical model was developed to jointly estimate the prevalence of two or more infections. This work proceeded with the restrictive assumption that the assay used for diagnosis is perfect. Also, the work used only pooled responses. Tebbs, McMahan, and Bilder (2013) overcame these limitations using the expectation-maximization algorithm, where the model can use error-prone test responses from a Dorfman-type two-stage hierarchical protocol. Li, Liu, and Xiong (2017) also allowed for imperfect diagnosis but studied only some aspects of optimal estimation strategies with pooled responses. Zhang, Bilder, and Tebbs (2013b) and Lin et al. (2019) developed regression methods with multiple infections where the goals were only parameter estimation (i.e., optimization was not their goal). There is no work in the pooled testing literature that provides optimal strategies for estimation from multistage pooling. Therefore, when the prevalence of multiple infections is to be estimated from pooling data, there are limited or no statistical methods that can guide one to use the optimal pooling design.

Our work aims at filling this important research gap. We study optimality of pooled testing estimation with only pooled responses as well as both pooled and individual retest responses. As in Li et al. (2017) and Tebbs et al. (2013), we use the maximum likelihood framework, where our goal is to maximize the precision of the maximum likelihood estimates. Because testing cost is an important consideration, we account for the number of tests expended so the estimates can be most cost effective. Another goal of our work is to develop a user-

friendly software application. We do so because the statistical models for pooled testing are complicated, especially when individual retesting data from the two-stage protocol is involved. Our software application can be used to easily implement the optimization methods presented in this article. These methods are illustrated using data on infection with *Theileria orientalis* and *Anaplasma marginale* collected from a large cattle population.

The subsequent sections are organized as follows. In Section 2, we describe a motivating example of animal disease surveillance. In Section 3, we describe the pooling protocol and model framework considered in this article. In Section 4, we study the optimality of estimation based on pooling data. A brief discussion is provided in Section 5.

2 Animal testing data

Theileriosis and anaplasmosis are tick-borne bovine infections caused by the parasite and bacterial agents *Theileria orientalis* and *Anaplasma marginale*. Theileriosis and anaplasmosis have similar clinical presentations in cattle and co-occur in the eastern United States (Oakes et al. 2022). Simultaneously testing to differentiate these diseases is important because the bacteria causing anaplasmosis responds to treatment with antimicrobials, while there is no Food and Drug Administration (FDA) approved treatment for theileriosis (USDA 2021). As part of a surveillance study of these infections, blood samples were collected through collaboration with the Virginia Department of Agriculture and Consumer Sciences from 1736 adult market cattle in 2018-2020 from different counties in Virginia and tested at the Virginia-Maryland College of Veterinary Medicine. Cattle samples were individually tested using a duplex real-time quantitative PCR assay (qPCR) for both infections simultaneously. Pooled testing would potentially allow wider future surveillance that could be conducted more efficiently. However, before implementing a new pooling design, it is important to determine the optimal design, and a substantial gain is possible from pooled testing.

[Table 1 near here]

Table 1 summarizes the test outcomes, which we view as historical data. The sensitivity and specificity of the duplex assay for *A. marginale* are 0.97 and 1.0, respectively, optimized at 37 amplification cycles. For *T. orientalis*, both sensitivity and specificity are 1.0, measured at 45 cycles (Oakes et al. 2022). For qPCR, each cycle doubles the DNA concentration in the sample, with a threshold of cycles selected to optimize sensitivity and specificity for pathogen detection (Kralik and Ricchi 2017). Based on the individual test outcomes and assay accuracy information, we first estimate the coinfection probabilities p_{00}, p_{10}, p_{01}, and p_{01}, where

p_{00} = proportion of cattle negative for both *T. orientalis* and *A. marginale*

p_{10} = proportion of cattle positive for *T. orientalis* but negative for *A. marginale*

p_{01} = proportion of cattle negative for *T. orientalis* but positive for *A. marginale*

p_{11} = proportion of cattle positive for both *T. orientalis* and *A. marginale*

and $p_{00} + p_{10} + p_{01} + p_{11} = 1$. The coinfection prevalence estimates are depicted in Table 1. These estimates were calculated accounting for testing error, using methods as in Warasi et al. (2016, Web Appendix D).

3 Preliminaries

Consider testing a random sample of N cattle with the qPCR assay for *T. orientalis* and *A. marginale* simultaneously. Let $\mathbf{p} = (p_{00}, p_{10}, p_{01})'$ denote the vector of coinfection probabilities to be estimated from pooled testing data, where $p_{11} = 1 - p_{00} - p_{10} - p_{01}$. We consider use of either only pooled responses or both pooled and individual retest responses observed from a hierarchical testing protocol. In this section, we describe the pooling protocol and how the maximum likelihood estimate of $\mathbf{p}$ can be calculated from pooling data.

The first stage in the pooling protocol involves drawing individual blood samples from the cattle and forming pools by mixing k individual blood samples together prior to testing (i.e., k is the pool size). The maximum k for optimization is limited by the potential loss of test sensitivity due to dilution. Maximum k can be determined in the laboratory by sequentially testing a set of representative positive samples pooled with an increasing number of negative samples to find the largest pool size which still maintains high sensitivity (Bateman et al. 2021). Maximum k can alternatively be based on estimating its effects on the weakest positive samples, since these are most susceptible to becoming false negatives through dilution (Mutesa et al. 2021). For animal diagnostics, funding for conducting serial pooling and testing of representative cattle samples may not be available. Thus, we have initially estimated the maximum k based on the weakest positives. For *A. marginale* positives, the assay should maintain a sensitivity of 97% for detecting one positive sample in a pool of size 16 or smaller, if the threshold for positives is changed from 37 to 40 cycles (Oakes et al. 2022). For *T. orientalis* positives, the assay should maintain a sensitivity of 100% with a pool size of up to 16 using a threshold of 43 cycles. We selected 10 as the maximum pool size for optimization throughout to stay within the limits of dilution effects with these qPCR thresholds.

Hierarchical pooling yields $m = N/k$ pooled blood samples, which are tested by the duplex assay in stage 1, followed by individual retesting in stage 2. The steps are described below.

1. At stage 1, m pools are tested. If a pool tests negatively for both *T. orientalis* and *A. marginale*, the pool members are diagnosed as negative; i.e., no further test is performed.

2. At stage 2, the members of the pools that test positively for at least one infection are retested one by one for case identification. All samples are retested with the duplex assay rather than separately for *T. orientalis* and *A. marginale*.

This type of hierarchical procedure has been commonly used in screening human populations. Tebbs et al. (2013) examined the screening accuracy, efficiency, and other characteristics of

hierarchical testing with a duplex assay for chlamydia and gonorrhea. They developed an estimation technique for this protocol but did not focus on optimizing the estimates. We use the same estimation framework but, unlike these authors, we study the optimization aspects of estimation (e.g., minimzing the mean squared error while reducing the testing cost). We use an animal disease diagnostic scenario because optimized pooling approaches have rarely been applied in veterinary diagnostics. Differences in diagnostic scenarios include that surveillance is often applied at the herd level and cost is often a serious constraint to disease detection as testing is not covered by insurance. Additionally, our team includes a veterinary diagnostician and veterinary epidemiologist working to enhance the strength of testing approaches in veterinary medicine. This includes optimizing pooling for these two serious diseases to improve animal health using the described methodologies.

[Figure 1 near here]

Let $\mathbf{Z} = (Z_1, Z_2)'$ denote a pooled response from stage 1, where $Z_1 = 1$ ($Z_2 = 1$) if a pool tests positively for *T. orientalis* (*A. marginale*) and $Z_1 = 0$ ($Z_2 = 0$) if a pool tests negatively for *T. orientalis* (*A. marginale*). Similarly, denote an individual response from stage 2 by $\mathbf{Y} = (Y_1, Y_2)'$. Figure 1 shows the four possible scenarios that one can see when testing a pool. Figure 1 also shows how the test outcomes (both pooled and individual) are coded by 0 and 1. We assume that the assay sensitivity and specificity are unaffected by the pool size. This is true over the range of k values explored in this application. We also assume that the test responses observed in stages 1-2 are mutually independent conditional on their true statuses. These assumptions are commonly adopted in pooled testing (Kim et al. 2007).

We consider two approaches for estimation. The first uses only pooled responses, $\mathbf{Z}$'s, from stage 1. Using the method in Li et al. (2017), the maximum likelihood estimates and covariance matrix are calculated. Our second approach takes advantage of individual retest responses in addition to the pooled responses. In this case, the maximum likelihood estimates

and covariance matrix are calculated based on the work in Tebbs et al. (2013). Estimation using additional individual data is more precise but also more challenging because the pooled and individual responses are potentially correlated (i.e., an individual may be tested in both stages). Tebbs et al. (2013) addressed this issue using a "missing data" technique, where the maximum likelihood estimates are calculated by the expectation-maximization (EM) algorithm and the observed-data Fisher information is calculated by Louis's (1982) method.

Denote by $\widehat{\mathbf{p}} = (\widehat{p}_{00}, \widehat{p}_{10}, \widehat{p}_{01})'$ the maximum likelihood estimate (MLE) of $\mathbf{p} = (p_{00}, p_{10}, p_{01})'$, which is calculated using only pooled responses or both pooled and individual responses. Then one can find the MLE of p_{11} as $\widehat{p}_{11} = 1-\widehat{p}_{00}-\widehat{p}_{10}-\widehat{p}_{01}$ and the MLE of the marginal prevalence of *T. orientalis* (μ_1), marginal prevalence of *A. marginale* (μ_2), and correlation (ρ) as

$$\widehat{\mu}_1 = \widehat{p}_{10} + \widehat{p}_{11},\ \widehat{\mu}_2 = \widehat{p}_{01} + \widehat{p}_{11},\ \widehat{\rho} = \frac{\widehat{p}_{11} - \widehat{\mu}_1\widehat{\mu}_2}{\sqrt{\widehat{\mu}_1(1-\widehat{\mu}_1)\widehat{\mu}_2(1-\widehat{\mu}_2)}}.$$

Let $\mathcal{I}(\mathbf{p})$ denote the expected Fisher information matrix of $\mathbf{p}$. The covariance matrix $\mathcal{I}(\mathbf{p})^{-1}$, an inverse of the Fisher information $\mathcal{I}(\mathbf{p})$, is estimated at the MLE $\widehat{\mathbf{p}}$. The variance of $\widehat{p}_{11}$, $\widehat{\mu}_1$, $\widehat{\mu}_2$, and $\widehat{\rho}$ can be estimated by using the delta method. Then, the Wald confidence interval for each parameter can be easily calculated in the usual manner.

4 Efficiency measures

Pool size plays an important role in gaining savings in testing cost and in realizing the precision of estimates from pooled testing. With these aspects in mind, we consider optimizing the following measures of efficiency: (a) $E[T]$, the expected number of tests expended, (b) $E[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$, the mean squared error associated with the MLE $\widehat{\mathbf{p}}$, and (c) $E[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$, the cost per unit information, where T is the number of tests expended. Note that T is the number of pools m when only pooled testing is used. For hierarchical testing, T is the total number of

tests used in stages 1-2 (i.e., m plus the number of individual retests). While $E[T]$ concerns testing cost, $E[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ captures the cost in estimation. Then $E[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ can be viewed as a compromise between the two.

We can examine the measures of efficiency for pooled testing relatively to individual testing. Therefore, we use relative test efficiency (RTE), relative estimation efficiency (REE), and relative cost efficiency (RCE), which are defined as

$$\begin{aligned}
\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) &= \frac{E_G[T]}{E_I[T]} \\
\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) &= \frac{E_G[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]}{E_I[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]} \\
\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) &= \frac{E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]}{E_I[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]},
\end{aligned}$$

where $\widehat{\mathbf{p}}_G$ and $\widehat{\mathbf{p}}_I$ are the MLEs from group or pooled testing (either hierarchical or pools only) and individual testing (i.e., one-at-a-time testing), respectively. For all three criteria, smaller is better (i.e., more efficient/precise). Thus, our goal is to identify the pool sizes that minimize these measures, which we do under three scenarios. In the first, the number of cattle that we can collect is limited or fixed. In the second, the number of total tests that we can afford to run is fixed. In the third, our target is a desired level of precision of our estimates.

4.1 The sample size is fixed

When the number of cattle N is fixed (i.e., a constant) and only pooled testing is used, $T = N/k$ is a constant for a given pool size k. Thus, the expected number of tests $E_G[T] = N/k$ so that $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) = 1/k$. Note that the expected number of tests expended with individual testing is always N, so $E_I[T] = N$. Because $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) = 1/k$, one can easily see that testing efficiency increases with pool size k.

However, dividing the N samples into larger pools might adversely affect estimation efficiency

because estimates would be based on fewer test responses. To understand the loss or gain in estimation efficiency, we examine the mean squared error $E[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$. In doing so, we first recognize that the MLE $\widehat{\mathbf{p}} = (\widehat{p}_{00}, \widehat{p}_{10}, \widehat{p}_{01})'$ has a large-sample multivariate normal distribution with mean $\mathbf{p}$ and covariance matrix $\mathcal{I}(\mathbf{p})^{-1}$ (Boos and Stefanski 2013, Section 2.5). Then the mean squared error for pooled testing can be expressed as

$$E_G[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})] = \text{var}(\widehat{p}_{00}) + \text{var}(\widehat{p}_{10}) + \text{var}(\widehat{p}_{01}),$$

where $\text{var}(\widehat{p}_{00})$, $\text{var}(\widehat{p}_{10})$, and $\text{var}(\widehat{p}_{01})$ are the diagonal elements of the covariance matrix $\mathcal{I}(\mathbf{p})^{-1}$. For individual testing, the mean squared error $E_I[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ is calculated in the same manner using pool size $k = 1$. Then we find $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, which we minimize as a function of the pool size k for optimal estimation efficiency.

Because T is a constant (for pooled only testing), we find an explicit expression of

$$E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})] = T[\text{var}(\widehat{p}_{00}) + \text{var}(\widehat{p}_{10}) + \text{var}(\widehat{p}_{01})],$$

and $E_I[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ can be calculated analogously. Then, $\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ can be minimized with respect to k for optimal cost efficiency. Note that, with pooled testing only, we calculate $E[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ and $E[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ analytically; see Web Appendix A.

For hierarchical testing, T, the number of tests used, is not fixed because individual retesting in stage 2 depends on whether a pool tests negatively or positively for the infections. Tebbs et al. (2013) provided a closed-form expression of $E_G[T]$, which we use for calculating $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, where $E_I[T] = N$ as before.

For hierarchical testing, the large-sample mean squared error is again

$$E_G[(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})] = \text{var}(\widehat{p}_{00}) + \text{var}(\widehat{p}_{10}) + \text{var}(\widehat{p}_{01}).$$

Unfortunately, there is no extant work to calculate the expected information matrix $\mathcal{I}(\mathbf{p})$, and thus evaluation of the variance components $\text{var}(\widehat{p}_{00})$, $\text{var}(\widehat{p}_{10})$, and $\text{var}(\widehat{p}_{01})$ is not possible based on the methods available in the literature. Because T is random, finding an explicit expression of $E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$ is even more difficult. To overcome these challenges, we have developed a computation algorithm that can be used to approximate $\mathcal{I}(\mathbf{p})$, $E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$, or any other quantities that are hard to calculate analytically. Then, the variance components $\text{var}(\widehat{p}_{00})$, $\text{var}(\widehat{p}_{10})$, and $\text{var}(\widehat{p}_{01})$ can be found as before. The computation algorithm is described below, where hierarchical testing data that consists of both pooled and individual responses is denoted by $\mathcal{D}$.

1. Specify a value of $\mathbf{p} = (p_{00}, p_{10}, p_{01})'$ from historical or pilot study data.

2. Simulate $\mathcal{D}$ according to the hierarchical testing protocol at the parameter value $\mathbf{p}$ given in step 1.

3. Do the following.

   (a) Calculate the observed information matrix $I(\mathbf{p})$ at the given value $\mathbf{p}$.

   (b) Find the MLE $\widehat{\mathbf{p}}$ and the number of tests T from the simulated data $\mathcal{D}$, and then calculate $T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})$.

4. Repeat steps 2-3 G times, where G is large, and do the following.

   (a) Take an average of the values of $I(\mathbf{p})$ found in step 3(a). This average is an estimate of the expected information matrix $\mathcal{I}(\mathbf{p})$.

   (b) Take an average of the values of $T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})$ found in step 3(b). This average is an estimate of $E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$.

As mentioned in Section 3, we calculate the MLE $\widehat{\mathbf{p}}$ and observed information matrix $I(\mathbf{p})$ in steps 3(a)-3(b) based on the work in Tebbs et al. (2013). The MLE is calculated using the

EM algorithm where a Gibbs sampler is implemented to approximate the E-step. Also, the same Gibbs sampler is used to approximate the expectations involved in $I(\mathbf{p})$. In both steps 3(a)-3(b), we use 3000 Gibbs iterates after discarding the initial 1000 iterates as a burn-in period; for more information about the MLE and information matrix, refer to Web Appendix A. Step 4 is justified by the law of large numbers; i.e., when the number of repetitions G is large, the averages in step 4 are reasonable approximations. In Appendix B of the web-based supplementary material, we discuss how large G should be to achieve a reasonable approximation of $\mathcal{I}(\mathbf{p})$ and $E_G[T(\widehat{\mathbf{p}}-\mathbf{p})'(\widehat{\mathbf{p}}-\mathbf{p})]$.

[Figure 2 near here]

To identify the optimal pooling configurations for our cattle data example, we compute $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, and $\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ with sample size $N = 500, 1000, 2000$ and parameter $\mathbf{p} = (0.834, 0.075, 0.078)'$ taken from Table 1. In Figure 2, the relative efficiency results are plotted against pool sizes $k = 2, 3, ..., 10$. For pooled testing only, none of the relative measures depends on N because N is canceled out from the numerator and denominator; i.e, all three relative measures are independent of N but we show them for each choice of N in Figure 2 for comparison. For hierarchical testing, we calculate $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, which is also independent of N. For hierarchical testing, we calculate $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ and $\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ using our computation algorithm with $G = 5000$ repetitions to approximate the expected values in step 4. We made this choice of G based on the convergence tests in Web Appendix B, where one can see that the sample averages converge at $G = 5000$ or faster. We have found that the approximations (not shown) are reasonable even with much smaller G, such as 1000.

For each of the relative measures, the optimal pool size occurs where the measures have the smallest values. When using only pools, the smallest pool size, 2, provides the best precision and the largest pool size, 10, requires the fewest number of tests. When these are simultaneously considered, the optimal pool size is 8. For hierarchical testing, pool size 3 offers

the fewest number of tests. One can see in Figure 2 that the relative estimation efficiency curves are flat; i.e., pool size does not have a noticeable impact on precision. This is likely because positive pools are retested so each positive case is identified. Because pool size does not affect estimation efficiency, one has the freedom to choose the pool size that yields the highest cost savings without compromising precision in estimation. This is reflected in the cost-efficiency graphs in Figure 2, which shows that optimal cost-efficiency is achieved at pool size 3. Even though different choices of N are used, the overall patterns of $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ and $\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ are not affected, which is expected when the large-sample assumption is reasonable.

Pooled testing can be compared with individual testing from the results in Figure 2. Interestingly, hierarchical testing provides somewhat better estimation precision (the flat graphs are below 1) even using fewer tests, when compared to individual testing. Pooled testing only provides coarse estimates, which is not unexpected because only pools are used and the assay sensitivities/specificities are very high. However, the precision loss can be mitigated by using the optimally chosen pool size (which is 2), which still provides a 50% decrease in the number of tests with only about a 10% loss in precision. Also, when cost-saving is of interest, pooled testing only can be the most attractive option, as is evident in the cost-efficiency graphs.

4.2 The number of tests is fixed

A second scenario is when the number of tests to be expended, T, is fixed; for example, when the testing budget is limited but the goal is to test a larger number of cattle than could be tested using the same number of tests only individually. We can do this only for testing of pools because T is not deterministic for hierarchical testing; i.e., individual retesting in stage 2 is uncertain as it depends on the pooled responses in stage 1.

Because T is fixed, the same number of individual and pooled tests must be used in the

relative measures. In this case, $\text{RTE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$, by definition, is 1.0, and T does not play any role in $\text{RCE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ as well. Therefore, we focus on only the relative estimation efficiency $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$. For each pool size $k = 2, 3, ..., 10$, we calculate $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$ at the true value $\mathbf{p} = (0.834, 0.075, 0.078)'$. The results are shown in Table 2, where pool sizes 7 and 8 provide the smallest $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I)$. Thus, either 7 or 8 can be used as the optimal pool size.

[Table 2 near here]

In this scenario, the number of cattle that can be tested using the same number of tests that individual testing requires is $k * N$. This can yield a substantial gain in estimator precision. For example, when $N = 100$, individual testing expends $T = 100$ tests. Then even with $k = 2$, 200 cattle could be tested with the same T and precision would be almost twice as high, $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) = 0.564$. With 700 or 800 cattle (i.e., the optimal configuration), this gain is much higher, $\text{REE}(\widehat{\mathbf{p}}_G, \widehat{\mathbf{p}}_I) = 0.303$. However, increasing the number of cattle is not always better. This can be seen in Table 2, where precision starts to become worse at pool size 9.

4.3 The minimal level of precision is fixed

Our final consideration of optimality is when a certain level of precision needs to be maintained in estimation. This is especially useful when estimation accuracy is of primary interest in disease surveillance or population-based decision making. By doing this, we can find the pool size that provides the best precision with the lowest testing cost. As in Section 4.2, we use only pooled responses because individual retesting in stage 2 is uncertain.

Let E_0 denote the upper bound of the mean squared error $E[(\widehat{\mathbf{p}} - \mathbf{p})'(\widehat{\mathbf{p}} - \mathbf{p})]$; i.e., E_0 can be viewed as the maximum amount of error in estimation. Let T^* denote the minimum number of tests needed to reach the desired precision. Then we find T^* when the maximum value of

$E[(\widehat{\mathbf{p}} - \mathbf{p})'(\widehat{\mathbf{p}} - \mathbf{p})]$ is smaller than E_0. That is, T^* is the T that satisfies

$$\max E[(\widehat{\mathbf{p}} - \mathbf{p})'(\widehat{\mathbf{p}} - \mathbf{p})] < E_0.$$

We perform this calculation using our software where, under each pool size k, we keep increasing T. When $E[(\widehat{\mathbf{p}} - \mathbf{p})'(\widehat{\mathbf{p}} - \mathbf{p})]$ becomes smaller than E_0, the program stops and records T as T^*.

[Table 3 near here]

Table 3 displays the number of tests needed to achieve a desired level of precision ($E_0 = 0.001, 0.003, ..., 0.015$) using pools of size 1 (individual testing) to size 10. The optimal pool size for each is the one that corresponds to the smallest T^*. As one might expect, more tests are required for a smaller E_0 (i.e., higher level of precision). The number of required tests for an E_0 decreases quickly as pool size increases, but plateaus at a pool size of 6 or 7. With $E_0 = 0.001$, for instance, pooled testing with pool size 7 takes only 87 tests to achieve the precision whereas individual testing requires 285 tests. Although the total number of tests is minimized, the number of cattle being tested ($k * T$) with pooling is larger than for individual testing, allowing 609 cattle to be tested with 87 tests in this example.

4.4 Software application for optimization

Determining the optimal pooling strategy for field use is directly beneficial to diagnosticians. Thus, we provide an interactive software application using the `shiny` package (Chang et al. 2021) in `R` (R Core Team 2021). From pooled tests alone or both pooled and individual retests, the software can provide the efficiency results presented in this article. Note that the computation algorithm in Section 4.1 is computationally intensive because the EM algorithm and observed information matrix that it uses in step 3 involve a Gibbs sampler. Therefore,

we strategically use compiled `Fortran` and `C` code to enhance computing efficiency of the software. With $G = 1$ repetition and sample size $N = 1000$, the computation algorithm is completed in 0.9 seconds for a pool size in an Intel 3.6GHz 32GB RAM machine. With $G = 1000$ repetitions, the computing time is about 15 minutes.

The software application will be disseminated in two ways. We will make it available at the free distribution site `https://www.shinyapps.io`. This distribution will be most portable and accessible; i.e., this can be accessed from any place without installing `R` or any other package. We will also provide a script file for the `R` users. Additional information about the software application is provided in Web Appendix C.

5 Discussion

Pooling samples for testing is a useful technique when disease prevalence is low and testing retains high sensitivity and specificity. Determining the optimal pooling strategy is complex because it depends on many factors, such as the goal of testing, disease epidemiology, precision of the outcome, urgency of results, number of available samples, and cost (Laurin et al. 2019). However, one of the challenges to widespread adoption of pooled testing is availability of the methods that determine the ideal pool size. The methods reported herein provide optimization techniques under several constraints with illustration through a current field application. To make implementation of our work effortless, we provide a software application.

Three constraints that could guide the design of a pooled sampling program were examined. The first was to design a strategy for a fixed number of individual samples. In this case, hierarchical testing gives the same precision in estimation with any pool size. Thus, we can choose the pool size that offers the highest cost savings. Methods were also developed to select the optimal pool size to limit the number of tests used. Here, we set our total number of available tests to be the number required to test every sample individually. We then optimized

the estimation precision to determine the largest number of animals that could be pool-tested with this same number of tests. A valuable generalization is that the optimal pool size is less than merely pooling as many animals as possible. The third option was to determine the fewest number of tests that could be used to estimate the prevalence in a population with a predetermined level of precision. When estimating the prevalence based on results from only the pools, pooled testing requires less than a third of the tests needed by individual testing to achieve the same precision.

Our work makes several assumptions that can be relaxed for more flexibility and generality. We use a duplex assay for testing the cattle where $k = 10$ is used as the maximum pool size to safeguard against potential loss of sensitivity due to dilution. We also assume that sensitivity and specificity of the assay are constant over the limited range of pool sizes. While these assumptions are reasonable in the surveillance application of the animal diseases, more general methods such as those that allow for differential misclassification errors (Hung and Swallow 1999; Zhang et al. 2020b) or pooled dilution effects (McMahan, Tebbs, and Bilder 2013; Warasi et al. 2017; Mokalled et al. 2020) could be important future work. These would be especially useful when the disease prevalence is very low or the pool size to be used is large. We take a likelihood-based approach assuming that a reasonable prior value of the parameter $\mathbf{p} = (p_{00}, p_{10}, p_{01})'$ is available from historical or pilot study data. Understandably, this approach may provide incorrect optimality results when $\mathbf{p}$ is misspecified; see Web Appendix D, where we explored the effects of misspecification of $\mathbf{p}$. Thus, a more flexible approach such as adaptive algorithm (Hughes-Oliver and Swallow 1994; Hughes-Oliver and Rosenberger 2000) that relies less on the prior value of $\mathbf{p}$ would be a valuable future extention of our work.

Although surveillance of *T. orientalis* and *A. marginale* among cattle in Virginia is used as an example, this work can also be useful in applications in human medicine, environmental testing, and other fields. The software that we provide allows animal and public health professionals to easily optimize pooling scenarios for their needs. For example, when an

optimal pooling design is to be identified for estimating the prevalence of any co-occurring pathogens, one only needs to update the value of $\mathbf{p}$ and assay sensitivity/specificity. The computation algorithm introduced in this article can be extended to accommodate data arising from more complex pooling protocols, such as hierarchical testing with three or more stages (Hou et al. 2017) and array testing (Hou et al. 2020; Bilder, Tebbs, and McMahan 2021).

Acknowledgments

We are grateful to the Editor, Associate Editor, and two reviewers for many helpful suggestions. We also thank our colleagues at the Virginia Department of Agriculture and Consumer Sciences and Virginia-Maryland College of Veterinary Medicine for their collaboration in sampling, testing, and compiling data on *T. orientalis* and *A. marginale* for the cattle.

Disclosure statement

There are no conflicts.

Data availability statement

Data used in this study were only counts of results from the duplex *T. orientalis* and *A. marginale* qPCR assay on cattle and are completely presented in Table 1.

References

Bateman A, Mueller S, Guenther K, Shult P (2021) Assessing the dilution effect of specimen pooling on the sensitivity of SARS-CoV-2 PCR tests. *Journal of Medical Virology* **93**:1568-72. https://doi.org/10.1002/jmv.26519

Bilder C, Tebbs J, McMahan C (2021) Informative array testing with multiplex assays. *Statistics in Medicine* **40**:3021-34. https://doi.org/10.1002/sim.8954

Boos D, Stefanski L (2013) Essential statistical inference, theory and methods. Springer, New York.

Chang W, Cheng L, Allaire J et al (2021) shiny: Web application framework for R. R package version 1.7.1. https://cran.r-project.org/web/packages/shiny/index.html

Daniel E, Esakialraj B, Muthuramalingam A et al (2021) Pooled testing strategies for SARS-CoV-2 diagnosis: A comprehensive review. *Diagnostic Microbiology and Infectious Disease* **101**:115432. https://doi.org/10.1016/j.diagmicrobio.2021.115432

Dorfman R (1943) The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**:436-40. https://doi.org/10.1214/aoms/1177731363

Hou P, Tebbs J, Bilder C, McMahan C (2017) Hierarchical group testing for multiple infections. *Biometrics* **73**:656-65. https://doi.org/10.1111/biom.12589

Hou P, Tebbs J, Wang D, Bilder C, McMahan C (2020) Array testing for multiplex assays. *Biostatistics* **21**:417-31. https://doi.org/10.1093/biostatistics/kxy058

Hughes-Oliver J, Rosenberger W (2000) Efficient estimation of the prevalence of multiple rare traits. *Biometrika* **87**:315-27. https://doi.org/10.1093/biomet/87.2.315

Hughes-Oliver J, Swallow W (1994) A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* **89**:982-93.

https://doi.org/10.2307/2290924

Hung M, Swallow W (1999) Robustness of group testing in the estimation of proportions. *Biometrics* **55**:231-37. https://doi.org/10.1111/j.0006-341x.1999.00231.x

Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C (2007) Comparison of group testing algorithms for case identification in the presence of testing error. *Biometrics* **63**:1152-63. https://doi.org/10.1111/j.1541-0420.2007.00817.x

Kralik P, Ricchi M (2017) A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Frontiers in Microbiology* **8**:108. https://doi.org/10.3389/fmicb.2017.00108

Laurin E, Thakur K, Mohr P et al (2019) To pool or not to pool? Guidelines for pooling samples for use in surveillance testing of infectious diseases in aquatic animals. *Journal of Fish Diseases* **42**:1471-91. https://doi.org/10.1111/jfd.13083

Lewis J, Lockary V, Kobic S (2012) Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases* **39**:46-8. https://doi.org/10.1097/OLQ.0b013e318231cd4a

Li Q, Liu A, Xiong W (2017) D-optimality of group testing for joint estimation of correlated rare diseases with misclassification. *Statistica Sinica* **27**:823-38. https://doi.org/10.5705/ss.202015.0178

Lin J, Wang D, Zheng Q (2019). Regression analysis and variable selection for two-stage multiple-infection group testing data. *Statistics in Medicine* **38**:4519-33. https://doi.org/10.1002/sim.8311

Liu A, Liu C, Zhang Z, Albert P (2012) Optimality of group testing in the presence of misclassification. *Biometrika* **99**:245-51. https://doi.org/10.1093/biomet/asr064

Liu Y, McMahan C, Tebbs J, Gallagher C, Bilder C (2021) Generalized additive regression for group testing data. *Biostatistics* **22**:873-89. https://doi.org/10.1093/biostatistics/kxaa003

Louis T (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodology)* **44**:226-33. https://doi.org/10.1111/j.2517-6161.1982.tb01203.x

McMahan C, Tebbs J, Bilder C (2013) Regression models for group testing data with pool dilution effects. *Biostatistics* **14**:284-98. https://doi.org/10.1093/biostatistics/kxs045

McMahan C, Tebbs J, Hanson T, Bilder C (2017) Bayesian regression for group testing data. *Biometrics* **73**:1443-52. https://doi.org/10.1111/biom.12704

Mokalled S, McMahan C, Tebbs J, Brown D, Bilder C (2020) Incorporating the dilution effect in group testing regression. *Statistics in Medicine* **40**:2540-55. https://doi.org/10.1002/sim.8916

Mutesa L, Ndishimye P, Butera Y et al (2021) A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* **589**:276-80. https://doi.org/10.1038/s41586-020-2885-5

Oakes V, Todd S, Carbonello A, Michalak P, Lahmers K (2022) Coinfection of cattle in Virginia with *Theileria orientalis* Ikeda genotype and *Anaplasma marginale*. *Journal of Veterinary Diagnostic Investigation* **34**:36-41. https://doi.org/10.1177/10406387211057627

Pilcher C, Fiscus S, Nguyen T et al (2005) Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine* **352**:1873-83. https://doi.org/10.1056/NEJMoa042291

R Core Team (2021) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Tebbs J, McMahan C, Bilder C (2013) Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics* **69**:1064-73. https://doi.org/10.1111/biom.12080

USDA [United States Department of Agriculture] (2021) Emerging risk notice: *Theileria orientalis* Ikeda. Washington, DC: USDA, Animal and Plant Health Inspection Service. https://www.aphis.usda.gov (visited on January 18, 2022)

Vansteelandt S, Goetghebeur E, Verstraeten T (2000) Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**:1126-33. https://doi.org/10.1111/j.0006-341X.2000.01126.x

Wang D, McMahan C, Gallagher M, Kulasekera B (2014) Semiparametric group testing regression models. *Biometrika* **101**:587-98. https://doi.org/10.1093/biomet/asu007

Wang D, McMahan C, Gallagher C (2015) A general parametric regression framework for group testing data with dilution effects. *Statistics in Medicine* **34**:3606-21. https://doi.org/10.1002/sim.6578

Warasi M (2021) groupTesting: an R package for group testing estimation. *Communications in Statistics - Simulation and Computation*. Published online: 09 Dec 2021. https://doi.org/10.1080/03610918.2021.2009867

Warasi M, Tebbs J, McMahan C, Bilder C (2016) Estimating the prevalence of multiple diseases from two-stage hierarchical pooling. *Statistics in Medicine* **35**:3851-64. https://doi.org/10.1002/sim.6964

Warasi M, Tebbs J, McMahan C, Bilder C (2017) Group testing regression models with dilution submodels. *Statistics in Medicine* **36**:4860-72. https://doi.org/10.1002/sim.7455

Xie M (2001) Regression analysis of group testing samples. *Statistics in Medicine* **20**:1957-69. https://doi.org/10.1002/sim.817

Zhang B, Bilder C, Tebbs J (2013a) Group testing regression model estimation when case identification is a goal. *Biometrical J.* **55**:173-89. https://doi.org/10.1002/bimj.201200168

Zhang B, Bilder C, Tebbs J (2013b) Regression analysis for multiple-disease group testing data. *Statistics in Medicine* **32**:4954-66. https://doi.org/10.1002/sim.5858

Zhang W, Liu A, Li Q, Albert P (2020a) Incorporating retesting outcomes for estimation of disease prevalence. *Statistics in Medicine* **39**:687-97. https://doi.org/10.1002/sim.8439

Zhang W, Liu A, Li Q, Albert P (2020b) Nonparametric estimation of distributions and diagnostic accuracy based on group-tested results with differential misclassification. *Biometrics* **76**:1147-56. https://doi.org/10.1111/biom.13236

Table 1: Historical data and estimates from a sample of 1736 cattle in Virginia.

T. orientalis	*A. marginale*	Test outcomes	Coinfection prevalence
$-$	$-$	1449	$p_{00} = 0.834$
$+$	$-$	133	$p_{10} = 0.075$
$-$	$+$	132	$p_{01} = 0.078$
$+$	$+$	22	$p_{11} = 0.013$

Table 2: Relative estimation efficiency (REE) with pooled testing only.

Pool size	2	3	4	5	6	7	8	9	10
REE	0.564	0.424	0.361	0.327	0.310	0.303	0.303	0.309	0.319

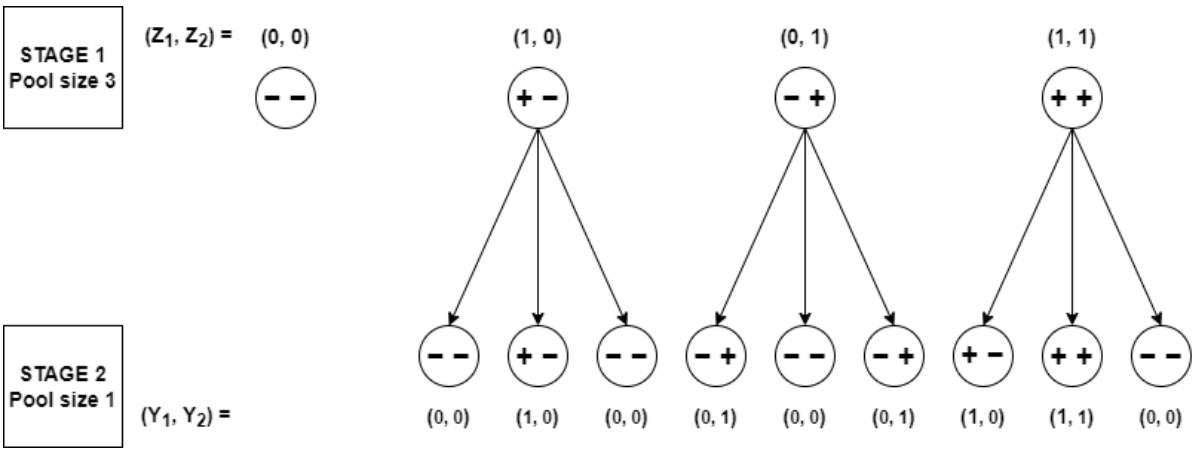

Figure 1: Two-stage hierarchical testing with pool size 3, showing an example of possible test outcomes. Pools that test positively in stage 1 for at least one infection are resolved in stage 2 by duplex testing of individual samples.

Table 3: The number of tests, T^*, required to estimate $\mathbf{p}$ with a maximum error of E_0. The minimum value of T^* over the pool sizes is marked by asterisk ($*$). When multiple pools are optimal, the smaller one is marked. Pool size 1 refers to individual testing.

Pool size	$E_0 = 0.001$	0.003	0.005	0.007	0.009	0.011	0.013	0.015
1	285	95	57	41	32	26	22	19
2	161	54	33	23	18	15	13	11
3	121	41	25	18	14	11	10	9
4	103	35	21	15	12	10	8	7
5	94	32	19	14	11	9	8	7
6	89	30	18^*	13^*	10^*	9	7^*	6^*
7	87^*	29^*	18	13	10	8^*	7	6
8	87	29	18	13	10	8	7	6
9	89	30	18	13	10	9	7	6
10	91	31	19	13	11	9	7	7

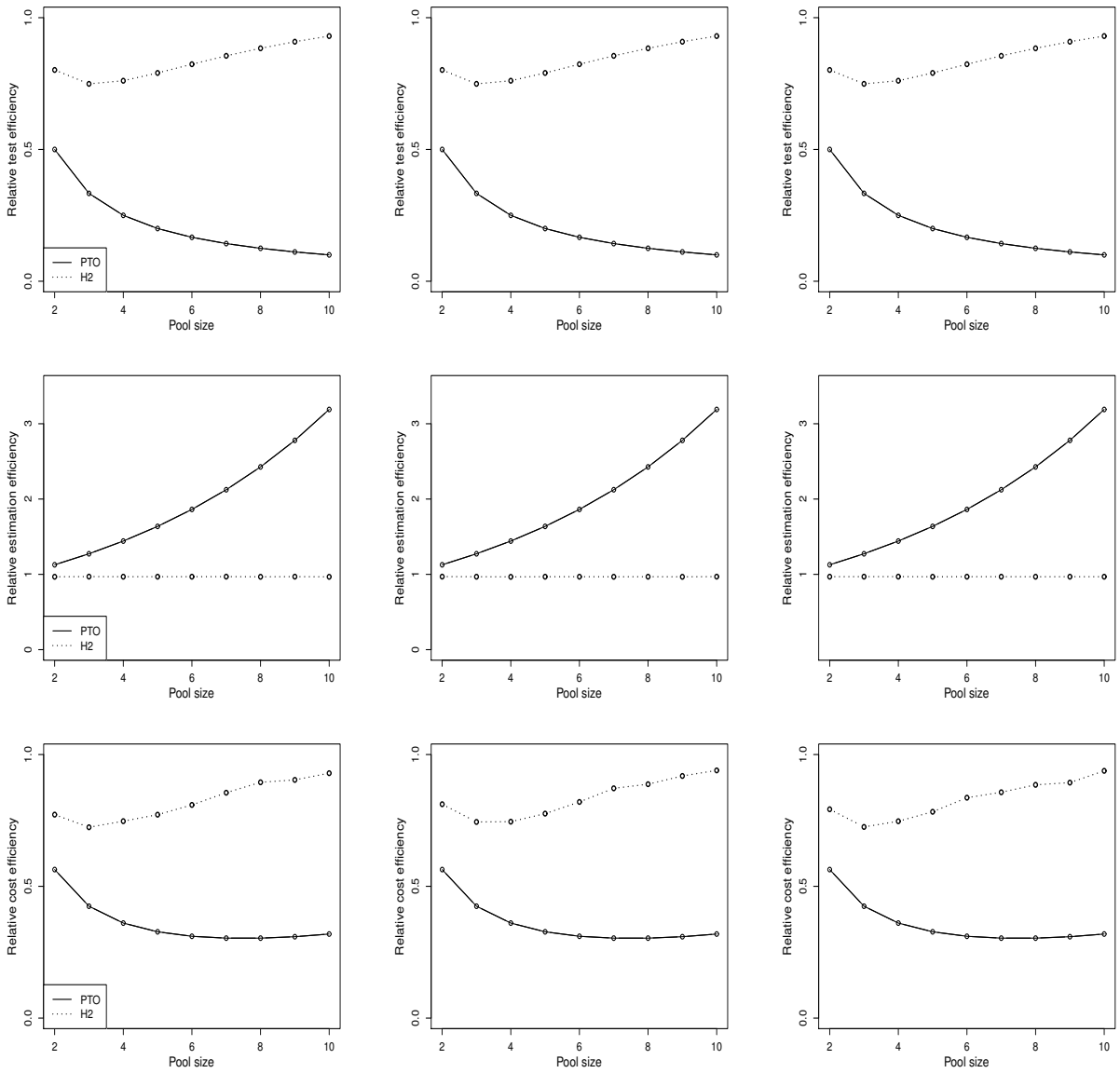

Figure 2: Relative test efficiency, relative estimation efficiency, and relative cost efficiency as a function of the pool size with sample sizes $N = 500$ (left), 1000 (middle), and 2000 (right). Results are shown for pooled testing only (PTO) and two-stage hierarchical testing (H2).