3-1-2023

Estimating the prevalence of two or more diseases using outcomes from multiplex group testing

Md S. Warasi

Joshua M. Tebbs

Christopher S. McMahan

Christopher R. Bilder

Received: 29 September 2022 | Revised: 26 February 2023 | Accepted: 1 March 2023

DOI: 10.1002/bimj.202200270

Biometrical Journal

RESEARCH ARTICLE

Estimating the prevalence of two or more diseases using outcomes from multiplex group testing

Md S. Warasi[1] | **Joshua M. Tebbs**[2] | **Christopher S. McMahan**[3] | **Christopher R. Bilder**[4]

[1]Department of Mathematics and Statistics, Radford University, Radford, Virginia, USA

[2]Department of Statistics, University of South Carolina, Columbia, South Carolina, USA

[3]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA

[4]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Correspondence
Md S. Warasi, Department of Mathematics and Statistics, Radford University, 801 East Main St., Radford, VA, 24142, USA.
Email: msarker@radford.edu

Funding information
National Institutes of Health,
Grant/Award Number: R01 AI121351

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

Abstract

When screening a population for infectious diseases, pooling individual specimens (e.g., blood, swabs, urine, etc.) can provide enormous cost savings when compared to testing specimens individually. In the biostatistics literature, testing pools of specimens is commonly known as group testing or pooled testing. Although estimating a population-level prevalence with group testing data has received a large amount of attention, most of this work has focused on applications involving a single disease, such as human immunodeficiency virus. Modern methods of screening now involve testing pools and individuals for multiple diseases simultaneously through the use of multiplex assays. Hou et al. (2017, *Biometrics*, *73*, 656–665) and Hou et al. (2020, *Biostatistics*, *21*, 417–431) recently proposed group testing protocols for multiplex assays and derived relevant case identification characteristics, including the expected number of tests and those which quantify classification accuracy. In this article, we describe Bayesian methods to estimate population-level disease probabilities from implementing these protocols or any other multiplex group testing protocol which might be carried out in practice. Our estimation methods can be used with multiplex assays for two or more diseases while incorporating the possibility of test misclassification for each disease. We use chlamydia and gonorrhea testing data collected at the State Hygienic Laboratory at the University of Iowa to illustrate our work. We also provide an online R resource practitioners can use to implement the methods in this article.

KEYWORDS
Bayesian estimation, latent response, multiplex assay, pooled testing, screening

1 | INTRODUCTION

In group testing applications, individual specimens are combined into pools and tests are performed on the pools for a binary outcome (e.g., positive/negative, etc.). Individuals from negative pools are diagnosed to be negative, while individuals from positive pools are tested further to determine which ones are positive. Dorfman (1943) is credited with introducing this method to test American soldiers for syphilis during World War II. In the seminal article, nonoverlapping pools of individual specimens are formed in the first stage of testing, and positive pools are resolved by testing each individual one by one in the second stage. When the probability of disease is small, group testing protocols that implement a larger number of stages (Pilcher et al., 2005; Quinn et al., 2000) and/or overlapping pools (Martin et al., 2013) can further reduce the number of tests needed to identify positive individuals. The infectious disease literature documents numerous applications of group testing, including for human immunodeficiency virus (HIV), hepatitis B, and hepatitis C (Hourfar et al., 2008; Stramer, Notari, et al., 2013; Westreich et al., 2008), chlamydia and gonorrhea (Lindan et al., 2005), and West Nile virus (Busch et al., 2005). More recently, group testing has played a critical role in reducing laboratory testing loads when diagnosing individuals for SARS-CoV-2 (Abdalhamid et al., 2020; Bilder et al., 2021; Pilcher et al., 2020).

Statistical research in group testing generally falls into one of two categories: case identification and estimation. In the case identification problem, the goal is to characterize the efficiency and accuracy of group testing protocols with the usual objectives of minimizing the expected number of tests and/or maximizing classification accuracy (Kim et al., 2007). On the other hand, the estimation problem involves estimating a population-level probability of disease (Huang et al., 2017; Liu et al., 2012) or covariate-adjusted probabilities by using regression (Delaigle & Meister, 2011; McMahan et al., 2017; Wang et al., 2014). In both problems, the performance of group testing and its ability to offer cost-effective screening and surveillance has been documented extensively. However, most of the existing research in group testing, including those articles referenced above, has focused on a single disease.

In this article, we consider the estimation problem in group testing when multiplex assays are used to test specimens for multiple diseases simultaneously. Our work is largely motivated by the screening practices for *Chlamydia trachomatis* (CT) and *Neisseria gonorrhoeae* (NG), the bacteria that lead to chlamydia and gonorrhea infections, respectively. For example, the Aptima Combo 2 Assay (Hologic) and the cobas CT/NG Assay (Roche) are two multiplex assays commonly used by laboratories to test individuals for these bacteria. Recent advances in technology have seen the development of multiplex assays for more than two infections. For example, the BD MAX CT/GC/*Trichomonas vaginalis* (TV) Assay (Becton, Dickinson and Company) tests for CT, NG, and TV simultaneously, and the Allplex STI Essential Assay (Seegene) detects CT, NG, TV, and *Mycoplasma genitalium* (de Salazar et al., 2019). Commonly used triplex assays for HIV, hepatitis B, and hepatitis C have been compared in Stramer, Krysztof, et al. (2013), and, more recently, multiplex assays have been authorized by the US Food and Drug Administration for simultaneous detection of influenza and SARS-CoV-2 (Roche, 2020).

When compared to research for single diseases, estimating multiple population-level disease probabilities from group testing data has received far less attention. The original work on this problem is attributed to Hughes-Oliver and Rosenberger (2000), who developed D-optimal designs for estimation when assays are 100% accurate. Ding and Xiong (2015) and Li et al. (2017) proposed optimal designs to estimate probabilities for multiple independent diseases and two correlated diseases, respectively, while allowing for testing error. A practical limitation in these articles is that the methods are based only on outcomes from initially formed master pools; that is, subsequent testing results from resolving positive pools are not incorporated. Another limitation is that the assay accuracy rates are assumed to be 100% for each disease or they are assumed to be known. In some applications, reasonable estimates may be available for disease-specific sensitivities and specificities. A more flexible approach is to regard these population-level parameters as unknown and then estimate them simultaneously with the disease probabilities. This is the approach we espouse in this article.

Estimation for group testing with multiplex assays is challenging. When incorporating test misclassification, (a) the true disease statuses of each specimen tested are latent and are likely correlated and (b) the available data from group testing protocols may include multiple (possibly misclassified) testing outcomes on the same individual. Tebbs et al. (2013) and Warasi et al. (2016) proposed frequentist and Bayesian approaches for estimation with multiplex assays, respectively, but limited their investigation to Dorfman testing, that is, hierarchical group testing protocols which implement two stages only. In this article, we propose a Bayesian framework to estimate population-level disease probabilities and assay accuracy probabilities from *any* group testing protocol which uses multiplex assays. This includes higher-stage hierarchical and array-based protocols recently proposed in Hou et al. (2017) and Hou et al. (2020), respectively, and any other protocol that might be used in practice. In other words, the estimation framework we present herein is invariant to how the multiplex outcomes are recorded. Therefore, we can compare the accuracy and precision of population-level estimates for different group testing protocols which use multiplex assays. Until now, such a comparison has been missing in the literature.

Subsequent sections of this article are organized as follows. In Section 2, we describe notation, state assumptions, and derive the observed data likelihood which is applicable for any group testing protocol using multiplex assays. In Section 3, we present the specifics of our Bayesian estimation approach when assay accuracy probabilities (sensitivities and specificities) are known, including prior model specification and data augmentation steps to construct an efficient posterior sampling algorithm. In Section 4, we then generalize our approach to allow for assay accuracy probabilities to be unknown. In Section 5, we provide simulation evidence to assess the performance of our estimation methods and provide a comparison of the estimates for different group testing protocols. In Section 6, we illustrate our work by using CT/NG data collected at the State Hygienic Laboratory (SHL) at the University of Iowa. In Section 7, we conclude with a summary discussion and describe future work. Additional details and simulation evidence are provided in the Supporting Information.

2 | NOTATION AND PRELIMINARIES

Suppose N individuals are to be tested for $K \geq 2$ diseases using a group testing protocol. We assume all diagnostic test results are obtained from a multiplex assay which provides a positive/negative diagnosis for each disease each time it is used (on specimen pools or on individual specimens). For example, the SHL at the University of Iowa uses the Aptima Combo 2 Assay (AC2A) to detect CT and NG simultaneously (see Section 6). The current multiplex protocol at the SHL tests specimen pools (usually of size 4) with the AC2A. Pools testing positively for either disease are then resolved by testing each individual specimen with the same assay. Disease diagnoses are then determined from the individual tests.

To focus our ideas, the presentation in this article will assume there are $K = 2$ diseases (e.g., CT/NG, etc.). Generalizing our approach to $K \geq 2$ diseases is straightforward and is thus relegated to the Supporting Information. Let $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, \widetilde{Y}_{i2})'$ denote a vector of binary random variables which encode the true disease statuses of the ith individual, with $\widetilde{Y}_{ik} = 1(0)$ denoting the individual is truly positive (negative) for the kth disease, for $i = 1, 2, \ldots, N$ and $k = 1, 2$. We assume the $\widetilde{\mathbf{Y}}_i$'s are mutually independent with probability mass function $\mathrm{pr}(\widetilde{Y}_{i1} = \widetilde{y}_1, \widetilde{Y}_{i2} = \widetilde{y}_2 | \mathbf{p}) = p_{00}^{(1-\widetilde{y}_1)(1-\widetilde{y}_2)} p_{10}^{\widetilde{y}_1(1-\widetilde{y}_2)} p_{01}^{(1-\widetilde{y}_1)\widetilde{y}_2} p_{11}^{\widetilde{y}_1\widetilde{y}_2}$, where $\widetilde{y}_1, \widetilde{y}_2 \in \{0, 1\}$, $\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11})'$, and the nonnegative cell probabilities p_{00}, p_{10}, p_{01}, p_{11} satisfy $p_{00} + p_{10} + p_{01} + p_{11} = 1$. Therefore, the joint distribution of the true disease status vectors for all N individuals; that is, $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1', \widetilde{\mathbf{Y}}_2', \ldots, \widetilde{\mathbf{Y}}_N')'$, is given by

$$\pi(\widetilde{\mathbf{Y}} | \mathbf{p}) = \prod_{i=1}^{N} p_{00}^{(1-\widetilde{Y}_{i1})(1-\widetilde{Y}_{i2})} p_{10}^{\widetilde{Y}_{i1}(1-\widetilde{Y}_{i2})} p_{01}^{(1-\widetilde{Y}_{i1})\widetilde{Y}_{i2}} p_{11}^{\widetilde{Y}_{i1}\widetilde{Y}_{i2}}. \tag{1}$$

Note that estimating $\mathbf{p}$ using Equation (1) is straightforward if individual testing is used and the multiplex assay is 100% accurate for each disease. Otherwise, the random vector $\widetilde{\mathbf{Y}}$ is best regarded as latent.

The observed data in group testing consist of diagnostic test results collected as part of a testing protocol. These protocols are typically completed over $S \geq 2$ stages, where, within each stage, pooled or individual specimens are tested in response to the results from the previous stage. For example, as noted earlier, the SHL uses an $S = 2$ stage protocol where pools of specimens are tested in the first stage and individual specimens from positive pools are tested in the second. Hou et al. (2017) evaluated the utility of hierarchical group testing protocols using a larger number of stages, showing that $S = 3$ stage protocols conferred the smallest number of tests when screening for CT/NG in four western states in the United States (Alaska, Idaho, Oregon, and Washington). A three-stage hierarchical protocol uses an intermediate second stage with smaller-sized subpools; for example, first-stage pools of size 9, three second-stage pools of size 3, individual testing in the third stage. Hou et al. (2020) later proposed $S = 2$ and $S = 3$ stage multiplex protocols which use array testing (AT). In these (nonhierarchical) protocols, testing results arise from pooling rows and columns of overlapping specimens arranged in an array-like configuration.

In this article, we propose an estimation framework which is applicable for *any* group testing protocol using multiplex assays. To maintain this level of generality, we need notation that helps us track pool membership. Define the index set $\mathcal{P}_j \subseteq \{1, 2, \ldots, N\}$, $j = 1, 2, \ldots, J$, which identifies which individuals contribute to the jth pool; that is, $i \in \mathcal{P}_j$ when the ith individual is in the jth pool. Let $\widetilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, \widetilde{Z}_{j2})'$ denote a vector of binary random variables encoding the true status of the jth pool, where $\widetilde{Z}_{jk} = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_{ik} > 0)$, for $j = 1, 2, \ldots, J$ and $k = 1, 2$. In other words, the jth pool is truly positive (truly negative) for the kth disease if the pool contains at least one (no) positive individual(s) for disease k. Again, due to the effects of imperfect testing, the $\widetilde{\mathbf{Z}}_j$'s are not observed. Instead, we observe $\mathbf{Z}_j = (Z_{j1}, Z_{j2})'$, a vector of binary random variables encoding the test results for the jth pool, where $Z_{jk} = 1(0)$ if the jth pool tests positively (negatively) for the kth disease.

To allow for imperfect testing, we need to relate the observed testing results in $\mathbf{Z}_j$ to the true disease statuses in $\widetilde{\mathbf{Z}}_j$. We assume $\text{pr}(Z_{jk} = 1|\widetilde{Z}_{jk} = 1) = S_{e:jk}$ and $\text{pr}(Z_{jk} = 0|\widetilde{Z}_{jk} = 0) = S_{p:jk}$, for $j = 1, 2, \ldots, J$ and $k = 1, 2$. That is, $S_{e:jk}$ ($S_{p:jk}$) is the sensitivity (specificity) of the multiplex assay when testing the jth pool for the kth disease. We assume these accuracy probabilities do not depend on the number of true positive individuals in the jth pool, that is, there is no effect due to the dilution of positive individuals. However, our notation emphasizes $S_{e:jk}$ and $S_{p:jk}$ can be "pool-specific," affording us the flexibility to allow for different multiplex assays to be used and/or to have testing accuracy of a multiplex assay be a function of the jth pool size (see Section 4). The conditional distribution of the observed data $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2', \ldots, \mathbf{Z}_J')'$ given the individuals' true disease statuses $\widetilde{\mathbf{Y}}$ is

$$\pi(\mathbf{Z}|\widetilde{\mathbf{Y}}, \boldsymbol{\delta}) = \prod_{j=1}^{J}\prod_{k=1}^{2} S_{e:jk}^{Z_{jk}\widetilde{Z}_{jk}}(1 - S_{e:jk})^{(1-Z_{jk})\widetilde{Z}_{jk}} S_{p:jk}^{(1-Z_{jk})(1-\widetilde{Z}_{jk})}(1 - S_{p:jk})^{Z_{jk}(1-\widetilde{Z}_{jk})}, \tag{2}$$

where $\boldsymbol{\delta}$ is a vector that contains all assay accuracy probabilities; that is, the $S_{e:jk}$'s and $S_{p:jk}$'s for $j = 1, 2, \ldots, J$ and $k = 1, 2$. Note that the right-hand side of Equation (2) is $\pi(\mathbf{Z}|\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2, \ldots, \widetilde{\mathbf{Z}}_J, \boldsymbol{\delta})$, but this equals $\pi(\mathbf{Z}|\widetilde{\mathbf{Y}}, \boldsymbol{\delta})$ because $\widetilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, \widetilde{Z}_{j2})'$ and $\widetilde{Z}_{jk} = I(\sum_{i\in\mathcal{P}_j} \widetilde{Y}_{ik} > 0)$ are uniquely determined when the true disease statuses $\widetilde{\mathbf{Y}}$ are given. When writing Equation (2), we assume that testing results in $\mathbf{Z}$ are conditionally independent given the true statuses in $\widetilde{\mathbf{Y}}$ and that the values of $S_{e:jk}$ and $S_{p:jk}$ for one disease do not depend on the true status of the other disease (see Hou et al., 2020). Combining Equations (1) and (2), we can express the distribution of the observed data from *any* group testing protocol as

$$\pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta}) = \sum_{\widetilde{\mathbf{Y}}\in\{0,1\}^{2N}} \pi(\widetilde{\mathbf{Y}}|\mathbf{p})\pi(\mathbf{Z}|\widetilde{\mathbf{Y}}, \boldsymbol{\delta}), \tag{3}$$

where $\{0, 1\}^{2N}$ denotes the collection of all possible realizations of $\widetilde{\mathbf{Y}}$. This distribution is obtained by marginalizing the joint distribution of the observed testing responses and the individuals' latent statuses, that is, by summing $\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}|\mathbf{p}, \boldsymbol{\delta}) = \pi(\widetilde{\mathbf{Y}}|\mathbf{p})\pi(\mathbf{Z}|\widetilde{\mathbf{Y}}, \boldsymbol{\delta})$ over $\widetilde{\mathbf{Y}}$. This marginalization process requires computing the sum over the 2^{2N} possible realizations of $\widetilde{\mathbf{Y}}$, which can be computationally prohibitive in practical settings. For example, the Iowa CT/NG data considered in Section 6 involves $N = 14,450$ individuals.

3 | ESTIMATION WITH KNOWN ASSAY ACCURACY PROBABILITIES

To incorporate prior knowledge about the disease probabilities in $\mathbf{p}$ and the assay accuracy probabilities in $\boldsymbol{\delta}$, we take a Bayesian approach as in Warasi et al. (2016) who considered two-stage hierarchical protocols only. Our methods herein are more general and apply to *any* group testing protocol with multiplex assays. In this section, we consider the simpler setting where the assay accuracy probabilities in $\boldsymbol{\delta}$ are known. This assumption is then relaxed in Section 4.

3.1 | Posterior sampling

We assume a priori that $\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha})$; that is, the prior distribution for $\mathbf{p}$ is given by

$$\pi(\mathbf{p}) = B(\boldsymbol{\alpha}) p_{00}^{\alpha_{00}-1} p_{10}^{\alpha_{10}-1} p_{01}^{\alpha_{01}-1} p_{11}^{\alpha_{11}-1},$$

where $B(\boldsymbol{\alpha})$ is a normalizing constant and $\boldsymbol{\alpha} = (\alpha_{00}, \alpha_{10}, \alpha_{01}, \alpha_{11})'$ is a vector of known hyperparameters. Based on the observed data $\mathbf{Z}$, we then update our knowledge about $\mathbf{p}$ through its posterior distribution, given by $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta}) \propto \pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta})\pi(\mathbf{p})$. Unfortunately, this distribution involves $\pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta})$ whose calculation in Equation (3) is generally infeasible. Therefore, to facilitate posterior estimation, we develop a Markov chain Monte Carlo sampling algorithm that can draw realizations from $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$.

At the crux of this development is a data augmentation step which involves introducing individuals' true disease statuses as "missing data." Define the vector $\widetilde{\mathbf{V}}_i = (\widetilde{V}_{(00)i}, \widetilde{V}_{(10)i}, \widetilde{V}_{(01)i}, \widetilde{V}_{(11)i})'$ so that $\widetilde{V}_{(00)i} = 1$ when $\widetilde{\mathbf{Y}}_i' = (0,0)$, $\widetilde{V}_{(10)i} = 1$ when $\widetilde{\mathbf{Y}}_i' = (1,0)$, and so on. We introduce $\widetilde{\mathbf{V}}_i$ because it uniquely encodes the true disease status of the ith individual, and we can work out its full conditional distribution. Specifically, $\widetilde{\mathbf{V}}_i|\widetilde{\mathbf{Y}}_{-i}, \mathbf{p}, \mathbf{Z}, \boldsymbol{\delta} \sim \text{multinomial}(p^*_{(00)i}, p^*_{(10)i}, p^*_{(01)i}, p^*_{(11)i})$, where $\widetilde{\mathbf{Y}}_{-i}$ aggregates all N true disease status vectors except $\widetilde{\mathbf{Y}}_i$. Closed-form expressions for $p^*_{(00)i}$, $p^*_{(10)i}$, $p^*_{(01)i}$, and $p^*_{(11)i}$ are given in Appendix A in the Supporting Information. In addition, from Equation (1) and the form of the prior $\pi(\mathbf{p})$, it is easy to verify the full conditional $\pi(\mathbf{p}|\widetilde{\mathbf{Y}})$ is also Dirichlet with parameter $\boldsymbol{\alpha}^* = (\alpha^*_{00}, \alpha^*_{10}, \alpha^*_{01}, \alpha^*_{11})'$, where $\alpha^*_{uv} = \alpha_{uv} + \sum_{i=1}^N \widetilde{V}_{(uv)i}$ and $\widetilde{V}_{(uv)i} = \widetilde{Y}_{i1}^u(1-\widetilde{Y}_{i1})^{1-u}\widetilde{Y}_{i2}^v(1-\widetilde{Y}_{i2})^{1-v}$, for $u, v \in \{0,1\}$. These two distributions, $\pi(\widetilde{\mathbf{V}}_i|\widetilde{\mathbf{Y}}_{-i}, \mathbf{p}, \mathbf{Z}, \boldsymbol{\delta})$ and $\pi(\mathbf{p}|\widetilde{\mathbf{Y}})$, can be used to construct an efficient algorithm to sample from $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$ as we now describe.

POSTERIOR SAMPLING ALGORITHM

1. Choose $\mathbf{p}^{(0)} = (p^{(0)}_{00}, p^{(0)}_{10}, p^{(0)}_{01}, p^{(0)}_{11})'$ as an initial value and simulate $\widetilde{\mathbf{Y}}^{(0)}_i = (\widetilde{Y}^{(0)}_{i1}, \widetilde{Y}^{(0)}_{i2})'$, for $i = 1, 2, \dots, N$, from the population-level multinomial model when $\mathbf{p} = \mathbf{p}^{(0)}$. Set $g = 1$.
2. For $i = 1, 2, \dots, N$, sample $\widetilde{\mathbf{V}}^{(g)}_i = (\widetilde{V}^{(g)}_{(00)i}, \widetilde{V}^{(g)}_{(10)i}, \widetilde{V}^{(g)}_{(01)i}, \widetilde{V}^{(g)}_{(11)i})'$ from

$$\widetilde{\mathbf{V}}_i|\widetilde{\mathbf{Y}}^{(g)}_{-i}, \mathbf{p}^{(g-1)}, \mathbf{Z}, \boldsymbol{\delta} \sim \text{multinomial}(p^*_{(00)i}, p^*_{(10)i}, p^*_{(01)i}, p^*_{(11)i}),$$

   where $\widetilde{\mathbf{Y}}^{(g)}_{-i} = (\widetilde{\mathbf{Y}}^{(g)'}_1, \dots, \widetilde{\mathbf{Y}}^{(g)'}_{i-1}, \widetilde{\mathbf{Y}}^{(g-1)'}_{i+1}, \dots, \widetilde{\mathbf{Y}}^{(g-1)'}_N)'$ and $\widetilde{\mathbf{Y}}^{(g)}_i = (\widetilde{V}^{(g)}_{(10)i} + \widetilde{V}^{(g)}_{(11)i}, \widetilde{V}^{(g)}_{(01)i} + \widetilde{V}^{(g)}_{(11)i})'$.
3. Sample $\mathbf{p}^{(g)}$ from $\mathbf{p}|\widetilde{\mathbf{Y}}^{(g)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^*)$, where $\widetilde{\mathbf{Y}}^{(g)} = (\widetilde{\mathbf{Y}}^{(g)'}_1, \widetilde{\mathbf{Y}}^{(g)'}_2, \dots, \widetilde{\mathbf{Y}}^{(g)'}_N)'$.
4. Set $g = g + 1$ and repeat steps 2–4 while $g < G$, the number of Gibbs iterates.

Two remarks are in order. First, it is worth emphasizing the multinomial cell probabilities in Step 2 are functions of the observed data in $\mathbf{Z}$ (see Appendix A). This is why the posterior sampling algorithm above can be implemented with *any* group testing protocol using multiplex assays. In other words, different protocols will give rise to different types of observed data $\mathbf{Z}$ but the sampling procedure remains unchanged. Second, in practice, we recommend selecting the number of Gibbs iterates G to be large; for example, $G = 10,000$, after discarding the first thousand or so iterates for burn-in purposes (see Sections 5 and 6 for specific illustrations). For inference, the sample mean of the G iterates can be used as an estimate of the posterior mean of $\mathbf{p}$; that is, $E(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$, and credible intervals can be constructed by using the appropriate sample quantiles.

3.2 | Maximum a posteriori (MAP) estimation

It is well known that group testing is most beneficial when the probability of disease is low. Otherwise, most initially formed master pools could test positively and the motivation for pooling specimens would quickly diminish. In our multiplex setting, this means p_{00}, the probability an individual is disease-free, may be close to unity, and the population-level parameters p_{10}, p_{01}, p_{11}, and marginal probabilities $p_{1+} = p_{10} + p_{11}$ and $p_{+1} = p_{01} + p_{11}$ may all be close to zero depending on the diseases under investigation. Because of these constraints on the parameter space, the marginal posterior distributions from $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$ may be heavily skewed and summarizing the posterior distribution with a mean (or median) estimate may be unwise.

In such instances, reporting a posterior mode may be more sensible. We therefore describe an approach to find the MAP estimate; that is, the mode of $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$. Using the same missing data conceptualization as in Section 3.1, we use the expectation-maximization (EM) algorithm to maximize $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$. This algorithm involves evaluating $Q(\mathbf{p}, \mathbf{p}^{(t)})$, the conditional expectation of the logarithm of the augmented posterior $\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}|\boldsymbol{\delta})\pi(\mathbf{p}) = \pi(\widetilde{\mathbf{Y}}|\mathbf{p})\pi(\mathbf{Z}|\widetilde{\mathbf{Y}}, \boldsymbol{\delta})\pi(\mathbf{p})$ given the observed data and current parameter value $\mathbf{p}^{(t)}$, and then maximizing it as a function of $\mathbf{p}$. One then iterates between these two steps until convergence. This can be accomplished by using the steps described below.

MAP ESTIMATION VIA EM ALGORITHM

1. Choose $\mathbf{p}^{(0)} = (p^{(0)}_{00}, p^{(0)}_{10}, p^{(0)}_{01}, p^{(0)}_{11})'$ as an initial value and simulate $\widetilde{\mathbf{Y}}^{(0)}_i = (\widetilde{Y}^{(0)}_{i1}, \widetilde{Y}^{(0)}_{i2})'$, for $i = 1, 2, \dots, N$, from the population-level multinomial model when $\mathbf{p} = \mathbf{p}^{(0)}$. Set $t = 0$.

2. (E-Step): For $i = 1, 2, \dots, N$,
   - sample $\widetilde{\mathbf{V}}_i^{(g)} = (\widetilde{V}_{(00)i}^{(g)}, \widetilde{V}_{(10)i}^{(g)}, \widetilde{V}_{(01)i}^{(g)}, \widetilde{V}_{(11)i}^{(g)})'$, for $g = 1, 2, \dots, G$, from $\widetilde{\mathbf{V}}_i | \widetilde{\mathbf{Y}}_{-i}^{(g)}, \mathbf{p}^{(t)}, \mathbf{Z}, \boldsymbol{\delta} \sim \text{multinomial}(p^*_{(00)i}, p^*_{(10)i}, p^*_{(01)i}, p^*_{(11)i})$, where G is the number of Gibbs iterates;
   - calculate the sample mean $G^{-1} \sum_{g=1}^{G} (\widetilde{V}_{(00)i}^{(g)}, \widetilde{V}_{(10)i}^{(g)}, \widetilde{V}_{(01)i}^{(g)}, \widetilde{V}_{(11)i}^{(g)})'$ as an estimate of the conditional expectation $E[(\widetilde{V}_{(00)i}, \widetilde{V}_{(10)i}, \widetilde{V}_{(01)i}, \widetilde{V}_{(11)i})' | \mathbf{Z}, \boldsymbol{\delta}; \mathbf{p}^{(t)}]$.
3. (M-Step): Calculate $\mathbf{p}^{(t+1)}$ using the solution in Appendix B in the Supporting Information; that is, this maximizer depends on $E[(\widetilde{V}_{(00)i}, \widetilde{V}_{(10)i}, \widetilde{V}_{(01)i}, \widetilde{V}_{(11)i})' | \mathbf{Z}, \boldsymbol{\delta}; \mathbf{p}^{(t)}]$ and exists in closed form.
4. Set $t = t + 1$, and repeat steps 2–4 until the maximum absolute difference in $\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}$ is less than ϵ, where ϵ is small.

We again make brief remarks. First, because Step 2 uses a Gibbs sampler to estimate the conditional expectation $E[(\widetilde{V}_{(00)i}, \widetilde{V}_{(10)i}, \widetilde{V}_{(01)i}, \widetilde{V}_{(11)i})' | \mathbf{Z}, \boldsymbol{\delta}; \mathbf{p}^{(t)}]$, calculating the MAP estimate of $\mathbf{p}$ takes longer than simply summarizing $\pi(\mathbf{p}|\mathbf{Z}, \boldsymbol{\delta})$ with the posterior mean from Section 3.1. However, because the M-Step solution exists in closed form, this additional time required is generally not prohibitive. We again recommend using a large number of Gibbs iterates in Step 2 (e.g., $G = 10,000$, etc.) after a sufficient burn in (see Sections 5 and 6). Second, when a uniform prior distribution $\pi(\mathbf{p})$ is used, that is, setting $\alpha_{00} = \alpha_{10} = \alpha_{01} = \alpha_{11} = 1$, the MAP estimate of $\mathbf{p}$ coincides with the maximum likelihood estimate (MLE) of $\mathbf{p}$, a potential preference for users wanting to report frequentist-based point estimates. Finally, when the assay accuracy probabilities in $\boldsymbol{\delta}$ are unknown, our simulation results in Section 5 and Appendix D in the Supporting Information demonstrate that MAP estimates of $\mathbf{p}$ and $\boldsymbol{\delta}$ can be more accurate than other posterior estimates. We now generalize our methodology to allow for this situation.

4 | ESTIMATION WITH UNKNOWN ASSAY ACCURACY PROBABILITIES

Our goal now is to estimate the population-level infection probabilities in $\mathbf{p}$ and the assay accuracy probabilities in $\boldsymbol{\delta}$ simultaneously. As we demonstrate, this can be accomplished by taking our algorithms in Section 3 and adding appropriate steps for the conditional distribution and MAP solution of $\boldsymbol{\delta}$. Such an extension is practically useful in incorporating the uncertainty in $\boldsymbol{\delta}$. For example, although manufacturers will typically report values of sensitivity and specificity for multiplex assays (for each disease) in their product literature, these values are usually obtained from small pilot studies involving specimens whose true disease statuses are known in advance. The practice of ostensibly regarding these values as "correct" can lead to two potential problems. First, doing so ignores the sampling error incurred from having to estimate these values in small feasibility experiments. Second, the population under investigation (e.g., high-risk females in Iowa, etc.) may differ substantially from the one which was used to validate the multiplex assay initially.

Extending the approach in McMahan et al. (2017) for single diseases, let $S_{e:(l)k}$ and $S_{p:(l)k}$ denote the sensitivity and specificity of the lth assay for the kth disease, for $k = 1, 2$ and $l = 1, 2, \dots, L$, and let $\mathcal{M}(l) = \{j : \text{the } l\text{th assay tests pool } j\}$ denote the index set of the specimens tested by the lth assay, for $j = 1, 2, \dots, J$. Our use of the set $\mathcal{M}(l)$ simply allows us to reparameterize the exposition in Section 2. For example, at the SHL in Iowa, the AC2A assay is used for all specimens tested in pools and individually. If this assay performs the same when testing pools and individuals, then $L = 1$ and the parameter vector $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2})'$. On the other hand, if the performance of the AC2A depends on whether pools or individuals are tested, one could envision one set of assay accuracy probabilities for pools ($l = 1$) and a separate set for individuals ($l = 2$). This situation would correspond to $L = 2$ and the parameter vector would become $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2}, S_{e:(2)1}, S_{e:(2)2}, S_{p:(2)1}, S_{p:(2)2})'$.

Under our reparameterization, the distribution of the observed data $\mathbf{Z}$ from *any* group testing protocol in Equation (3) can be written as

$$\pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta}) = \sum_{\widetilde{\mathbf{Y}} \in \{0,1\}^{2N}} \left[\pi(\widetilde{\mathbf{Y}}|\mathbf{p}) \prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} S_{e:(l)k}^{Z_{jk}\widetilde{Z}_{jk}} (1 - S_{e:(l)k})^{(1-Z_{jk})\widetilde{Z}_{jk}} \times S_{p:(l)k}^{(1-Z_{jk})(1-\widetilde{Z}_{jk})} (1 - S_{p:(l)k})^{Z_{jk}(1-\widetilde{Z}_{jk})} \right],$$

where now both $\mathbf{p}$ and the assay accuracy probabilities in $\boldsymbol{\delta}$ are regarded as unknown. To incorporate the uncertainty in $\boldsymbol{\delta}$, we use beta prior distributions for each sensitivity and specificity parameter, that is, $S_{e:(l)k} \sim \text{beta}(a_{lk}, b_{lk})$ and

$S_{p:(l)k} \sim \text{beta}(c_{lk}, d_{lk})$, for $k = 1, 2$ and $l = 1, 2, \dots, L$. If all prior distributions are independently specified, then the posterior distribution of $\mathbf{p}$ and $\boldsymbol{\delta}$ satisfies $\pi(\mathbf{p}, \boldsymbol{\delta}|\mathbf{Z}) \propto \pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta})\pi(\mathbf{p}) \prod_{l=1}^{L} \pi(S_{e:(l)k})\pi(S_{p:(l)k})$, where $\pi(\mathbf{Z}|\mathbf{p}, \boldsymbol{\delta})$ is given above and $\pi(S_{e:(l)k})$ and $\pi(S_{p:(l)k})$ denote the beta priors. As noted earlier, data from multiplex assay feasibility studies can be used to elicit informative prior distributions for $S_{e:(l)k}$ and $S_{p:(l)k}$. Of course, in the absence of any prior knowledge, uniform priors can also be used.

Both the posterior sampling and EM algorithms in Section 3 can be generalized to estimate $\mathbf{p}$ and $\boldsymbol{\delta}$ simultaneously. To sample from $\pi(\mathbf{p}, \boldsymbol{\delta}|\mathbf{Z})$, we note that $S_{e:(l)k}|\mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{beta}(a^*_{lk}, b^*_{lk})$ and $S_{p:(l)k}|\mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{beta}(c^*_{lk}, d^*_{lk})$, where $a^*_{lk} = a_{lk} + \sum_{j\in\mathcal{M}(l)} Z_{jk}\widetilde{Z}_{jk}$, $b^*_{lk} = b_{lk} + \sum_{j\in\mathcal{M}(l)}(1 - Z_{jk})\widetilde{Z}_{jk}$, $c^*_{lk} = c_{lk} + \sum_{j\in\mathcal{M}(l)}(1 - Z_{jk})(1 - \widetilde{Z}_{jk})$, and $d^*_{lk} = d_{lk} + \sum_{j\in\mathcal{M}(l)} Z_{jk}(1 - \widetilde{Z}_{jk})$. Therefore, because all other conditionals remain unchanged, one can take the posterior sampling algorithm described in Section 3.1 and insert one additional step. Similarly, to calculate the MAP estimate of $\mathbf{p}$ and $\boldsymbol{\delta}$, the EM algorithm in Section 3.2 can be easily amended. The conditional expectation of the logarithm of the augmented posterior given the observed data and current parameter value, now written $Q(\mathbf{p}, \boldsymbol{\delta}, \mathbf{p}^{(t)}, \boldsymbol{\delta}^{(t)})$, also has a closed-form solution in the M-step. The complete algorithms are given in Appendix C in the Supporting Information.

5 | SIMULATION EVIDENCE

We performed a comprehensive simulation study to evaluate the performance of our estimation methods. This study included examining three hierarchical group testing protocols (H2, H3, and H4) from Hou et al. (2017) and one AT protocol from Hou et al. (2020). We now briefly describe these protocols.

5.1 | Multiplex protocols and simulation description

A hierarchical group testing protocol is carried out by first testing a nonoverlapping master pool of individual specimens. If this pool tests negatively, then each individual in the pool is declared to be negative. If this pool tests positively, the master pool is divided into nonoverlapping subpools of specimens. Two-stage Dorfman protocols (H2) revert to individual testing in the second stage, while higher-stage protocols use smaller sized subpools during intermediate stages of testing before individual testing is used in the final stage. In AT, individual specimens are arranged in a square array configuration forming row and column master pools which are tested in the first stage. Individuals in positive row/column intersections are tested in the second stage along with other individuals whose statuses are potentially unknown because of testing errors (Hou et al., 2020). The overarching message from Hou et al. (2017, 2020) is that higher stage hierarchical protocols (H3, H4) and AT can substantially reduce the number of tests needed when compared to H2, especially when the probability of at least one disease $1 - p_{00}$ is small.

This prompts an obvious question. When compared to H2, how do H3, H4, and AT perform in terms of estimation? One might hypothesize that because H3, H4, and AT generally require fewer tests, fewer observations would be available and thus the estimation performance for these protocols would be degraded. On the other hand, it could be that H3 and H4 implement more tests "where it counts," that is, on individuals who are more likely to be positive, and AT uses master pools (rows and columns) that consist of overlapping individuals. In the presence of testing errors, more replicate tests on potentially positive individuals may actually improve estimation—despite H3, H4, and AT requiring fewer tests overall.

We simulated the execution of each protocol (H2, H3, H4, and AT) using two configurations of the disease probabilities, $\mathbf{p} = (0.95, 0.02, 0.02, 0.01)'$ and $(0.990, 0.004, 0.004, 0.002)'$; we henceforth call these Configurations I and II, respectively. The first configuration was chosen to represent the overall prevalence of CT/NG in higher risk populations, while the second configuration allows for two rarer diseases, each with a marginal probability of $0.004 + 0.002 = 0.006$. For each of H2, H3, H4, and AT, Table 1 lists the specific protocol which minimizes the expected number of tests when $S_{e:(1)k} = 0.95$ and $S_{p:(1)k} = 0.99$, for $k = 1, 2$. For example, the entry "5 : 1" for H2 under Configuration I means that master pools of size 5 reduce the number of tests as much as possible on average among all two-stage hierarchical protocols. Similarly, the entry "9 : 3 : 1" for H3 means that master pools of size 9 are used in the first stage, three subpools of size 3 are used in the second stage, and individual testing is used in the third. We determine these pool sizes using the optimization methods described in Hou et al. (2017, 2020).

For each protocol and disease probability configuration, we simulated the true disease statuses of $N = 5000$ individuals and randomly assigned these individuals to appropriately sized master pools. We then simulated the testing outcomes on

TABLE 1 Testing protocols in Section 5. Hierarchical protocols H2, H3, and H4 use two, three, and four stages, respectively (Hou et al., 2017). Array testing (AT) uses square arrays (Hou et al., 2020). The protocols listed below minimize the expected number of tests per individual specimen. "Configuration I" uses $\mathbf{p} = (0.95, 0.02, 0.02, 0.01)'$ and "Configuration II" uses $\mathbf{p} = (0.990, 0.004, 0.004, 0.002)'$.

Configuration I		**Configuration II**	
Protocol	**Pool sizes**	**Protocol**	**Pool sizes**
H2	5 : 1	H2	11 : 1
H3	9 : 3 : 1	H3	25 : 5 : 1
H4	18 : 6 : 3 : 1	H4	48 : 12 : 4 : 1
AT	11 × 11	AT	29 × 29

pools and individuals (allowing for potential testing error) to produce data that would be available for estimation purposes. This entire process was repeated $B = 500$ times, providing us with 500 independent data sets for each protocol under Configurations I and II. Note that in some cases smaller sized master pools were formed when there were remainder individuals. For example, 555 master pools of size 9 were formed for the "9 : 3 : 1" H3 protocol listed in Table 1; the remaining five individuals were tested in a master pool of size 5 and resolved using H2. This practice of using H2 for remainder pools was applied uniformly in all cases to ensure a fair comparison among the protocols.

In all simulations, we used $G = 10,000$ Gibbs iterates after discarding the first 2000 for burn-in purposes, and we used $\mathbf{p}^{(0)} = (0.92, 0.05, 0.02, 0.01)'$ and $\boldsymbol{\delta}^{(0)} = (0.96, 0.96, 0.98, 0.98)'$ as initial values for the disease probabilities and assay accuracy probabilities, respectively. All posterior measures of variability have been calculated by thinning with every fifth Gibbs iterate selected. For the simulations reported in this section, independent investigations on our part revealed that perturbing these selections (i.e., using different starting values, using different numbers of Gibbs iterates, thinning differently) did not have a large impact on the results.

5.2 | Simulation results

Tables 2 and 3 show the estimation results for both disease probability configurations when the assay accuracy probabilities in $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2})'$ are assumed to be unknown. Table 2 shows the results for estimating $\mathbf{p}$ while Table 3 presents those for estimating $\boldsymbol{\delta}$. When $\boldsymbol{\delta}$ is treated to be known (Section 3), we provide a table of estimation results for $\mathbf{p}$ in Appendix D. Unless otherwise stated, we used flat priors for both $\mathbf{p}$ and $\boldsymbol{\delta}$, that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1, 1)$, and $S_{p:(1)k} \sim \text{beta}(1, 1)$, for $k = 1, 2$. We selected these noninformative priors for two reasons. First, these distributions give us the most challenging case for estimation because no useful prior information is injected into the model. Second, our use of a flat prior for $\mathbf{p}$ produces MAP estimates which should largely coincide with the MLE of $\mathbf{p}$. When the Dirichlet($\mathbf{1}_4$) distribution is specified a priori, MAP and MLE of $\mathbf{p}$ will be identical when $\boldsymbol{\delta}$ is known and should be approximately equal otherwise.

In both Tables 2 and 3, we present the sample mean ("Est") of the posterior mean (Mean) and MAP estimates calculated from $B = 500$ independent data sets. We also report two measures of posterior variability: "SD," which is the sample standard deviation of the 500 posterior estimates, and "SE," which is the posterior standard deviation of the Gibbs iterates retained for one data set and then averaged over the 500 data sets. To compare the four protocols (H2, H3, H4, and AT) in terms of classification, we also recorded the average and standard deviation of the number of tests needed to classify each of the 5000 individuals as positive/negative for each disease. These quantities are denoted in Table 2 by $\overline{T}$ and S_T, respectively.

Table 2 reveals the averaged mean and MAP estimates of $\mathbf{p}$ are on target for both configurations of the disease probabilities across all four group testing protocols. An intriguing finding is that if one moves from Dorfman testing (H2) to one of the more complex protocols (H3, H4, or AT), the posterior variability gets no larger and may actually decrease slightly. This is interesting because H3, H4, and AT all require fewer tests to complete on average. For example, for Configuration I with $\mathbf{p} = (0.95, 0.02, 0.02, 0.01)'$, moving from H2 to H3 decreases the average number of tests by approximately 14.6% (2166.6 tests for H2; 1850.8 tests for H3), yet the posterior mean and MAP estimates for H3 are as good as or better than those for H2. Similar observations can be made for H4 and AT in terms of estimation performance, despite these protocols also offering a large reduction in the number of tests needed.

Moving to the assay accuracy probabilities in Table 3, one does observe a slight degradation in performance of Dorfman testing (H2) when attempting to estimate the sensitivity parameters $S_{e:(1)1}$ and $S_{e:(1)2}$ using a posterior mean, especially for Configuration II with $\mathbf{p} = (0.990, 0.004, 0.004, 0.002)'$. This should not be surprising because with so few positive individ-

TABLE 2 Simulation results for the posterior mean (Mean) and the maximum a posteriori (MAP) estimates of $\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11})'$ when assay accuracy probabilities are unknown. Estimates ("Est") are averages over $B = 500$ Monte Carlo data sets, "SD" is the sample standard deviation of the 500 estimates, and "SE" is the estimated posterior standard deviation as described in Section 5.2. Flat priors have been used for all parameters; that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1, 1)$, and $S_{p:(1)k} \sim \text{beta}(1, 1)$, for $k = 1, 2$. The mean $\overline{T}$ and standard deviation S_T of the number of tests are also shown; the percentage reduction in $\overline{T}$ is compared to H2 from Tebbs et al. (2013) and Warasi et al. (2016).

			H2			H3			H4			AT		
			Est	SD	SE	Est	SD	SE	Est	SD	SE	Est	SD	SE
Configuration I	$p_{00} = 0.95$	Mean	0.948	0.0037	0.0039	0.949	0.0033	0.0034	0.949	0.0033	0.0033	0.949	0.0034	0.0034
	$p_{10} = 0.02$		0.021	0.0024	0.0026	0.020	0.0023	0.0022	0.020	0.0021	0.0022	0.020	0.0022	0.0022
	$p_{01} = 0.02$		0.021	0.0025	0.0026	0.020	0.0023	0.0022	0.020	0.0022	0.0022	0.020	0.0023	0.0022
	$p_{11} = 0.01$		0.010	0.0015	0.0015	0.010	0.0014	0.0015	0.010	0.0014	0.0014	0.010	0.0014	0.0015
		MAP	0.950	0.0037	0.0039	0.950	0.0033	0.0034	0.950	0.0033	0.0033	0.950	0.0033	0.0034
			0.020	0.0024	0.0026	0.020	0.0023	0.0022	0.020	0.0021	0.0022	0.020	0.0022	0.0022
			0.020	0.0025	0.0026	0.020	0.0022	0.0022	0.020	0.0022	0.0022	0.020	0.0022	0.0022
			0.010	0.0015	0.0015	0.010	0.0014	0.0015	0.010	0.0014	0.0014	0.010	0.0014	0.0015
	$\overline{T}$		2166.6			1850.8 (14.6%)			1858.3 (14.2%)			1729.1 (20.2%)		
	S_T		66.8			74.7			87.8			77.1		
Configuration II	$p_{00} = 0.990$	Mean	0.988	0.0018	0.0020	0.989	0.0016	0.0016	0.989	0.0015	0.0016	0.989	0.0015	0.0016
	$p_{10} = 0.004$		0.005	0.0012	0.0013	0.004	0.0010	0.0011	0.004	0.0009	0.0010	0.004	0.0010	0.0010
	$p_{01} = 0.004$		0.005	0.0012	0.0013	0.004	0.0010	0.0011	0.004	0.0010	0.0010	0.004	0.0010	0.0010
	$p_{11} = 0.002$		0.002	0.0007	0.0007	0.002	0.0006	0.0007	0.002	0.0006	0.0007	0.002	0.0007	0.0007
		MAP	0.990	0.0019	0.0020	0.990	0.0016	0.0016	0.990	0.0015	0.0016	0.990	0.0016	0.0016
			0.004	0.0012	0.0013	0.004	0.0011	0.0011	0.004	0.0010	0.0010	0.004	0.0010	0.0010
			0.004	0.0012	0.0013	0.004	0.0010	0.0011	0.004	0.0010	0.0010	0.004	0.0010	0.0010
			0.002	0.0007	0.0007	0.002	0.0006	0.0007	0.002	0.0006	0.0007	0.002	0.0007	0.0007
	$\overline{T}$		1047.6			675.2 (35.5%)			582.9 (44.4%)			755.3 (27.9%)		
	S_T		78.9			64.5			63.1			90.1		

uals for either disease, posterior distributions are highly skewed and thus the mean may not be an ideal choice. However, Table 3 also shows that MAP estimation in this setting is much improved for H2, and that MAP estimation with H3, H4, or AT appears to recover $S_{e:(1)1}$ and $S_{e:(1)2}$ nearly perfectly on average. In addition, when compared to H2, the posterior distributions of $S_{e:(1)1}$ and $S_{e:(1)2}$ are less variable (smaller SD/SE) when using H3, H4, and AT under both disease probability configurations.

Following the recommendations of an anonymous reviewer, we have performed additional simulation studies which use a smaller sample size ($N = 1000$), a larger number of assays ($L = 2$), and we have investigated the performance of our methods when assay accuracy probabilities in $\boldsymbol{\delta}$ are substantially lower. In these more challenging settings for estimation, the use of informative prior distributions for $\mathbf{p}$ and/or $\boldsymbol{\delta}$ can be useful. Another reviewer astutely noted that when assays are perfect, that is, when $S_{e:(1)1} = S_{e:(1)2} = S_{p:(1)1} = S_{p:(1)2} = 1$, all group testing protocols will produce the same estimates of $\mathbf{p}$ because all true individual disease statuses are recoverable. Because retesting positive pools may confer limited utility in this setting (Chen & Swallow, 1990), this has motivated us to evaluate the use of master pool testing (MPT) as a means to estimate $\mathbf{p}$ in the presence of imperfect testing; see Tu et al. (1995) and McMahan et al. (2017). All additional simulation studies are summarized in Appendix D.

6 | CHLAMYDIA AND GONORRHEA APPLICATION

Chlamydia and gonorrhea are two of the most common sexually transmitted diseases in the United States (Centers for Disease Control and Prevention, 2022). Infected individuals can develop serious health-related complications, including pelvic inflammatory disease, infertility, and ectopic pregnancy. Public health laboratories across the United States continually perform surveillance for these diseases. For example, the SHL at the University of Iowa performs thousands of tests

TABLE 3 Simulation results for the posterior mean (Mean) and the maximum a posteriori (MAP) estimates of $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2})'$. Estimates ("Est") are averages over $B = 500$ Monte Carlo data sets, "SD" is the sample standard deviation of the 500 estimates, and "SE" is the estimated posterior standard deviation as described in Section 5.2. Flat priors have been used for all parameters; that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1,1)$, and $S_{p:(1)k} \sim \text{beta}(1,1)$, for $k = 1, 2$.

			H2			H3			H4			AT		
			Est	**SD**	**SE**	**Est**	**SD**	**SE**	**Est**	**SD**	**S3**	**Est**	**SD**	**SE**
Configuration I	$S_{e:(1)1} = 0.95$	Mean	0.938	0.024	0.027	0.947	0.015	0.015	0.948	0.011	0.012	0.948	0.014	0.015
	$S_{e:(1)2} = 0.95$		0.942	0.023	0.027	0.946	0.015	0.015	0.947	0.012	0.012	0.947	0.014	0.015
	$S_{p:(1)1} = 0.99$		0.990	0.003	0.004	0.989	0.004	0.004	0.989	0.003	0.004	0.989	0.005	0.005
	$S_{p:(1)2} = 0.99$		0.990	0.004	0.004	0.989	0.004	0.004	0.989	0.003	0.004	0.989	0.004	0.005
		MAP	0.946	0.025	0.027	0.950	0.014	0.015	0.950	0.011	0.012	0.950	0.014	0.015
			0.954	0.024	0.027	0.951	0.015	0.015	0.950	0.012	0.012	0.952	0.014	0.015
			0.990	0.004	0.004	0.990	0.003	0.004	0.990	0.003	0.004	0.990	0.004	0.005
			0.990	0.004	0.004	0.990	0.004	0.004	0.990	0.003	0.004	0.989	0.005	0.005
Configuration II	$S_{e:(1)1} = 0.95$	Mean	0.889	0.051	0.067	0.930	0.033	0.036	0.939	0.026	0.027	0.931	0.034	0.036
	$S_{e:(1)2} = 0.95$		0.889	0.055	0.067	0.930	0.032	0.037	0.936	0.026	0.028	0.930	0.031	0.037
	$S_{p:(1)1} = 0.99$		0.990	0.004	0.005	0.989	0.004	0.005	0.988	0.005	0.006	0.988	0.005	0.005
	$S_{p:(1)2} = 0.99$		0.990	0.004	0.005	0.989	0.004	0.005	0.988	0.005	0.006	0.988	0.005	0.005
		MAP	0.936	0.058	0.067	0.948	0.037	0.036	0.951	0.027	0.027	0.948	0.038	0.036
			0.938	0.061	0.067	0.949	0.036	0.037	0.948	0.027	0.028	0.947	0.035	0.037
			0.991	0.004	0.005	0.991	0.005	0.005	0.990	0.005	0.006	0.990	0.006	0.005
			0.990	0.004	0.005	0.990	0.005	0.005	0.990	0.005	0.006	0.990	0.005	0.005

TABLE 4 CT/NG testing protocols in Section 6. The configurations below minimize the expected number of tests per individual. AC2A accuracy probabilities for CT and NG are also shown.

	Protocol	**Pool sizes**		**Sensitivity**	**Specificity**
Urine	H2	4 : 1	CT	$S_{e:(1)1} = 0.947$	$S_{p:(1)1} = 0.989$
	H3	9 : 3 : 1	NG	$S_{e:(1)2} = 0.913$	$S_{p:(1)2} = 0.993$
	AT	8 × 8			
Swab	H2	4 : 1	CT	$S_{e:(1)1} = 0.942$	$S_{p:(1)1} = 0.976$
	H3	9 : 3 : 1	NG	$S_{e:(1)2} = 0.992$	$S_{p:(1)2} = 0.987$
	AT	8 × 8			

Abbreviations: AC2A, Aptima Combo 2 Assay; AT, array testing; CT, *Chlamydia trachomatis*; NG, *Neisseria gonorrhoeae*.

each year for Iowa residents using the Dorfman (H2) protocol with the AC2A. We use a data set provided to us by our collaborators at the SHL to further investigate the estimation methods in this article.

Our data consist of the CT/NG test results for 14,450 females tested during the 2014 calendar year. Urine and cervical swab specimens were collected from these individuals at different locations in Iowa and were transported to the SHL for testing. Among the 14,450 females, 4402 contributed urine specimens while 10,048 contributed cervical swab specimens. The AC2A was used to diagnose all specimens, whether in pools or individually, and Dorfman's H2 protocol was used to resolve all positive pools. In total, there were 2395 master pools of size 4, 12 pools of size 3, and 1 pool of size 2. All other specimens received at the SHL were tested individually. Based on the results of all individual tests performed, individuals were classified as positive or negative for each disease. A summary of these classification results is shown in Appendix E in the Supporting Information.

We use the classification results to perform a feasibility study comparing the estimation results from H2 to those from other group testing protocols. To do this, we treat the final disease classifications in the Iowa data as true statuses, assign individuals to initial master pools, and simulate the test responses for H2, H3, and AT, as described in Section 5. We do not use the H4 protocol in this illustration because the marginal disease probabilities of CT and NG are too large for the sample of individuals we have. The specific configurations of H2, H3, and AT are shown in Table 4, which were determined the same way as in Section 5, that is, by minimizing the expected number of tests per individual specimen. For verisimilitude,

TABLE 5 CT/NG pooling feasibility study. Posterior estimates of $\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11})'$ when assay accuracy probabilities are unknown. Estimates ("Est") are averages over $B = 500$ Monte Carlo data sets, and "SE" is the estimated posterior standard deviation as described in Section 5.2. Flat priors have been used for all parameters; that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1, 1)$, and $S_{p:(1)k} \sim \text{beta}(1, 1)$, for $k = 1, 2$. The mean $\overline{T}$ and standard deviation S_T of the number of tests are also shown; the percentage reduction in $\overline{T}$ is compared to H2 from Tebbs et al. (2013) and Warasi et al. (2016).

				H2		H3		AT	
Stratum		**CT**	**NG**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**
Urine	Mean	−	−	$\hat{p}_{00} = 0.907$	0.0065	$\hat{p}_{00} = 0.907$	0.0055	$\hat{p}_{00} = 0.909$	0.0054
$N = 4402$		+	−	$\hat{p}_{10} = 0.081$	0.0062	$\hat{p}_{10} = 0.081$	0.0053	$\hat{p}_{10} = 0.080$	0.0051
		−	+	$\hat{p}_{01} = 0.007$	0.0018	$\hat{p}_{01} = 0.006$	0.0013	$\hat{p}_{01} = 0.006$	0.0013
		+	+	$\hat{p}_{11} = 0.005$	0.0013	$\hat{p}_{11} = 0.005$	0.0012	$\hat{p}_{11} = 0.005$	0.0011
	MAP	−	−	$\hat{p}_{00} = 0.908$	0.0065	$\hat{p}_{00} = 0.908$	0.0055	$\hat{p}_{00} = 0.908$	0.0054
		+	−	$\hat{p}_{10} = 0.081$	0.0062	$\hat{p}_{10} = 0.081$	0.0053	$\hat{p}_{10} = 0.081$	0.0051
		−	+	$\hat{p}_{01} = 0.006$	0.0018	$\hat{p}_{01} = 0.006$	0.0013	$\hat{p}_{01} = 0.006$	0.0013
		+	+	$\hat{p}_{11} = 0.005$	0.0013	$\hat{p}_{11} = 0.005$	0.0012	$\hat{p}_{11} = 0.005$	0.0011
	$\overline{T}$			2489.6		2332.9 (6.3%)		2333.3 (6.3%)	
	S_T			23.7		25.4		26.3	
Swab	Mean	−	−	$\hat{p}_{00} = 0.907$	0.0052	$\hat{p}_{00} = 0.908$	0.0039	$\hat{p}_{00} = 0.908$	0.0038
$N = 10048$		+	−	$\hat{p}_{10} = 0.081$	0.0051	$\hat{p}_{10} = 0.081$	0.0038	$\hat{p}_{10} = 0.081$	0.0037
		−	+	$\hat{p}_{01} = 0.006$	0.0011	$\hat{p}_{01} = 0.006$	0.0008	$\hat{p}_{01} = 0.006$	0.0008
		+	+	$\hat{p}_{11} = 0.005$	0.0009	$\hat{p}_{11} = 0.005$	0.0007	$\hat{p}_{11} = 0.005$	0.0007
	MAP	−	−	$\hat{p}_{00} = 0.908$	0.0052	$\hat{p}_{00} = 0.909$	0.0039	$\hat{p}_{00} = 0.909$	0.0038
		+	−	$\hat{p}_{10} = 0.081$	0.0051	$\hat{p}_{10} = 0.081$	0.0038	$\hat{p}_{10} = 0.081$	0.0037
		−	+	$\hat{p}_{01} = 0.005$	0.0011	$\hat{p}_{01} = 0.005$	0.0008	$\hat{p}_{01} = 0.005$	0.0008
		+	+	$\hat{p}_{11} = 0.005$	0.0009	$\hat{p}_{11} = 0.005$	0.0007	$\hat{p}_{11} = 0.005$	0.0007
	$\overline{T}$			5802.8		5400.0 (6.9%)		5354.7 (7.7%)	
	S_T			46.4		47.9		53.8	

Abbreviations: CT, *Chlamydia trachomatis*; MAP, maximum a posteriori; NG, *Neisseria gonorrhoeae*.

we used values of $S_{e:(1)1}$ and $S_{p:(1)1}$ reported in the AC2A package insert for CT and similarly $S_{e:(1)2}$ and $S_{p:(1)2}$ for NG (see www.hologic.com). These values were used only to determine the protocols in Table 4 and to simulate all test responses in our feasibility study. To average over the effect of Monte Carlo simulation error, we created $B = 500$ sets of master pools for each protocol with the configurations in Table 4. Within each specimen type (urine/swab), random assignment of individuals to pools was used throughout.

For each set of master pools, we used simulation to create test outcomes one would observe had H2, H3, and AT been implemented at the Iowa SHL, and we calculated the posterior mean and MAP estimates of $\mathbf{p}$ and $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2})'$ under the assumption $\boldsymbol{\delta}$ is unknown. The disease probabilities in $\mathbf{p}$ in this application are

$$p_{00} = \text{proportion of individuals negative for both CT and NG}$$

$$p_{10} = \text{proportion of individuals positive for CT but negative for NG}$$

$$p_{01} = \text{proportion of individuals negative for CT but positive for NG}$$

$$p_{11} = \text{proportion of individuals positive for both CT and NG.}$$

As in Section 5, we used flat priors for all parameters, that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1, 1)$, and $S_{p:(1)k} \sim \text{beta}(1, 1)$, for $k = 1, 2$. We also continued to use the same starting values $\mathbf{p}^{(0)}$ and $\boldsymbol{\delta}^{(0)}$ as in Section 5 and the same selections for the number of Gibbs iterates and thinning. Trace plots were used to monitor convergence and to check posterior mixing. For one data set (out of 500) in the urine stratum, which includes the results for H2, H3, and AT, the simulation took 70 (271) s to determine the posterior mean estimates (MAP estimates). These same times for the larger swab stratum were 167 and

TABLE 6 CT/NG pooling feasibility study. Posterior estimates of $\boldsymbol{\delta} = (S_{e:(1)1}, S_{e:(1)2}, S_{p:(1)1}, S_{p:(1)2})'$. Estimates ("Est") are averages over $B = 500$ Monte Carlo data sets, and "SE" is the estimated posterior standard deviation as described in Section 5.2. Flat priors have been used for all parameters; that is, $\mathbf{p} \sim \text{Dirichlet}(\mathbf{1}_4)$, $S_{e:(1)k} \sim \text{beta}(1,1)$, and $S_{p:(1)k} \sim \text{beta}(1,1)$, for $k = 1, 2$.

			H2		H3		AT	
Stratum	**Accuracy**		**Est**	**SE**	**Est**	**SE**	**Est**	**SE**
Urine	$S_{e:(1)1} = 0.947$	Mean	$\widehat{S}_{e:(1)1} = 0.948$	0.022	$\widehat{S}_{e:(1)1} = 0.946$	0.012	$\widehat{S}_{e:(1)1} = 0.948$	0.011
$N = 4402$	$S_{e:(1)2} = 0.913$		$\widehat{S}_{e:(1)2} = 0.875$	0.066	$\widehat{S}_{e:(1)2} = 0.899$	0.034	$\widehat{S}_{e:(1)2} = 0.904$	0.032
	$S_{p:(1)1} = 0.989$		$\widehat{S}_{p:(1)1} = 0.987$	0.007	$\widehat{S}_{p:(1)1} = 0.987$	0.006	$\widehat{S}_{p:(1)1} = 0.985$	0.008
	$S_{p:(1)2} = 0.993$		$\widehat{S}_{p:(1)2} = 0.993$	0.002	$\widehat{S}_{p:(1)2} = 0.993$	0.002	$\widehat{S}_{p:(1)2} = 0.993$	0.002
		MAP	$\widehat{S}_{e:(1)1} = 0.948$	0.022	$\widehat{S}_{e:(1)1} = 0.947$	0.012	$\widehat{S}_{e:(1)1} = 0.948$	0.011
			$\widehat{S}_{e:(1)2} = 0.911$	0.066	$\widehat{S}_{e:(1)2} = 0.911$	0.034	$\widehat{S}_{e:(1)2} = 0.914$	0.032
			$\widehat{S}_{p:(1)1} = 0.989$	0.007	$\widehat{S}_{p:(1)1} = 0.989$	0.006	$\widehat{S}_{p:(1)1} = 0.989$	0.008
			$\widehat{S}_{p:(1)2} = 0.993$	0.003	$\widehat{S}_{p:(1)2} = 0.993$	0.002	$\widehat{S}_{p:(1)2} = 0.993$	0.002
Swab	$S_{e:(1)1} = 0.942$	Mean	$\widehat{S}_{e:(1)1} = 0.940$	0.019	$\widehat{S}_{e:(1)1} = 0.941$	0.009	$\widehat{S}_{e:(1)1} = 0.941$	0.008
$N = 10048$	$S_{e:(1)2} = 0.992$		$\widehat{S}_{e:(1)2} = 0.936$	0.040	$\widehat{S}_{e:(1)2} = 0.985$	0.009	$\widehat{S}_{e:(1)2} = 0.986$	0.009
	$S_{p:(1)1} = 0.976$		$\widehat{S}_{p:(1)1} = 0.976$	0.006	$\widehat{S}_{p:(1)1} = 0.976$	0.005	$\widehat{S}_{p:(1)1} = 0.975$	0.007
	$S_{p:(1)2} = 0.987$		$\widehat{S}_{p:(1)2} = 0.988$	0.002	$\widehat{S}_{p:(1)2} = 0.987$	0.002	$\widehat{S}_{p:(1)2} = 0.987$	0.002
		MAP	$\widehat{S}_{e:(1)1} = 0.942$	0.019	$\widehat{S}_{e:(1)1} = 0.942$	0.009	$\widehat{S}_{e:(1)1} = 0.942$	0.008
			$\widehat{S}_{e:(1)2} = 0.984$	0.040	$\widehat{S}_{e:(1)2} = 0.991$	0.009	$\widehat{S}_{e:(1)2} = 0.991$	0.009
			$\widehat{S}_{p:(1)1} = 0.976$	0.006	$\widehat{S}_{p:(1)1} = 0.976$	0.005	$\widehat{S}_{p:(1)1} = 0.976$	0.007
			$\widehat{S}_{p:(1)2} = 0.987$	0.002	$\widehat{S}_{p:(1)2} = 0.987$	0.002	$\widehat{S}_{p:(1)2} = 0.987$	0.002

Abbreviations: CT, *Chlamydia trachomatis*; MAP, maximum a posteriori; NG, *Neisseria gonorrhoeae*.

330 s, respectively. The study was performed on a computer that has an Intel Core i7-10750H CPU @ 2.60 GHz and 32 GB of RAM.

The posterior mean and MAP estimate summaries for $\mathbf{p}$ and $\boldsymbol{\delta}$ are shown in Tables 5 and 6, respectively. The "Est" values shown in the tables are averages over 500 data sets, and we continue to use "SE" as an estimate of the posterior standard deviation. We did not report sample standard deviations of the 500 posterior estimates in this study because, unlike those in Section 5, all Monte Carlo data sets have been constructed from one set of CT/NG diagnoses.

The first observation in Table 5 is that, for either specimen type, disease probability estimates are similar across the three protocols (H2, H3, and AT) for both the posterior mean and mode (MAP). In addition, posterior variability is reduced when moving from H2 to either H3 or AT. For example, when compared to H2, both H3 and AT confer a 6.3% reduction in the average number of tests needed to classify all urine specimens for CT and NG, yet all posterior estimates are slightly more precise. Similar reductions and improvements are seen for the swab specimens. Moving to the assay accuracy probabilities in Table 6, one observes the same phenomenon for $S_{e:(1)1}$ and $S_{e:(1)2}$, that is, moving from H2 to either H3 or AT reduces the posterior variability by roughly 50% for both specimen types. The benefits of using H3 and AT are well known in the case identification literature (see, e.g., Kim et al., 2007) because both protocols provide improved classification efficiency. What our study adds is that the Iowa SHL could switch to these more efficient protocols and not lose anything in terms of estimation accuracy or precision.

7 | DISCUSSION

We have developed estimation techniques for group testing data with two or more diseases, thereby generalizing the approaches in Warasi et al. (2016) and Tebbs et al. (2013) to accommodate data from *any* group testing protocol which uses multiplex assays. When compared to Dorfman testing through H2, our simulation studies demonstrate that estimation performance is not compromised when using higher stage hierarchical or AT protocols, despite the fact these protocols often require fewer tests. We provide R functions in Appendix F in the Supporting Information which describe posterior sampling and enable the practitioner to implement the methods presented in this article. Code with documentation and examples is also available at the first author's GitHub page https://github.com/mswarasi/General-MultiplexBayes.

We conclude with three remarks, all of which present avenues for future work. First, it is always of interest to think about optimal designs in group testing—not only for case identification but also for estimation. We have used protocols in Sections 5 and 6 on the basis of the former because these designs represent those which laboratories can implement to provide diagnoses to all individuals in the most efficient way possible. At the same time, one might also select those designs which provide the best estimation performance. As noted in Section 1, this has been investigated on frequentist grounds when resolving positive pools is not performed. More recently, Warasi et al. (2022) have extended this work to Dorfman testing (H2) when detecting multiple diseases in animal populations. Second, instead of specifying multiple sets of prior distributions for differently sized pools, perhaps to fend off fears of dilution (Warasi et al., 2017), it should be possible to elicit a secondary model which describes how pools might experience the dilution of positive specimens. Modeling assay sensitivities directly, for each disease separately or jointly, could provide a way to relax assumptions in Section 2, and it may provide a more parsimonious approach to estimation. Modifications of our EM algorithm to determine posterior maximizers, such as variational EM approaches, could be useful in the event of increased computational complexity. Finally, an anonymous reviewer has remarked that including covariates which are predictive of disease (e.g., number of sexual partners, race/ethnicity, etc.) may sharpen probability estimates on an individual level. A number of authors have looked at the regression analysis of group testing data for a single disease (see Section 1). Generalizing this work to accommodate outcomes from multiplex group testing is an excellent topic for future research.

ACKNOWLEDGMENTS

We are grateful to the Editor, Associate Editor, and two reviewers for their valuable comments and suggestions on earlier versions of this article. This research was funded by Grant R01 AI121351 from the National Institutes of Health.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Simulated data in Section 5 can be generated with the R code provided in the Supporting Information. The Iowa CT/NG data, in the form of disease counts, are summarized in Appendix E in the Supporting Information.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

ORCID

Md S. Warasi https://orcid.org/0000-0003-1740-4223
Christopher S. McMahan https://orcid.org/0000-0001-5056-9615

REFERENCES

Abdalhamid, B., Bilder, C., McCutchen, E., Hinrichs, S., Koepsell, S., & Iwen, P. (2020). Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *American Journal of Clinical Pathology*, *153*, 715–718.

Bilder, C., Tebbs, J., & McMahan, C. (2021). Discussion on “Is group testing ready for prime-time in disease identification.” *Statistics in Medicine*, *40*, 3881–3886.

Busch, M., Caglioti, S., Robertson, E., McAuley, J., Tobler, L., Kamel, H., Linnen, J., Venkatakrishna, S., Tomasulo, P., & Kleinman, S. (2005). Screening the blood supply for West Nile virus RNA by nucleic acid amplification testing. *New England Journal of Medicine*, *353*, 460–467.

Centers for Disease Control and Prevention. (2022). *Sexually transmitted disease surveillance 2020*. Department of Health and Human Services. Retrieved January 30, 2023, from https://www.cdc.gov/std/statistics/2020/default.htm

Chen, C., & Swallow, W. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*, *46*, 1035–1046.

Delaigle, A., & Meister, A. (2011). Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*, *106*, 640–650.

de Salazar, A., Espadafor, B., Fuentes-Lopez, A., Barrientos-Duran, A., Salvador, L., Alvarez, M., & Garcia, F. (2019). Comparison between Aptima Assays (Hologic) and the Allplex STI Essential Assay (Seegene) for the diagnosis of sexually transmitted infections. *PLoS One*, *14*, e022439.

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

Ding, J., & Xiong, W. (2015). Robust group testing for multiple traits with misclassification. *Journal of Applied Statistics*, *42*, 2115–2125.

Hou, P., Tebbs, J., Bilder, C., & McMahan, C. (2017). Hierarchical group testing for multiple infections. *Biometrics*, *73*, 656–665.

Hou, P., Tebbs, J., Wang, D., McMahan, C., & Bilder, C. (2020). Array testing for multiplex assays. *Biostatistics*, *21*, 417–431.

Hourfar, M., Jork, C., Schottstedt, V., Weber-Schehl, M., Brixner, V., Busch, M., Geusendam, G., Gubbe, K., Mahnhardt, C., Mayr-Wohlfar, W., Pichl, L., Roth, W., Schmidt, M., Seifried, E., & Wright, D. (2008). Experience of German Red Cross blood donor services with nucleic acid testing: Results of screening more than 30 million blood donations for human immunodeficiency virus, hepatitis C virus, and hepatitis B virus. *Transfusion*, *48*, 1558–1566.

Huang, S., Huang, M., Shedden, K., & Wong, W. (2017). Optimal group testing designs for estimating prevalence with uncertain testing errors. *Journal of the Royal Statistical Society: Series B*, *79*, 1547–1563.

Hughes-Oliver, J., & Rosenberger, W. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, *87*, 315–327.

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*, 1152–1163.

Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., & Moncada, J. (2005). Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Virology*, *43*, 1674–1677.

Li, Q., Liu, A., & Xiong, W. (2017). D-optimality of group testing for joint estimation of correlated rare diseases with misclassification. *Statistica Sinica*, *27*, 823–838.

Liu, A., Liu, C., Zhang, Z., & Albert, P. (2012). Optimality of group testing in the presence of misclassification. *Biometrika*, *99*, 245–251.

Martin, E., Salaru, G., Mohammed, D., Coombs, R., Paul, S., & Cadoff, E. (2013). Finding those at risk: Acute HIV infection in Newark, NJ. *Journal of Clinical Virology*, *58*, e24–e28.

McMahan, C., Tebbs, J., Hanson, T., & Bilder, C. (2017). Bayesian regression for group testing data. *Biometrics*, *73*, 1443–1452.

Pilcher, C., Fiscus, S., Nguyen, T., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J., McPherson, J., Stalzer, B., Hightow, L., Miller, W., Eron, J., Cohen, M., & Leone, P. (2005). Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine*, *352*, 1873–1883.

Pilcher, C., Westreich, D., & Hudgens, M. (2020). Group testing for SARS-CoV-2 to enable rapid scale-up of testing and real-time surveillance of incidence. *Journal of Infectious Diseases*, *222*, 903–909.

Quinn, T., Brookmeyer, R., Kline, R., Shepherd, M., Paranjape, R., Mehendale, S., Gadkari, D., & Bollinger, R. (2000). Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. *AIDS*, *14*, 2751–2757.

Roche. (2020). Roche receives FDA emergency use authorization for the cobas SARS-CoV-2 & influenza A/B test for use on the cobas 6800/8800 systems. Retrieved January 30, 2023, from https://www.roche.com/media/releases/med-cor-2020-09-04

Stramer, S., Krysztof, D., Brodsky, J., Fickett, T., Reynolds, B., Dodd, R., & Kleinman, S. (2013). Comparative analysis of triplex nucleic acid test assays in United States blood donors. *Transfusion*, *53*, 2525–2537.

Stramer, S., Notari, E., Krysztof, D., & Dodd, R. (2013). Hepatitis B virus testing by minipool nucleic acid testing: Does it improve blood safety? *Transfusion*, *53*, 2449–2458.

Tebbs, J., McMahan, C., & Bilder, C. (2013). Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics*, *69*, 1064–1073.

Tu, X., Litvak, E., & Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika*, *82*, 287–297.

Wang, D., McMahan, C., Gallagher, C., & Kulasekera, K. (2014). Semiparametric group testing regression models. *Biometrika*, *101*, 587–598.

Warasi, M., Hungerford, L., & Lahmers, K. (2022). Optimizing pooled testing for estimating the prevalence of multiple diseases. *Journal of Agricultural, Biological and Environmental Statistics*, *27*, 713–727.

Warasi, M., McMahan, C., Tebbs, J., & Bilder, C. (2017). Group testing regression models with dilution submodels. *Statistics in Medicine*, *36*, 4860–4872.

Warasi, M., Tebbs, J., McMahan, C., & Bilder, C. (2016). Estimating the prevalence of multiple diseases from two-stage hierarchical pooling. *Statistics in Medicine*, *35*, 3851–3864.

Westreich, D., Hudgens, M., Fiscus, S., & Pilcher, C. (2008). Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. *Journal of Clinical Microbiology*, *46*, 1785–1792.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Warasi, M. S., Tebbs, J. M., McMahan, C. S., & Bilder, C. R. (2023). Estimating the prevalence of two or more diseases using outcomes from multiplex group testing. *Biometrical Journal*, 2200270. https://doi.org/10.1002/bimj.202200270