Received: 4 October 2021 | Revised: 1 July 2022 | Accepted: 5 July 2022

DOI: 10.1002/sim.9532

RESEARCH ARTICLE

Statistics in Medicine WILEY

Capturing the pool dilution effect in group testing regression: A Bayesian approach

Stella Self[1] | **Christopher McMahan**[2] | **Stefani Mokalled**[2]

[1]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA

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

Correspondence
Stella Self, Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Discovery Building, 901 Greene St, Columbia, SC 29208, USA.
Email: scwatson@mailbox.sc.edu

Funding information
National Institutes of Health, Grant/Award Numbers: P20GM130420, R01-AI121351, R25AI164581; National Science Foundation, Grant/Award Number: OIA-1826715; U.S. Department of Defense, Grant/Award Number: N00014-19-1-2295

Group (pooled) testing is becoming a popular strategy for screening large populations for infectious diseases. This popularity is owed to the cost savings that can be realized through implementing group testing methods. These methods involve physically combining biomaterial (eg, saliva, blood, urine) collected on individuals into pooled specimens which are tested for an infection of interest. Through testing these pooled specimens, group testing methods reduce the cost of diagnosing all individuals under study by reducing the number of tests performed. Even though group testing offers substantial cost reductions, some practitioners are hesitant to adopt group testing methods due to the so-called *dilution effect*. The dilution effect describes the phenomenon in which biomaterial from negative individuals dilute the contributions from positive individuals to such a degree that a pool is incorrectly classified. Ignoring the dilution effect can reduce classification accuracy and lead to bias in parameter estimates and inaccurate inference. To circumvent these issues, we propose a Bayesian regression methodology which directly acknowledges the dilution effect while accommodating data that arises from any group testing protocol. As a part of our estimation strategy, we are able to identify pool specific optimal classification thresholds which are aimed at maximizing the classification accuracy of the group testing protocol being implemented. These two features working in concert effectively alleviate the primary concerns raised by practitioners regarding group testing. The performance of our methodology is illustrated via an extensive simulation study and by being applied to Hepatitis B data collected on Irish prisoners.

KEYWORDS
Bayesian models, biomarkers, dilution effect, group testing regression, measurement error

1 | INTRODUCTION

Group (pooled) testing was first proposed by Dorfman[1] as a strategy that could be used to screen United States Army inductees for syphilis during the Second World War. The strategy outlined by this seminal work suggested that pooled specimen be formed, by amalgamating biomaterial (eg, blood, urine) collected from individuals, and tested for the

infection of interest. Based on the outcomes of the pool tests, individuals are either diagnosed as being negative or are subjected to further testing. In particular, as a part of Dorfman's original strategy, individuals contributing to pools that test negative would be diagnosed as such, while positive pools would be resolved by retesting contributing individuals one-by-one. If the infection/disease of interest is relatively rare, it is easy to see that this testing strategy can confer a substantial reduction in testing cost, that is, in such settings a majority of the pools will be negative allowing practitioners to diagnose all contributing individuals at the expense of a single diagnostic test per pool. Given these potential cost savings, group testing has been adopted to screen for a variety of infectious diseases (eg, HIV,[2] Zika,[3] influenza,[4] SARS-CoV-2[5]) as well as in alternate application areas (eg, animal testing,[6,7] bio-terrorism detection,[8] drug discovery,[9] genetics[10]).

In addition to case identification (ie, determining which specific individuals have a disease, prior infection, etc.), group testing has also been posited as a tool that can be used to reduce costs associated with conducting surveillance. This is accomplished by designing statistical methodologies that can estimate population level characteristics based on data arising from implementing a group testing protocol. The origins of the group testing estimation problem can be traced to Thompson[11] and Chiang and Reeves,[12] who independently developed a prevalence estimator based on test outcomes taken solely on pooled specimens. Since this proposal, the prevalence estimation problem has received considerable attention, for example, see Hung and Swallow[13] for a nice review. Extending these works to allow for the inclusion of covariate information, a number of regression procedures have been developed, which include parameteric,[14-16] semi-parameteric,[17,18] and non-parametric[19-21] techniques. A common limitation among the aforementioned methodologies is that they do not account for the *dilution effect*, which, if present and unaccounted for, can lead to bias in parameter estimates and inaccurate inference.

To understand the underpinnings of the dilution effect, one must consider the underlying mechanism by which diagnostic tests classify the infection status of a specimen (pooled or unpooled). Most assays render a binary diagnosis based on the measured concentration of a continuous biomarker (eg, antibody level, antigen concentration) which is indicative of the infection of interest. Thus, a diagnosis is levied based on whether a measured biomarker concentration exceeds a diagnostic threshold, with elevated concentrations typically being indicative of infection. With this in mind, the dilution effect describes the phenomenon by which an assay's sensitivity (ability to classify a truly positive sample as such) is adversely impacted by pooling multiple biospecimens. This impact is due to the biomarker concentration of a positive specimen being diluted when pooled with several negative ones.

To account for the dilution effect, a number of regression methodologies have been developed. McMahan et al[22] was the first to propose such a procedure, though this proposal incorporates data from master pools only. Wang et al[23] expanded on this approach by incorporating data arising from retesting protocols. Other approaches include Delaigle and Hall[24] and Warasi et al.[25] More recently, Mokalled et al[26] developed a regression methodology which acknowledges the dilution effect by assuming that the observed testing outcomes are continuous biomarker concentrations. A primary strength of this work is that the underlying biomarker distributions for the positive and negative individuals are assumed to be unknown and are estimated as part of the regression procedure, thus circumventing restrictive assumption made by previous proposals. However, this approach cannot accommodate data observed from resolving positive pools and it ignores the potential for measurement error.

To overcome these limitations, herein we develop a Bayesian group testing regression methodology which specifically accounts for the dilution effect. Our approach, unlike Mokalled et al,[26] can accommodate testing data arising from any group testing protocol, can easily be implemented under any biomarker distributional assumptions, and directly acknowledges the error associated with measuring the biomarker concentrations. In developing our approach, we consider four commonly encountered settings; namely, (1) the information available on the biomarker distributions is poor quality, (2) there is limited information available, (3) there is a great deal of information available, and (4) the distributions are known. Through analyzing continuous outcomes measured on pools, our approach can estimate both a regression function and the distributions of the biomarker concentrations of positive and negative individuals. Further, in settings where limited/poor information is available for the biomarker distributions, we propose a two-stage procedure under which our proposed modeling framework can be used to set diagnostic thresholds to minimize misclassifications, thus merging the classification and estimation problems. To facilitate model fitting, an easy to implement Markov chain Monte Carlo (MCMC) posterior sampling algorithm is developed. The finite sample performance of our approach is illustrated through in-depth numerical studies and by being applied to Hepatitis B virus (HBV) data collected on Irish prisoners.

The remainder of this article is organized as follows. In Section 2, we develop our Bayesian regression methodology for group testing data. This includes deriving the observed data likelihood, developing the proposed MCMC posterior sampling algorithm, and outlining our two-stage procedure for estimating pool-specific thresholds. Section 3 covers case

identification. In Section 4, we use simulation to assess the performance of both the proposed estimation and case identification aspects of our work. In Section 5, we apply our methods to the HBV data. In Section 6, we conclude with a summary discussion. Additional details are provided in the Web Appendix.

2 | METHODOLOGY

2.1 | The model

Suppose we are screening N individuals for a disease/infection of interest via a group testing protocol. Let Y_i denote the true infection status of the ith individual, for $i = 1, \ldots, N$, with the convention that $Y_i = 1$ indicates that the individual is infected and $Y_i = 0$ otherwise. We assume that an individual's true disease status can be associated with a set of Q individual-level covariates, denoted $\boldsymbol{x}_i$, through the following model

$$Y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli}\{g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})\}, \quad \text{for } i = 1, \ldots, N, \tag{1}$$

where $\boldsymbol{\beta}$ is a Q-dimensional vector of regression coefficients and $g(\cdot)$ is a known binary link function, for example, logistic or probit. As is common in the literature, we assume that the individuals' statuses (ie, the Y_i's) are conditionally independent given the covariate information. Moreover, it is important to note that in the presence of imperfect testing the Y_i's are unobservable, even under individual level testing.

In our proposed modeling framework, we relate the individuals' true statuses to the outcomes measured on pools through their true biomarker concentrations. To this end, let ζ_i denote the true biomarker concentration of the ith individual, and we assume that these variables are conditionally (given Y_i) distributed as

$$\zeta_i|Y_i, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \sim (1 - Y_i)f_{\zeta^-}(\zeta|\boldsymbol{\theta}_0) + Y_i f_{\zeta^+}(\zeta|\boldsymbol{\theta}_1), \tag{2}$$

where $f_{\zeta^-}(\cdot|\boldsymbol{\theta}_0)$ and $f_{\zeta^+}(\cdot|\boldsymbol{\theta}_1)$ are the probability density functions of the biomarker concentrations of the negative and positive individuals, respectively, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)'$ is a vector of parameters governing these distributions. A few comments are warranted. First, given the individuals' true statuses, we assume that the ζ_i are conditionally independent of each other and the covariates. Second, as with the individuals' true statuses, in the presence of imperfect testing the ζ_i's are unobservable. Lastly, in some settings it may be reasonable to assume $\boldsymbol{\theta}$ is known, while in others it may be unknown. When $\boldsymbol{\theta}$ is unknown, some information (of various quality and quantity) might nevertheless be available. Our method facilitates the inclusion of such information if it exists.

To develop a general methodology, we note that many group testing protocols have been proposed for classification[1,27-29] and/or quality control purposes.[30,31] Most of these protocols require a number of the individuals under study to be tested in multiple pools. To track pool membership, we introduce the index set $\mathcal{P}_j$ which identifies the individuals who contributed to the jth pool, for $j = 1, \ldots, J$, that is, $i \in \mathcal{P}_j$ if and only if individual i contributes biomaterial to pool j. Herein, unlike previous proposals, we assume that the observed data collected from assaying the jth pool consists of its measured biomarker concentration, which we denote by C_j, for $j = 1, \ldots, J$. To relate the individuals' true biomarker concentrations to those measured on pools, we assume that the (true) biomarker concentration of the jth pool, $\zeta_{\mathcal{P}_j}$, is the arithmetic average of the concentrations of the contributing individuals, that is, $\zeta_{\mathcal{P}_j} = |\mathcal{P}_j|^{-1} \sum_{i \in \mathcal{P}_j} \zeta_i$. This assumption is common among the literature[23,26] and is reasonable as long as pools are formed from equal volume aliquots. To relate these two variables, we assume the following classical measurement error model

$$C_j|\boldsymbol{\zeta}, \tau^2 \overset{ind}{\sim} \text{Normal}(\zeta_{\mathcal{P}_j}, \tau^2) \quad \text{for } j = 1, \ldots, J,$$

where $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_N)'$, and $\tau^2 > 0$ is an unknown error variance. A few comments are warranted. First, it would be relatively easy to allow for other types of measurement error. For example, one could allow the magnitude of the error variance to depend on the biomarker concentration being measured by assuming that

$$C_j|\boldsymbol{\zeta}, \tau^2 \overset{ind}{\sim} \text{Normal}(\zeta_{\mathcal{P}_j}, \zeta_{\mathcal{P}_j}\tau^2) \quad \text{for } j = 1, \ldots, J.$$

Conceptually, this would be easy to incorporate into our approach, however for simplicity, we restrict our attention to constant error variance scenario. Second, given that a number of the individuals may contribute to multiple pools, the C_j's are generally not independent. Although, under our assumed measurement error model, the C_j's are conditionally independent given the true biomarker concentrations of the pools (or equivalently $\boldsymbol{\zeta}$).

This observation forms the crux of our model fitting strategy. In particular, by exploiting the conditional independence, one can express the conditional distribution of the observed data as:

$$\begin{aligned}\pi(\boldsymbol{C}|\boldsymbol{\delta}) &= \int \pi(\boldsymbol{C}|\boldsymbol{\zeta},\boldsymbol{\delta})\pi(\boldsymbol{\zeta}|\boldsymbol{\delta})d\boldsymbol{\zeta} \\ &= \int \pi(\boldsymbol{C}|\boldsymbol{\zeta},\boldsymbol{\delta})\sum_{\boldsymbol{Y}\in\mathcal{Y}}\pi(\boldsymbol{\zeta}|\boldsymbol{Y},\boldsymbol{\delta})\pi(\boldsymbol{Y}|\boldsymbol{\delta})d\boldsymbol{\zeta} \\ &= \int \prod_{j=1}^{J} f(C_j|\zeta_{\mathcal{P}_j},\tau^2)\sum_{\boldsymbol{Y}\in\mathcal{Y}}\prod_{i=1}^{N} f_{\zeta^-}(\zeta_i|\boldsymbol{\theta}_0)^{1-Y_i} f_{\zeta^+}(\zeta_i|\boldsymbol{\theta}_1)^{Y_i} g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})^{Y_i}\{1-g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})\}^{1-Y_i} d\boldsymbol{\zeta},\end{aligned}$$

where $\boldsymbol{C} = (C_1, C_2, \ldots, C_J)'$, $\boldsymbol{\delta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \tau^2)'$, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)'$, $\mathcal{Y} = \{\boldsymbol{y} = (y_1, y_2, \ldots, y_N)' : y_i \in \{0,1\}\}$, and $f(\cdot|a,b)$ is the probability density function of a normal random variable with mean a and variance b. Given the observed data model, we can complete our Bayesian model by assigning prior distributions for the model parameters. To this end, we specify normal and inverse gamma priors for $\boldsymbol{\beta}$ and τ^2, respectively, that is, $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\beta)$ and $\tau^2 \sim IG(\alpha_\tau, \beta_\tau)$. In practice, the hyperparameters of these priors are selected so that they are diffuse,[32] though if the magnitude of measurement error is well-understood for the laboratory procedure under consideration, an informative prior distribution could be used; for more information on how this can be accomplished see Klauenberg et al.[33] The prior specifications for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are intrinsically tied to the assumed distributional families of f_{ζ^-} and f_{ζ^+}. To avoid loss of generality, we will denote these prior distributions by $\pi(\boldsymbol{\theta}_0)$ and $\pi(\boldsymbol{\theta}_1)$ while leaving their particular form unspecified. It is important to note that the model in (2) hierarchically represents a mixture model and in therefore subject to the "label switching" problem, where the rolls of f_{ζ^-} and f_{ζ^+} can be reversed. This issue is common in mixture models[34,35] and is typically resolved by applying a relabeling algorithm,[35-37] imposing constraints on the parameters,[38] or by assigning informative prior distributions.[39,40] Herein, we adopt the last strategy, and note that the role of the informative priors is primarily to discourage the label switching problem by diminishing the interchangeability of the biomarker distributions. However when high quality biomarker information is available, the informative priors also provide a means of incorporating this information into the model. As we show in our simulation study and data application, these informative prior distributions do not need to be *correctly* specified in order for our method to perform well. In fact, our method performs well in our data application even when $\pi(\boldsymbol{\theta}_0)$ and $\pi(\boldsymbol{\theta}_1)$ are egregiously misspecified and strongly informative. Web Appendix A provides further discussion of other strategies for preventing or resolving the label switching problem, as well as guidance for selecting the most appropriate strategy for particular applications.

2.2 | Estimation

While the observed data likelihood simplifies considerably under certain group testing protocols, obtaining a "closed-form" simplified expression for the general case can prove to be quite cumbersome, if at all possible. Moreover, evaluating the likelihood via numerical integration is computationally impractical since it would require computing an N-dimensional integral whose integrand includes a 2^N-dimensional sum. To circumvent these difficulties, we propose a two-stage data augmentation approach which begins by introducing $\boldsymbol{\zeta}$ and $\boldsymbol{Y}$ as latent random variables. Proceeding in this fashion yields the following augmented data likelihood

$$\pi(\boldsymbol{C},\boldsymbol{\zeta},\boldsymbol{Y}|\boldsymbol{\delta}) = \prod_{j=1}^{J} f(C_j|\zeta_{\mathcal{P}_j},\tau^2)\prod_{i=1}^{N} f_{\zeta^-}(\zeta_i|\boldsymbol{\theta}_0)^{1-Y_i} f_{\zeta^+}(\zeta_i|\boldsymbol{\theta}_1)^{Y_i} g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})^{Y_i}\{1-g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})\}^{1-Y_i}. \quad (3)$$

The next stage in our data augmentation strategy introduces $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots \psi_N)'$ as a means to decompose the binary regression model, thus making the posterior sampling of $\boldsymbol{\beta}$ straightforward. The distribution that ψ_i obeys is specifically tied to the specified link function. Herein, we focus on the case in which $g(\cdot)$ is either the probit or logistic link, which leads to ψ_i being specified as a truncated normal or Pólya-Gamma random variable, respectively; for further details see

Albert and Chib[41] and Polson et al.[42] Under either link function, the augmented data likelihood after the second stage can be expressed as

$$\pi(\boldsymbol{C},\boldsymbol{\zeta},\boldsymbol{Y},\boldsymbol{\psi}|\boldsymbol{\delta}) = \prod_{j=1}^{J} f(C_j|\zeta_{\mathcal{P}_j},\tau^2)\prod_{i=1}^{N} f_{\zeta^-}(\zeta_i|\boldsymbol{\theta}_0)^{1-Y_i} f_{\zeta^+}(\zeta_i|\boldsymbol{\theta}_1)^{Y_i} f(h_i|\boldsymbol{x}_i\boldsymbol{\beta},\omega_i) h(\psi_i), \tag{4}$$

where $h_i = \psi_i$, $\omega_i = 1$, and $h(\psi_i) = I(\psi_i > 0, Y_i = 1) + I(\psi_i < 0, Y_i = 0)$ under the probit link, while under the logistic link, $h_i = \kappa_i/\psi_i$, $h(\psi_i) = \exp\{\kappa_i^2/(2\psi_i)\}\varphi(\psi_i)$, $\kappa_i = Y_i - 0.5$, $\omega_i = \psi_i^{-1}$, and $\varphi(\cdot)$ denotes the density function of a Pólya-Gamma(1,0) random variable; for further details see Polson et al.[42]

To develop our posterior sampling algorithm, we note that based on the forms provided in (3) and (4), it is easy to identify the full conditional distributions of several parameters. In particular, we have that

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\psi} &\sim N(\boldsymbol{\mu}_\beta^*, \boldsymbol{\Sigma}_\beta^*),\\
\tau^2|\boldsymbol{C},\boldsymbol{\zeta} &\sim IG\left(\alpha_\tau^*, \beta_\tau^*\right),\\
Y_i|\zeta_i,\boldsymbol{\beta},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1 &\sim Bernoulli\left(p_i^*\right),
\end{aligned}$$

where the specific forms of $\boldsymbol{\mu}_\beta^*$, $\boldsymbol{\Sigma}_\beta^*$, α_τ^*, β_τ^*, and p_i^* are provided in Web Appendix B. Further, when $g(\cdot)$ is taken to be the probit or logistic link we have that the full conditional distribution of ψ_i is

$$\begin{aligned}
\psi_i|\boldsymbol{\beta},Y_i &\sim \text{TN}\{\boldsymbol{x}_i\boldsymbol{\beta},1,\mathcal{S}(Y_i)\},\\
&\text{or}\\
\psi_i|\boldsymbol{\beta} &\sim \text{PG}(1,\boldsymbol{x}_i\boldsymbol{\beta}),
\end{aligned}$$

respectively, where $\mathcal{S}(\cdot)$ controls the support of the truncated normal distribution, with $\mathcal{S}(0) = (-\infty, 0)$ and $\mathcal{S}(1) = (0,\infty)$, and $PG(a,b)$ denotes the Pólya-Gamma distribution with parameters a and b.

We now turn attention toward the remaining parameters, namely ζ_i, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$. The full conditional distributions of these variables are given by

$$\begin{aligned}
\pi(\zeta_i|\boldsymbol{C},\boldsymbol{\zeta},\boldsymbol{Y},\boldsymbol{\delta}) &\propto \prod_{j\in\mathcal{A}_i} f(C_j|\zeta_{\mathcal{P}_j},\tau^2) f_{\zeta^-}(\zeta_i|\boldsymbol{\theta}_0)^{1-Y_i} f_{\zeta^+}(\zeta_i|\boldsymbol{\theta}_1)^{Y_i},\\
\pi(\boldsymbol{\theta}_0|\boldsymbol{\zeta},\boldsymbol{Y}) &\propto \prod_{i=1}^{N} f_{\zeta^-}(\zeta_i|\boldsymbol{\theta}_0)^{1-Y_i}\pi(\boldsymbol{\theta}_0),\\
\pi(\boldsymbol{\theta}_1|\boldsymbol{\zeta},\boldsymbol{Y}) &\propto \prod_{i=1}^{N} f_{\zeta^+}(\zeta_i|\boldsymbol{\theta}_1)^{Y_i}\pi(\boldsymbol{\theta}_1),
\end{aligned}$$

where $\mathcal{A}_i = \{j : i \in \mathcal{P}_j\}$. Regretfully, under common biomarker distributional assumptions (eg, gamma, log-normal) these full conditionals are not recognizable as a member of a common family. For this reason, and generality, we make use of a Metropolis-Hastings (MH) algorithm to sample these terms. Thus, the proposed Markov chain Monte Carlo (MCMC) sampling algorithm consists of a Metropolis-Hastings-within-Gibbs sampling scheme. For a detailed implementation of our posterior sampling algorithm, see Web Figures 1-4. Note, in implementing the proposed approach, if the biomarker distributions are known with certainty, then $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ can be treated as fixed constants instead of being sampled from their posterior distributions. For example, if high quality data is available on a large number of positive and negative individuals, then $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ can be estimated from this data a priori and held constant during the model fitting process.

3 | CASE IDENTIFICATION

The dilution effect can also adversely impact the classification accuracy of group testing protocols. This effect can be mitigated by setting diagnostic thresholds that acknowledge the dilution effect based on *a priori* knowledge of the biomarker distributions; for further discussion see Wang et al.[43] In practice, the information available for the biomarker distributions

can be of varying quality and quantity. In the context of a rare or newly emerging infectious disease, the available information might be limited to a handful of measurements on positive and negative individuals. Even in more established diseases, if the causative pathogen mutates rapidly, available information may quickly become outdated. Alternatively, if the disease exhibits heterogeneity among different populations, information available from one population may not generalize well to other populations. In an effort to guard against the dilution effect and to improve classification accuracy of group testing strategies in settings where limited/poor information is available for the biomarker distributions, we propose a two-stage procedure which leverages testing information to guide case identification. In the first stage, our estimation methodology is used to estimate "optimal" pool specific diagnostic thresholds which acknowledge the dilution effect. Based on these diagnostic thresholds, retesting is completed via a group testing protocol. In the second stage, we make use of the retesting information to refine the estimated thresholds with an eye toward identifying pools (or individuals) who were potentially misclassified in the first stage. It is important to note that we consider the setting in which limited information is available about the biomarker distributions and researchers do not feel comfortable specifying these distributions exactly, that is, there is considerable uncertainty regarding f_{ζ^+} and f_{ζ^-}. If these distributions were known *a priori* and researchers were interested in classification only, the work of Wang et al could be used to set the diagnostic thresholds.[43]

3.1 | Optimal threshold selection

In what follows, we seek to identify the optimal diagnostic threshold for a pool consisting of c individuals, which we denote by $t^{\triangledown}(c)$. If the biomarker distributions were known, one can identify $t^{\triangledown}(c)$ as

$$t^{\triangledown}(c) = \underset{t}{\operatorname{argmax}}\{S_p(c,t) + S_e(c,t) - 1\},$$

where $S_p(c,t)$ is the probability that a pool consisting of c negative individuals will test negative under a diagnostic threshold t and $S_e(c,t)$ is the probability that a pool consisting of 1 positive and $c-1$ negative individuals will test positive under a diagnostic threshold t. Formally, under the model formulation discussed above, we have that

$$S_p(c,t) = \int_{-\infty}^{t}\int_{-\infty}^{\infty} cf(u|v,\tau^2) f_{\zeta}^{c(0)}(cv)dvdu,$$
$$S_e(c,t) = \int_{t}^{\infty}\int_{-\infty}^{\infty} cf(u|v,\tau^2) f_{\zeta}^{c(1)}(cv)dvdu,$$

where $f_{\zeta}^{c(q)} = f_{\zeta^-}^{(c-q)*} * f_{\zeta^+}^{(q)*}$, "$*$" denotes the usual convolution operator, and $f_{\zeta}^{(q)*}$ denotes the q-fold convolution of f_{ζ} with itself. For further discussion on the derivation of these expressions, see McMahan et al.[22] A few comments are warranted. First, the objective function used to identify the thresholds was inspired by the Youden index, which is commonly adopted for setting diagnostic thresholds for individual level testing. Second, as a strategy for setting thresholds, this approach has been well explored by Wang et al.[43] Lastly, computing the thresholds as discussed above requires one to know f_{ζ^+} and f_{ζ^-}, or equivalently $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$.

In settings where such knowledge about the biomarker distributions is unavailable, one can estimate the thresholds described above by replacing the unknown parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ by their estimates. To this end, we assume that we have access to $\mathcal{C}_j$, for $j = 1, \ldots, J$, arising from the first stage of a group testing protocol, for example, the biomarker concentrations measured on master pools as a part of the first stage of Dorfman testing or row and column pool results from implementing array testing. Based on these assessments, the estimation methodology described above can be utilized to estimate $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ and hence the optimal thresholds. Admittedly, especially in high volume settings, it is expected that the estimates of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ will stabilize after enough data is collected and analyzed by the proposed approach. After this has occurred, one could use the method of Wang et al to set diagnostic thresholds treating the biomarker distributions as known quantities.

3.2 | Quality control stage

Once the thresholds have been determined by the approach outlined above, the group testing protocol can be completed, that is, initially tested pools can be classified and retesting can be conducted as necessary. As a part of the retesting

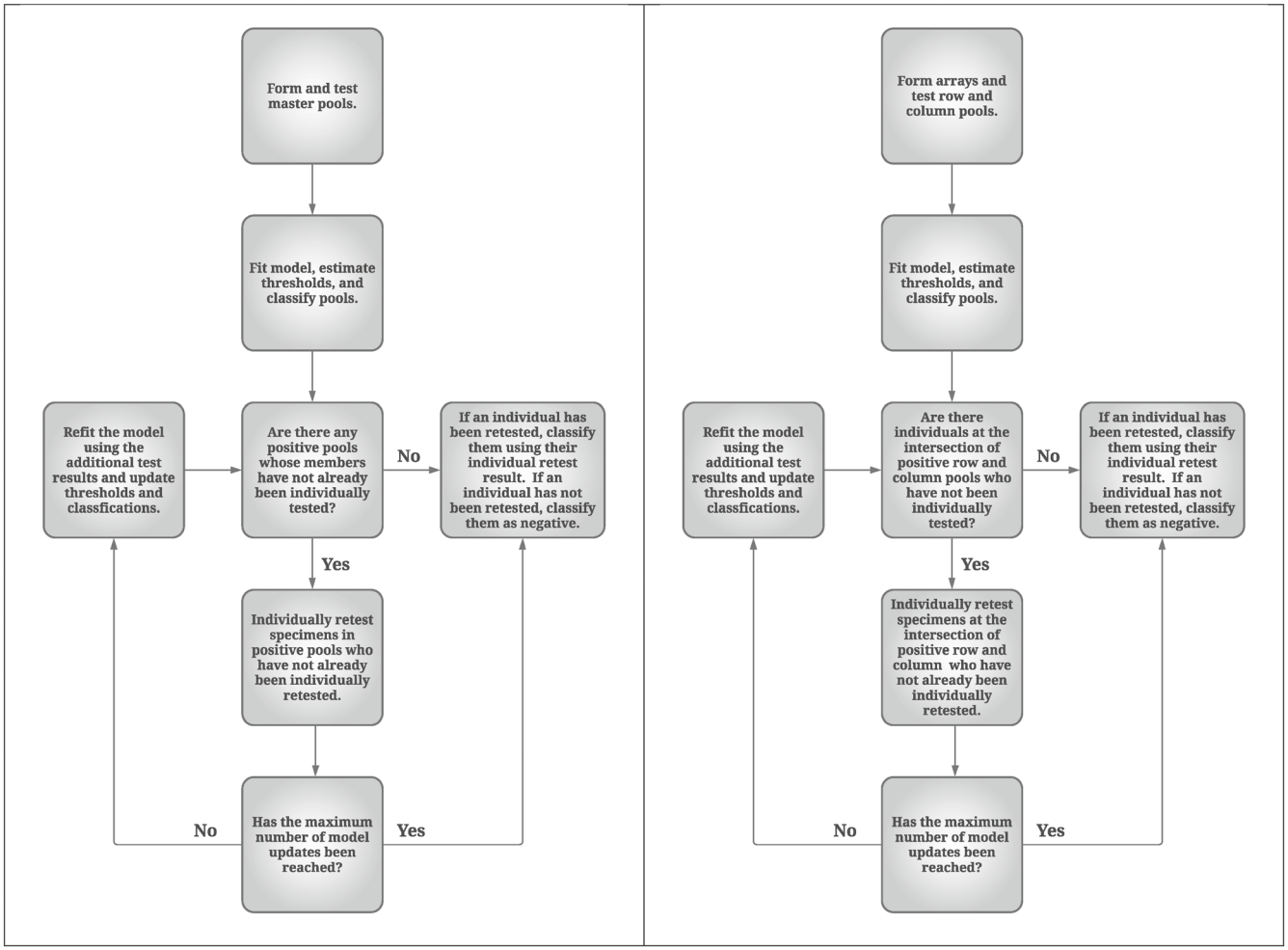

FIGURE 1 This figure presents a flowchart that outlines our quality control adapted variants of Dorfman (left panel) and array (right panel) testing. These adaptations require iterating between testing, classification, and retesting. At each iteration, estimation results are used to update classification thresholds

process, we gain more information by resolving positive pools which can be assimilated into our model to refine our understanding about the unknown model parameters. Thus, as a part of the second stage of our proposed procedure, we assert that our model should be re-fit to this extended dataset and that the diagnostic thresholds be re-estimated. Once this process is complete, one can use the updated diagnostic thresholds to identify discrepancies, for example, pools/individuals initially diagnosed to be negative that are re-classified as positive. In some cases, this could require that additional pools be resolved. Given the numerous group testing protocols that have been proposed, it is hard to enumerate how all possible discrepancies could arise and whether they would necessitate further retesting. That is to say, the retesting process would have to be protocol specific. Figure 1 provides flowcharts depicting the implementation of two such strategies under Dorfman and array testing. This cycle of retesting and re-estimation could be allowed to continue until no discrepancies remain, or until a pre-specified number of updates have been performed.

4 | SIMULATION STUDY

4.1 | Simulation configuration

To demonstrate the performance of the proposed methodology, we conducted an extensive simulation study. As a part of this study, we generated true disease statuses for N individuals, for $N \in \{900, 1800\}$, from the following population-level

models

$$\text{M1: } P(Y_i = 1|\boldsymbol{x}_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}); \boldsymbol{x}_i = (x_{i1}, x_{i2})', \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (-5, 2, 1)',$$
$$\text{M2: } P(Y_i = 1|\boldsymbol{x}_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}); \boldsymbol{x}_i = (x_{i1}, x_{i2})', \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (-3, 0.5, 1.5)',$$

where $g(\cdot)$ is the logistic link, $x_{i1} \sim N(0, 1)$, and $x_{i2} \sim$Bernoulli(0.5). Models M1 and M2 yield overall disease prevalences of 5% and 12%, respectively. To simulate biomarker concentrations for the positive and negative individuals, we consider 2 separate specifications:

$$\text{D1: } \zeta_i|Y_i = y \sim \text{Gamma}(\alpha_y, \phi_y); \boldsymbol{\delta} = (\alpha_0, \gamma_0, \alpha_1, \gamma_1, \boldsymbol{\beta}')' = (2.5, 0.5, 80, 2, \boldsymbol{\beta}')',$$
$$\text{D2: } \zeta_i|Y_i = y \sim \text{Gamma}(\alpha_y, \phi_y); \boldsymbol{\delta} = (\alpha_0, \gamma_0, \alpha_1, \gamma_1, \boldsymbol{\beta}')' = (2.5, 0.5, 20, 1, \boldsymbol{\beta}')',$$

where the specification in D1 allows for near perfect separation, while D2 allows for overlap between the 2 distributions; see Figure 2.

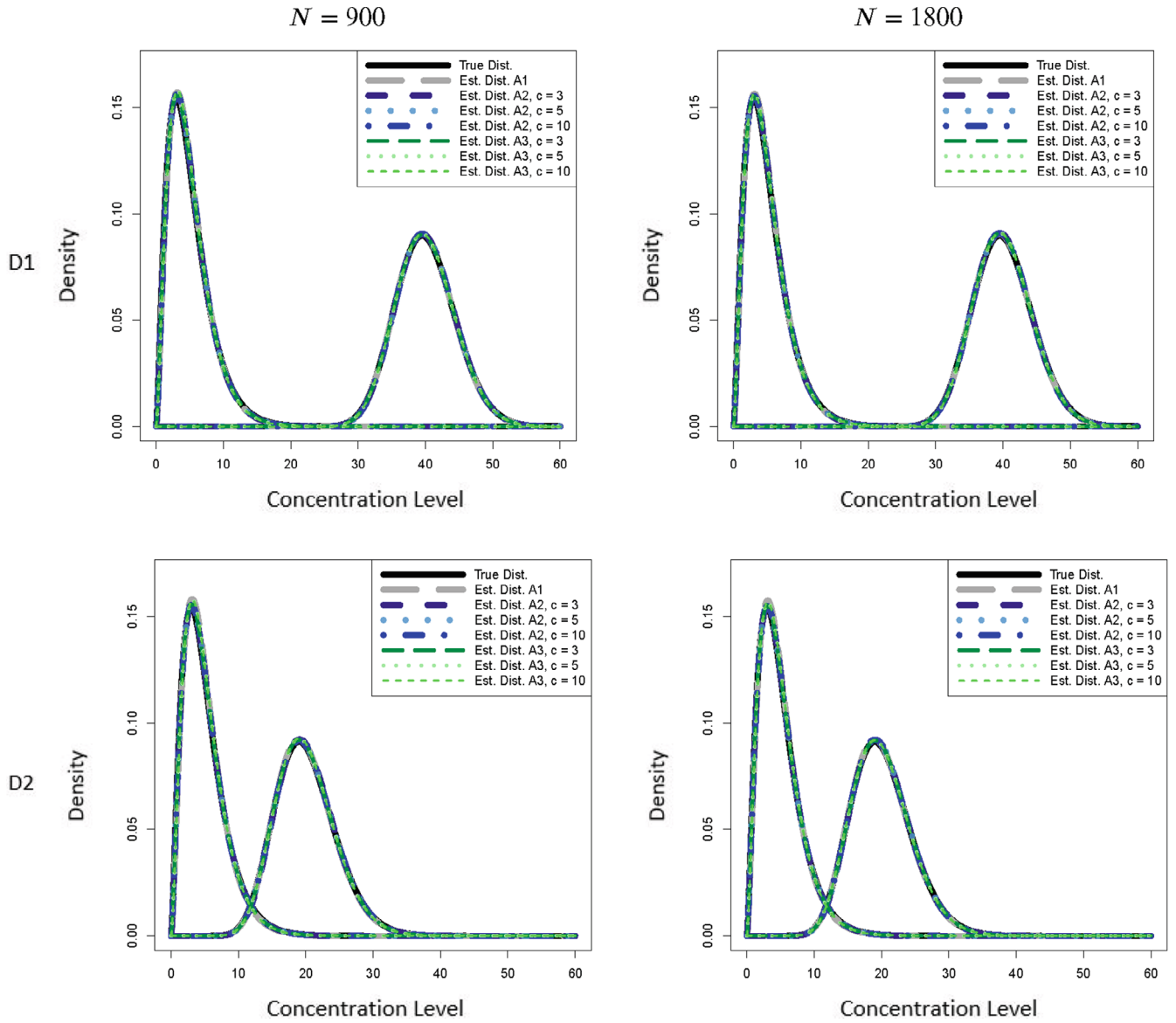

FIGURE 2 Simulation results: Summary of the posterior mean estimates of $\hat{\boldsymbol{\theta}}$ from model M1, under biomarker distributions D1 (top row) and D2 (bottom row) obtained from individual (A1), Dorfman (A2), and array (A3) testing. The curves represent the average estimate of the biomarker distributions, plotted against the true densities. The left column corresponds to $N = 900$ and the right to $N = 1800$

To illustrate our methodology, we simulate the process of testing the N individuals using our quality control adapted variants of 2 common group testing procedures, that is, Dorfman and array testing, with one round of retesting and one associated model update. The step-by-step implementation of these protocols are outlined in Figure 1. In this study, we randomly assigned individuals to pools of size c for Dorfman testing and $c \times c$ arrays for array testing, where $c \in \{3, 5, 10\}$. To provide a baseline for comparison, we also simulated individual level testing. For all three testing protocols the observed biomarker concentration for pools and individuals were simulated as $C_j = |P_j|^{-1} \sum_{i \in P_j} \zeta_i + \epsilon_j$, where $\epsilon_j \sim N(0, \tau^2 = 0.005^2)$ provides the measurement error. This process was repeated 500 times for each combination of sample size, population model, biomarker model, and testing protocol leading to a total of 28 000 datasets.

To complete our Bayesian model, prior distributions for $\boldsymbol{\beta}$ and τ^2 were assigned as described in Section 2 with $\boldsymbol{\Sigma}_\beta = 100\boldsymbol{I}$ and $\alpha_\tau = \beta_\tau = 3$. The biomarker distribution parameters were assigned independent gamma priors whose parameters were specified according to the strategy outlined in Web Appendix A, reflecting the more difficult scenario in which limited and imperfect information is available about the biomarker distributions. We found that this approach all but eliminated the label switching problem, with only 0.02% of model fits exhibiting the problem. The results from these datasets were removed and replaced. To analyze each dataset, a posterior sample of 10 000 realizations was generated using the algorithm outlined in Section 2, after discarding a burn-in sample. Most scenarios required a burn-in period of 20 000 iterations with a few of the individual testing and Dorfman testing with pools of size 10 scenarios requiring 50 000. Convergence of the MCMC chains were assessed in the usual manner (eg, trace plots). Based on the posterior sample, we obtain point estimates (estimated posterior means) of the model parameter and associated measures of uncertainty (estimated posterior SD). Further, based on the estimated diagnostic thresholds, we also classify each individual.

4.2 | Simulation results

Table 1 provides a summary of our estimates of $\boldsymbol{\beta}$ under population model M1 and biomarker concentration model D2. This combination represents the most difficult estimation setting considered, that is, M1 provides for the lowest prevalence and D2 provides for overlapping biomarker distributions. Web Tables 1-3 provide the same summary under the other considered simulation configurations. The presented summary includes the empirical bias, average estimated posterior SD, and the SD of our estimators, along with the empirical coverage probability associated with 95% credible intervals. From these results, one will first note that the proposed approach provides both accurate point estimates as well as reliable inference. That is, using the results from individual level testing as a baseline for comparison, we first note the bias in point estimates are relatively small. Moreover, this bias tends to disappear as the sample size increases, as one should expect. Further, the variability of the estimates obtained by the group testing procedures are roughly equivalent to those attained from individual level testing and the coverage probabilities attain their nominal level. In making these comparisons, it is important to remember that it takes approximately twice as many tests to collect the individual level data than the group testing data; see the average number of tests reported in the right hand column of Table 1. Attention is now turned to classification accuracy, Table 1 also provides the empirical true positive, true negative, false positive, and false negative classification rates that were obtained based on our estimated diagnostic thresholds. For comparative purposes, the same accuracy measures are provided for the case in which the true diagnostic threshold was known. From these results, we see that our quality control step provides near "oracle" like performance, that is, our approach, which has to estimate the diagnostic threshold, classifies the individuals with the same level of precision as the approach that is given the true diagnostic threshold. This is made possible by the fact that our approach is capable of precisely estimating the biomarker distributions of the positives and negatives, which is demonstrated by Figure 2. That is, this figure displays a summary of the estimated biomarker distributions for all simulation configurations under model M1. Similar results under model M2 are provided in Web Figure 5. In summary, the findings from this simulation study suggests that the proposed approach can simultaneously estimate the regression model and the biomarker distributions, as well as provide a path for precise classification, all while directly accounting for the dilution effect and measurement error.

To further explore the performance of the proposed approach, several complementary simulations studies were performed. In particular, we consider the performance of our methodology under several different biomarker distributional settings. These include normal and log-normal specifications. Further, we also examine the case in which the biomarker distribution for the negative individuals is known and concentrates around zero, which is indicative of the absence of the biomarker for negative individuals. To consider other group testing protocols, we also simulated testing under rectangular arrays and array testing with master pool testing. For further details on these additional studies and a summary of the results see Web Appendix C. Briefly, the findings from these studies reinforce all of the conclusions discussed above. That

TABLE 1 Simulation results: The table provides the bias of the posterior mean estimate, average estimated posterior SD, empirical coverage probability of 95% credible intervals (CP), and sample standard deviation (SSD) of the estimated regression coefficients from regression model M1 with biomarker model D2 obtained from individual (A1), Dorfman (A2), and array (A3) testing. The testing accuracy using the estimated optimal threshold (ET) is reported in the form of true negative (TN), true positive (TP), false negative (FN), and false positive (FP) rates. The testing accuracy using the true optimal threshold (TT) is included for comparison. The average number of tests used (#) is also reported

				Estimation				Classification				
N	***A***	***c***		β_0	β_1	β_2		**TN**	**TP**	**FN**	**FP**	**#**
900	A1		Bias (SD)	0.22(0.65)	0.09(0.32)	0.07(0.46)	ET	0.96	0.98	0.02	0.04	900.00
			CP95(SSD)	0.94(0.69)	0.94(0.33)	0.94(0.50)	TT	0.96	0.99	0.01	0.04	
	A2	3	Bias (SD)	−0.37(0.72)	0.15(0.35)	0.11(0.50)	ET	0.97	0.93	0.07	0.03	519.16
			CP95(SSD)	0.94(0.75)	0.94(0.38)	0.94(0.53)	TT	0.97	0.93	0.07	0.03	
		5	Bias (SD)	−0.38(0.74)	0.13(0.36)	0.14(0.53)	ET	0.97	0.89	0.11	0.03	494.96
			CP95(SSD)	0.94(0.88)	0.95(0.37)	0.93(0.66)	TT	0.97	0.89	0.11	0.03	
		10	Bias (SD)	−0.43(0.80)	0.17(0.39)	0.15(0.56)	ET	0.97	0.86	0.14	0.03	547.86
			CP95(SSD)	0.93(0.90)	0.93(0.41)	0.96(0.61)	TT	0.97	0.86	0.14	0.03	
	A3	3	Bias (SD)	−0.34(0.69)	0.14(0.34)	0.11(0.48)	ET	0.98	0.89	0.11	0.02	690.82
			CP95(SSD)	0.91(0.84)	0.92(0.41)	0.94(0.52)	TT	0.98	0.89	0.11	0.02	
		5	Bias (SD)	−0.31(0.71)	0.12(0.36)	0.11(0.49)	ET	0.98	0.82	0.18	0.02	495.16
			CP95(SSD)	0.94(0.76)	0.94(0.39)	0.94(0.51)	TT	0.98	0.84	0.16	0.02	
		10	Bias (SD)	−0.55(0.84)	0.20(0.40)	0.20(0.58)	ET	0.98	0.76	0.24	0.02	417.58
			CP95(SSD)	0.90(1.00)	0.92(0.47)	0.93(0.69)	TT	0.98	0.76	0.24	0.02	
1800	A1		Bias (SD)	−0.01(0.41)	0.02(0.21)	−0.01(0.30)	ET	0.96	0.98	0.02	0.04	1800.00
			CP95(SSD)	0.95(0.43)	0.95(0.21)	0.95(0.31)	TT	0.96	0.99	0.01	0.04	
	A2	3	Bias (SD)	−0.17(0.45)	0.09(0.23)	0.04(0.32)	ET	0.97	0.93	0.07	0.03	1038.26
			CP95(SSD)	0.94(0.45)	0.94(0.23)	0.97(0.31)	TT	0.97	0.93	0.07	0.03	
		5	Bias (SD)	−0.16(0.46)	0.07(0.24)	0.04(0.33)	ET	0.97	0.89	0.11	0.03	981.2
			CP95(SSD)	0.95(0.46)	0.95(0.23)	0.94(0.35)	TT	0.97	0.89	0.11	0.03	
		10	Bias (SD)	−0.17(0.48)	0.07(0.25)	0.05(0.35)	ET	0.97	0.86	0.14	0.03	1084.78
			CP95(SSD)	0.95(0.49)	0.95(0.25)	0.96(0.34)	TT	0.97	0.86	0.14	0.03	
	A3	3	Bias (SD)	−0.12(0.44)	0.05(0.23)	0.02(0.32)	ET	0.98	0.89	0.11	0.02	1379.68
			CP95(SSD)	0.94(0.46)	0.95(0.22)	0.94(0.32)	TT	0.98	0.89	0.11	0.02	
		5	Bias (SD)	−0.09(0.45)	0.04(0.23)	0.02(0.32)	ET	0.98	0.82	0.18	0.02	989.47
			CP95(SSD)	0.94(0.48)	0.94(0.24)	0.94(0.33)	TT	0.98	0.84	0.16	0.02	
		10	Bias (SD)	−0.16(0.48)	0.05(0.25)	0.07(0.35)	ET	0.98	0.77	0.23	0.02	837.33
			CP95(SSD)	0.95(0.52)	0.94(0.26)	0.96(0.36)	TT	0.98	0.76	0.24	0.02	

Abbreviations: A, retesting protocol; c, pool size; N, sample size.

is, that the proposed methodology can simultaneously estimate the regression model and the biomarker distributions while directly accounting for the dilution effect and measurement error.

5 | DATA APPLICATION

To further assess the performance of our methodology, we make use of a Hepatitis B dataset collected on an Irish prisoner population. The dataset contains 1193 individuals, though 95 individuals were missing key variables and were excluded leaving 1098 individuals for analysis. Available information included hepatitis B virus (HBV) infection status, age, sex, and continuous optimal density (OD) reading from a Murex ICE enzyme immunoassay on oral fluid samples. For complete details regarding the data and associated study protocol, see Allwright et al.[44] This dataset contains individual patient OD readings, allowing us to construct, test, and decode master pools according to the three classification approaches outlined in Section 3. In so doing, we follow the approach of McMahan et al,[22] Delaigle and Hall,[24] and Mokalled et al,[26] who also used this data to demonstrate the performance of various group testing methodologies. We consider the performance of our method under four different information settings (limited information, inaccurate information, high quality information, and perfect information). Our goal is threefold: to estimate the logistic regression model linking the individuals' disease statuses to their associated risk factors, to estimate the underlying distribution of the OD values of the positive and negative individuals, and to correctly classify each individual as positive or negative for HBV. To relate the patients' ages and sexes to their infection statuses, we assume

$$P(Y_i = 1|x_{1i}, x_{2i}) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}),$$

where x_{1i} and x_{2i} denote the age and sex of the ith patient, respectively.

We evaluated the performance our method on the HBV data using the three classification approaches outlined in Section 3: individual, Dorfman, and array testing. Our quality control variant of Dorfman testing was implemented with pools of size 3 and 5 and our array based testing procedure with 3×3 and 5×5 arrays. We note that $N = 1098$ does not divide equally into pools of size 5 or 5×5 arrays. One approach to address this is to utilize "remainder pools" or to test "left over" individuals individually. However, both of these options make it difficult to compare performance fairly across testing methods and pool sizes. Instead, we constructed and simulated testing of 500 datasets, each of which consisted of 900 randomly selected individuals. Proceeding in this fashion allows for a direct comparison across the three classification techniques. In implementing these techniques, individuals were assigned to pools/arrays at random and pooled responses were simulated by taking the arithmetic mean of OD readings of the members of each pool. We applied the retesting approaches depicted in Figure 1 using one round of retesting and one associated model update. If an individual was retested, their individual OD reading was used as their retest observation.

To fit our model, we assume that the distributions of the OD readings for positive and negative individuals are well-approximated by gamma distributions; Mokalled et al[26] concluded that this assumption was reasonable for these data. The parameters of these distributions were assigned independent gamma priors whose parameters were selected to reflect four different settings; namely, a limited information, inaccurate information, high quality information, and perfect information setting. For specific details of these specifications see Web Appendix A. This was done to gauge the performance of our methodology across a broad spectrum of potential settings a practitioner may face. In all scenarios, prior distributions for $\boldsymbol{\beta}$ and τ^2 were assigned as described in Section 2 with $\boldsymbol{\Sigma}_\beta = 100\boldsymbol{I}$ and $\alpha_\tau = \beta_\tau = 3$. Our posterior sampling algorithm was used to obtain a sample of size 10 000 from the posterior distribution, after discarding a burn-in sample to ensure convergence. The individual testing scenarios and the high quality information setting for Dorfman testing with pools of size 5 required a burn-in period of 50 000 iterations with the other scenarios requiring only 20 000 iterations. Convergence was assessed via trace plots. The label switching problem occurred in approximately 2% of the datasets, with the overwhelming majority of label switching (all but 8 of the instances) occurring in the unreliable prior scenarios. These instances were replaced with additional model fits. Based on the posterior sample, we obtain point estimates (estimated posterior means) of the model parameter and associated measures of uncertainty (estimated posterior SD). Further, based on the estimated diagnostic thresholds, we also classify each individual.

Table 2 contains the average posterior mean estimate and the average estimated posterior SD. The table also summarizes the average number of tests and the classification performance, using the HBV statuses in the data as the true statuses. From these results, we see that analyzing data arising from our variants of Dorfman and array testing provide for approximately the same level of accuracy as analyzing individual level data, though there was a modest increase in the

TABLE 2 Data application results: The presented summary includes the average posterior mean estimate (Est.) and the average estimated posterior SD for the regression coefficients $\boldsymbol{\beta}$ obtained from individual (A1), Dorfman (A2), and array (A3) testing from the HBV data application

			β_0	β_1	β_2					
Prior specification	***A***	***c***	**Est. (SD)**	**Est. (SD)**	**Est. (SD)**	**TN**	**TP**	**FN**	**FP**	**#**
Limited information	A1	1	−4.30 (0.44)	0.05 (0.01)	−0.73 (0.99)	0.98	1.00	0.00	0.02	900.00
	A2	3	−4.29 (0.44)	0.06 (0.01)	−0.99 (1.13)	0.98	1.00	0.00	0.02	479.85
		5	−5.41 (0.89)	0.04 (0.03)	−3.02 (2.96)	0.99	0.99	0.01	0.01	416.26
	A3	3	−4.32 (0.44)	0.06 (0.01)	−0.84 (1.05)	0.99	1.00	0.00	0.01	672.17
		5	−4.42 (0.46)	0.06 (0.01)	−0.93 (1.12)	0.99	0.99	0.01	0.01	460.48
Inaccurate information	A1	1	−4.15 (0.41)	0.05 (0.01)	−0.52 (0.89)	0.96	1.00	0.00	0.04	900.00
	A2	3	−4.14 (0.43)	0.05 (0.01)	−0.59 (0.95)	0.98	1.00	0.00	0.02	429.82
		5	−4.43 (0.46)	0.06 (0.01)	−1.01 (1.15)	0.98	1.00	0.00	0.02	432.27
	A3	3	−4.16 (0.42)	0.05 (0.01)	−0.61 (0.97)	0.98	1.00	0.00	0.02	681.82
		5	−4.38 (0.47)	0.05 (0.02)	−1.02 (1.14)	0.99	1.00	0.00	0.01	469.854
High quality information	A1	1	−4.27 (0.43)	0.05 (0.01)	−0.90 (1.10)	0.98	1.00	0.00	0.02	900.00
	A2	3	−4.62 (0.56)	0.06 (0.02)	−1.60 (1.77)	0.99	1.00	0.00	0.01	469.07
		5	−4.95 (0.79)	0.05 (0.03)	−2.30 (2.36)	0.99	1.00	0.00	0.01	432.38
	A3	3	−4.27 (0.44)	0.05 (0.01)	−0.89 (1.10)	0.99	1.00	0.00	0.01	674.72
		5	−4.36 (0.45)	0.06 (0.03)	−0.95 (1.07)	0.99	1.00	0.00	0.01	462.66
Perfect information	A1	1	−4.32 (0.44)	0.05 (0.01)	−0.96 (1.10)	0.98	1.00	0.00	0.02	900.00
	A2	3	−4.45 (0.49)	0.06 (0.01)	−1.29 (1.39)	0.99	1.00	0.00	0.01	464.97
		5	−5.07 (0.84)	0.04 (0.02)	−2.40 (2.68)	0.99	1.00	0.00	0.01	428.69
	A3	3	−4.36 (0.45)	0.06 (0.01)	−1.06 (1.13)	0.99	1.00	0.00	0.01	716.85
		5	−4.37 (0.45)	0.06 (0.01)	−1.05 (1.12)	0.99	1.00	0.00	0.01	459.52

Abbreviations: A, retesting protocol; c, pool size.

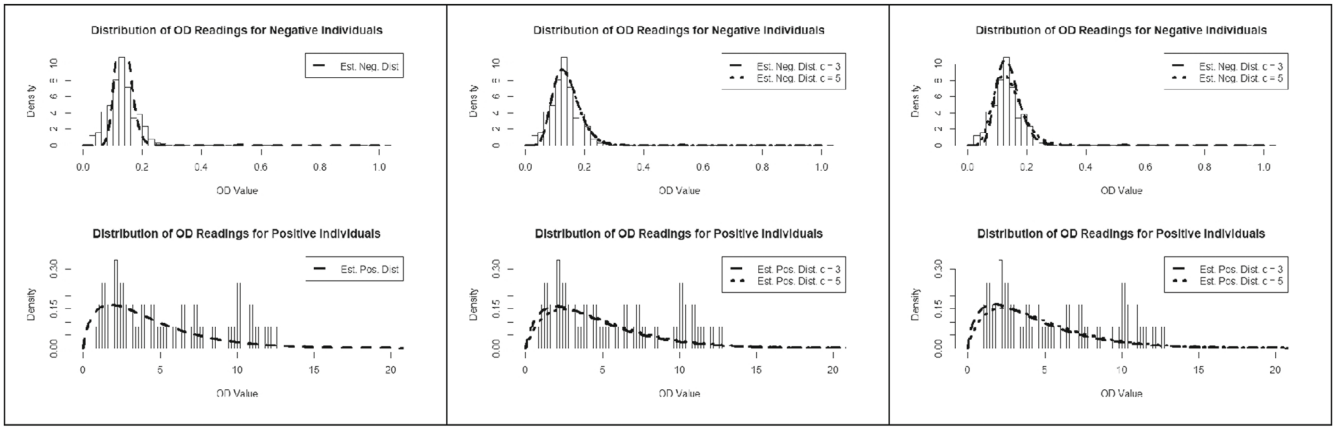

FIGURE 3 Data application results: The figure displays both the empirical and the average estimated biomarker distributions for negative and positive individuals obtained from individual (left), Dorfman (center) and array (right) testing using inaccurate information for the HBV data application. Note that the axis scales differ between the histograms of the negative and positive individuals

size of the SEs for Dorfman testing with pools of size 5. Further, it took approximately half the number of tests to attain these estimates, with virtually no loss in classification accuracy. Notably, the precision and accuracy under the misspecified priors are comparable to that observed for the other prior configurations, indicating that our method is robust to prior misspecification. Figure 3 displays the estimated biomarker distribution for positive and negative individuals from the inaccurate information configurations, plotted against a histogram of the OD reading from positive and negative individuals. Results for the other three configurations are displayed in Web Figures 6-8. From this figure we again see that the proposed approach is capable of well estimating the biomarker distributions, even when strongly informative misspecified prior distributions are used. In summary, the findings from this analysis reinforce the findings from our simulation studies. That is, the proposed approach can estimate both the regression model linking the individuals' disease statuses to their associated risk factors and the underlying biomarker distributions, and in so doing our approach provides a path for precise classification, all while directly accounting for the dilution effect and measurement error.

6 | DISCUSSION

We have developed a Bayesian group testing regression methodology which can be applied to continuous observations arising from any group testing procedure (Dorfman testing, array testing, etc.) for the purposes of estimating both a regression function and the underlying biomarker distributions of the positive and negative individuals. Our modeling technique allows us to directly account for the dilution effect as well as measurement error. Further, our approach can be used to estimate optimal pool and individual classification thresholds via the estimated biomarker distributions. We have assessed the performance of our method under a variety of conditions with an in-depth simulation study and we have further demonstrated our technique by applying it to HBV data collected on Irish prisoners. Given the multitude of group testing protocols which have been proposed, an exploration of which protocol is most efficient for our method would be a worthwhile pursuit. To further disseminate our work, code (written in R) which implements every aspect of our approach has been developed and has been made freely available at https://github.com/scwatson812/GT_Dilution.

A number of possible extensions of this methodology are possible. While an appropriate distributional family exists for many biomarker concentrations, a semi-parametric or non-parametric approach would be desirable for scenarios in which researchers are uncomfortable making assumptions about the underlying form of the biomarker distributions. Further, the work described herein may not be applicable to polymerase chain reaction (PCR) testing. That is, PCR tests render a diagnosis via a cycle threshold (CT) value, which represents the number of amplification cycles required for the signal from the targeted genetic sequence to cross the detection threshold. While CT values are related to the amount of targeted genetic material present in the original sample, the relationship is more complex than the relationship assumed in Section 2. That said, once the distributional relationship for PCR testing is made, our general framework can be applied seamlessly. Based on this realization, coupled with the widespread use of PCR based testing, we believe extending our proposed methodology in this manner would be a worthwhile pursuit. Further, given that our approach can estimate

the parameters of the biomarker distributions through the analysis of group testing data. These estimates can then be used to set diagnostic thresholds. Over time, as more information becomes available it might be reasonable to treat these estimates as "known" quantities. Proceeding in this fashion would allow one to identify optimal thresholds based on the approach of Wang et al.[43] Given this potential, further work is needed to determine the optimal time for transitioning to fixed thresholds, and the consequences of transitioning too early. Lastly, another direction for future work could involve developing methods that can be used to assess goodness-of-fit for regression models that are fit based on group testing data. This could be particularly challenging given that the individuals' true infection statuses are latent.

ACKNOWLEDGEMENTS

Stella Self was partially supported by NIGMS award number P20GM130420 and NIAID award number R25AI164581. Christopher McMahan was partially supported by NSF award number OIA-1826715, NIAID award number R01 AI121351 and Office of Naval Research award number N00014-19-1-2295. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the National Science Foundation, or the Office of Naval Research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The Hepatitis B data were provided to us by Dr. Shane Allwright (Trinity College Dublin).

ORCID

Stella Self https://orcid.org/0000-0002-5956-5220
Christopher McMahan https://orcid.org/0000-0001-5056-9615

REFERENCES

1. Dorfman R. The detection of defective members of large populations. *Ann Math Stat*. 1943;14(4):436-440. doi:10.1214/aoms/1177731363
2. Krajden M, Cook D, Mak A, et al. Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *J Clin Virol*. 2014;61(1):132-137.
3. Saá P, Proctor M, Foster G, et al. Investigational testing for ZIKA virus among U.S. blood donors. *N Engl J Med*. 2018;378(19):1778-1788. doi:10.1056/NEJMoa1714977
4. Van TT, Miller J, Warshauer DM, et al. Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *J Clin Microbiol*. 2012;50(3):891-896. doi:10.1128/JCM.05631-11
5. Mutesa L, Ndishimye P, Butera Y, et al. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature*. 2020;589:276-280. doi:10.1038/s41586-020-2885-5
6. Muñoz-Zanzi CA, Johnson WO, Thurmond MC, Hietala SK. Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle. *J Vet Diagn Investig*. 2000;12(3):195-203. doi:10.1177/104063870001200301
7. Ly A, Dhand NK, Sergeant ESG, Marsh I, Plain KM. Determining an optimal pool size for testing beef herds for Johne's disease in Australia. *PLoS One*. 2019;14(11):1-18. doi:10.1371/journal.pone.0225524
8. Schmidt M, Roth WK, Meyer H, Seifried E, Hourfar MK. Nucleic acid test screening of blood donors for orthopoxviruses can potentially prevent dispersion of viral agents in case of bioterrorism. *Transfusion*. 2005;45(3):399-403.
9. Kainkaryam R, Woolf PJ. Pooling in high-throughput drug screening. *Curr Opin Drug Discov Devel*. 2009;12(3):339-350.
10. Amos CI, Frazier ML, Wang W. DNA pooling in mutation detection with reference to sequence analysis. *Am J Hum Genet*. 2000;5:1689-1692. doi:10.1086/302894
11. Thompson KH. Estimation of the proportion of vectors in a natural population of insects. *Biometrics*. 1962;18(4):568-578.
12. Chiang CL, Reeves WC. Statistical estimation of virus infection rates in mosquito vector populations. *Am J Hyg*. 1962;75(3):377-391. doi:10.1093/oxfordjournals.aje.a120259
13. Hung M, Swallow WH. Robustness of group testing in the estimation of proportions. *Biometrics*. 1999;55(1):231-237.
14. Farrington C. Estimating prevalence by group testing using generalized linear models. *Stat Med*. 1992;11(12):1591-1597.
15. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000;56(4):1126-1133.
16. Xie M. Regression analysis of group testing samples. *Stat Med*. 2001;20(13):1957-1969.
17. Wang D, McMahan C, Gallagher C, Kulasekera K. Semiparametric group testing regression models. *Biometrika*. 2014;101(3):587-598.
18. Delaigle A, Hall P, Wishart J. New approaches to non-and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*. 2014;101:567-585.
19. Delaigle A, Meister A. Nonparametric regression analysis for group testing data. *J Am Stat Assoc*. 2011;106(494):640-650.

20. Delaigle A, Hall P. Nonparametric regression with homogeneous group testing data. *Ann Stat*. 2012;40(1):131-158. doi:10.1214/11-AOS952
21. Wang D, Zhou H, Kulasekera KB. A semi-local likelihood regression estimator of the proportion based on group testing data. *Journal of Nonparametric Statistics*. 2013;25(1):209-221. doi:10.1080/10485252.2012.750726
22. McMahan C, Tebbs J, Bilder C. Regression models for group testing data with pool dilution effects. *Biostatistics*. 2013;14(2):284-298.
23. Wang D, McMahan CS, Gallagher CM. A general regression framework for group testing data, which incorporates pool dilution effects. *Stat Med*. 2015;34(27):3606-3621. doi:10.1002/sim.6578
24. Delaigle A, Hall P. Nonparametric methods for group testing data, taking dilution into account. *Biometrika*. 2015;102(4):871-887. doi:10.1093/biomet/asv049
25. Warasi MS, McMahan C, Tebbs J, Bilder C. Group testing regression models with dilution submodels. *Stat Med*. 2017;36(30):4860-4872.
26. Mokalled SC, McMahan CS, Tebbs JM, Andrew Brown D, Bilder CR. Incorporating the dilution effect in group testing regression. *Stat Med*. 2021;40:2540-2555. doi:10.1002/sim.8916
27. Phatarfod R, Sudbury A. The use of a square array scheme in blood testing. *Stat Med*. 1994;13(22):2337-2343.
28. Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*. 2007;63(4):1152-1163.
29. Kim H, Hudgens M. Three-dimensional array-based group testing algorithms. *Biometrics*. 2009;65(3):903-910.
30. Gastwirth J, Johnson W. Screening with cost-effective quality control: potential applications to HIV and drug testing. *J Am Stat Assoc*. 1994;89(427):972-981.
31. Johnson W, Gastwirth J. Dual group screening. *J Stat Plan Infer*. 2000;83(2):449-473.
32. Bolstad WM, Curran JM. *Introduction to Bayesian Statistics*. New York: John Wiley & Sons; 2016.
33. Klauenberg K, Walzel M, Ebert B, Elster C. Informative prior distributions for ELISA analyses. *Biostatistics*. 2015;16(3):454-464. doi:10.1093/biostatistics/kxu057
34. Diebolt J, Robert CP. Estimation of finite mixture distributions through Bayesian sampling. *J Royal Stat Soc Ser B (Methodol)*. 1994;56(2):363-375.
35. Stephens M. Dealing with label switching in mixture models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2000;62(4):795-809. doi:10.1111/1467-9868.00265
36. Rodríguez CE, Walker SG. Label switching in Bayesian mixture models: deterministic relabeling strategies. *J Comput Graph Stat*. 2014;23(1):25-45. doi:10.1080/10618600.2012.735624
37. Papastamoulis P, Iliopoulos G. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J Comput Graph Stat*. 2010;19(2):313-331. doi:10.1198/jcgs.2010.09008
38. Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J Royal Stat Soc Ser B (Stat Methodol)*. 1997;59(4):731-792. doi:10.1111/1467-9868.00095
39. Chung H, Loken E, Schafer JL. Difficulties in drawing inferences with finite-mixture models. *Am Stat*. 2004;58(2):152-158. doi:10.1198/0003130043286
40. Kunkel D, Peruggia M. Anchored Bayesian Gaussian mixture models. *Electron J Stat*. 2020;14(2):3869-3913. doi:10.1214/20-EJS1756
41. Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc*. 1993;88(422):669-679.
42. Polson N, Scott J, Windle J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J Am Stat Assoc*. 2013;108(504):1339-1349.
43. Wang D, McMahan C, Tebbs J, Bilder C. Group testing case identification with biomarker information. *Comput Stat Data Anal*. 2018;122:156-166.
44. Allwright S, Bradley F, Long J, Barry J, Thornton L, Parry JV. Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: results of a national cross sectional survey. *BMJ*. 2000;321(7253):78-82. doi:10.1136/bmj.321.7253.78

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: Self S, McMahan C, Mokalled S. Capturing the pool dilution effect in group testing regression: A Bayesian approach. *Statistics in Medicine*. 2022;41(23):4682-4696. doi: 10.1002/sim.9532